**Running head:** Machine Learning of Pain Outcomes Using EEG.

# Systematic Review of the Effectiveness of Machine Learning Algorithms for Classifying Pain Intensity, Phenotype or Treatment Outcomes Using Electroencephalogram Data

Tyler Mari[1]

Jessica Henderson[1]

Michelle Maden[2]

Sarah Nevitt[2]

Rui Duarte[2#]

Nicholas Fallon[1#]

[1]Department of Psychology, University of Liverpool, Liverpool, UK

[2] Department of Health Data Science, Liverpool Reviews and Implementation Group, University of Liverpool, Liverpool, UK

[#] These authors contributed equally to this manuscript.

Address for correspondence: Tyler Mari, Department of Psychology, University of Liverpool, 2.21 Eleanor Rathbone Building, Bedford Street South, Liverpool L69 7ZA, UK. E-mail: Tyler.Mari@liverpool.ac.uk

Conflict of interest statement: The authors declare that they have no conflicts of interest.

Sources of financial support: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

**Abstract**

Recent attempts to utilise machine learning (ML) to predict pain-related outcomes from Electroencephalogram (EEG) data demonstrate promising results. The primary aim of this review was to evaluate the effectiveness of ML algorithms for predicting pain intensity, phenotypes or treatment response from EEG. Electronic databases MEDLINE, EMBASE, Web of Science, PsycINFO and The Cochrane Library were searched. A total of 44 eligible studies were identified, with 22 presenting attempts to predict pain intensity, 15 investigating the prediction of pain phenotypes and seven assessing the prediction of treatment response. A meta-analysis was not considered appropriate for this review due to heterogeneous methods and reporting. Consequently, data were narratively synthesised. The results demonstrate that the best performing model of the individual studies allows for the prediction of pain intensity, phenotypes and treatment response with accuracies ranging between 62% to 100%, 57% to 99% and 65% to 95.24%, respectively. The results suggest that ML has the potential to effectively predict pain outcomes, which may eventually be used to assist clinical care. However, inadequate reporting and potential bias reduce confidence in the results. Future research should improve reporting standards and externally validate models to decrease bias, which would increase the feasibility of clinical translation.

Perspective: This systematic review explores the state-of-the-art machine learning methods for predicting pain intensity, phenotype or treatment-response from EEG data. Results suggest that machine learning may demonstrate clinical utility, pending further research and development. Areas for improvement, including standardised processing, reporting and the need for better methodological assessment tools, are discussed.

**Keywords:** Machine Learning, Pain Intensity, Pain Phenotypes, Systematic Review, Treatment Response

## Introduction

Accurate assessment of pain is challenging due to the complex interplay between biological and psychological processes, but it is vital for understanding the effectiveness of clinical pain management [19,80,106]. Traditionally, pain is evaluated using interviews, observations, psychological screening and rating scales [11,19,38,101]. Whilst behavioural tools are valuable, developments are needed to individualise clinical care further, as many conventional methods fail in individuals who cannot accurately communicate their pain, such as infants and those with dementia [11,39]. Moreover, imperfect tools, coupled with the complexity of pain, also inhibit accurate diagnoses and treatment, further limiting the management of clinical pain [11,27,92]. Consequently, improved pain assessment is required to individualise clinical pain care.

Recent attempts at improving the detection of pain outcomes using neuroimaging and Machine Learning (ML) have seen promising results [59]. ML refers to an algorithm that learns complex data patterns and makes predictions without being explicitly programmed [75]. Supervised learning is the most applicable method to pain prediction, whereby labelled input data are propagated through an algorithm, which then learns patterns associated with each label [48,56,89,96]. This is achieved by altering internal weights; minimising the error between the input and the predicted label using optimisation algorithms, such as gradient descent [49,56,100]. Therefore, the algorithm learns from experience and can then be used to predict the labels of novel, unseen data [55]. We focus on the application of ML on Electroencephalogram (EEG), as it is inexpensive and accessible, making it an excellent candidate for clinical applications [31,84]. However, neuroimaging methods of pain classification are not the only promising

approach within this line of research. Alternative approaches such as pain prediction from facial expressions also demonstrate promising results and can be identified elsewhere[8,52,71]. Additionally, due to the technicality of ML and the corresponding algorithms, we also provide reference to comprehensive overviews of ML, which can be retrieved to make ML more accessible and provide an intuition regarding the underlying mechanisms of ML algorithms [6,24,45,55,76,89].

By applying supervised ML, researchers have successfully decoded patterns of neuronal activation arising from pain-related outcomes [59]. The development of computational methods of pain assessment may allow for the prediction of pain intensity, phenotype or response to treatment should research demonstrate its effectiveness. Pain intensity reflects self-reported pain ratings arising from experimental pain stimulation or naturally occurring pain. Pain phenotypes broadly reflect characteristics of pain conditions, suggesting the presence of a condition, whilst treatment response involves predicting the effect of pain treatments. The validation of ML and EEG for clinical use may improve clinical provision and mitigate current limitations by introducing objective markers, which could guide individualised treatment and diagnosis [20,22]. For example, predicting treatment effectiveness could reduce ineffective trial-and-error treatment and improve patient outcomes [30–32]. Despite their potential, pain biomarkers have not significantly impacted public health or clinical practice to-date [104]. Therefore, throughout this systematic review, we discuss the effectiveness of ML for predicting pain outcomes from EEG whilst concurrently discussing the benefits and challenges, alluding to the potential for clinical translation. We address the research question: how effective are machine learning algorithms for predicting pain intensity, phenotype or response to treatment from EEG data? We included research on

healthy participants or chronic pain populations. To achieve this, we complete the following objectives:

(i) To evaluate the effectiveness of ML by comparing performance metrics.

(ii) To explore the benefits and challenges of ML, alluding to the feasibility of clinical translation.

(iii) To evaluate the quality of these studies.

## Methods

This systematic review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [61]. The review protocol was registered on PROSPERO on June 5th, 2020 as CRD42020172091.

### Search Strategy

Electronic databases MEDLINE, EMBASE, Web of Science, PsycINFO and The Cochrane Library were searched from inception to May 4th, 2020 and updated on May 10th, 2021, using a combination of free text and thesaurus terms and restricted to English language. The searches were comprised of terms relating to pain, ML and EEG. Pain terms included pain conditions (e.g., neuralgia) and pain synonyms (e.g., nociception), whilst ML terms included methods (e.g., decision tree) and ML synonyms (e.g., classification) and EEG mostly included unabbreviated terminology (e.g., electroencephalogram). Reference lists of eligible studies

and similar publications were hand-searched to identify further potentially relevant studies. The complete search strategy is presented in supplementary material 1.

**Study Selection**

Firstly, two reviewers (TM and JH) independently screened the title and abstracts of all the unique search results to identify all potentially relevant studies to be retrieved for full-text review. Secondly, full-text articles retrieved in stage one were reviewed for inclusion independently by two reviewers (TM and JH). The screening stages were guided by the eligibility criteria outlined in Table 1. Reviewer discrepancies at either stage were resolved through discussion or consultation with a third reviewer (NF), who acted as an arbiter, if necessary.

*Table 1. Eligibility Criteria*

| Inclusion criteria (included if all of the following are satisfied) | Exclusion criteria (excluded if any of the following are met) |
|---|---|
| 1. Published peer-reviewed studies presenting original data predicting pain intensity, phenotype or response to treatment. | 1. Non-peer reviewed citations (abstracts or conference proceedings, letters and commentaries). Non-original data or case reports. |
| 2. Human participants ≥ 18 years old. | 2. Non-human sample, or human participants < 18 years old. |
| 3. EEG study. | 3. Non-EEG study. |
| 4. Applied supervised ML. | 4. Did not apply supervised ML. |
| 5. English full text. | 5. Non-English texts. |

EEG, electroencephalogram; ML, machine learning.

**Data Extraction**

A data extraction form was developed to retrieve data regarding the study authors, participant demographics, type of painful stimuli, treatment type (where applicable), pain condition (where applicable), EEG array, model features, prediction type (binary, multiclass or continuous), the algorithm used, model validation and the performance metrics for the best performing model. The data extraction was performed independently by one reviewer (TM) and checked for accuracy by a second reviewer (JH). Disagreements were resolved through discussion or consultation with a third reviewer (NF), who acted as an arbiter, if necessary.

The model we report is intended to reflect the best performing algorithm, which is deemed as the one with the greatest performance metrics (e.g., highest accuracy), as several models are typically developed in each study. If the authors attempt different classifications (binary, multiclass or continuous prediction), we report the best performing model of each classification type. The model reported is defined as the best performing either in the original studies or based on our judgement when the original studies did not define the best performing model. The majority of the studies implement cross-validation methods. The cases where cross-validation was not performed or was unclear are highlighted in the respective tables. Through reporting the best performing model, we hope to present the current state-of-the-art methods, which may eventually be candidates for clinical translation. A definition of the typical performance metrics reported in this review can be seen in Table 2. A comprehensive discussion of the performance metrics has been reported elsewhere [15,42,67,82,85,102].

*Table 2. General definitions of ML metrics*

| Metric | Notation | Explanation |
|---|---|---|
| Accuracy | $\frac{tp + tn}{tp + fp + tn + fn}$ | The algorithm's overall effectiveness. Reflects the ratio of correctly classified data points over all data points. |
| AUC (BCA) | $\frac{1}{2}\left(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}\right)$ | The AUC represents the ability of the classifier to avoid incorrect classification. |
| F1 | $\frac{2tp}{2tp + fp + fn}$ | Represents the harmonic mean of PPV (Precision) and Sensitivity (Recall, TPR). |
| FPR | $\frac{fp}{fp + tn}$ | Represents the ratio of negative classes incorrectly labelled as positive cases over the total number of negative labels. |
| MAE | $\frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$ | Represents average absolute error between the actual output value $(y_i)$ and the predicted output value$(\hat{y}_i)$ . |
| Misclassification | $\frac{fp + fn}{tp + fp + tn + fn}$ | Represents the ratio of incorrectly labelled predictions over all data points. |
| NPV | $\frac{tn}{tn + fn}$ | Represents the ratio of correctly labelled negative cases over the total negative predictions made. |
| PPV (Precision) | $\frac{tp}{tp + fp}$ | Represents the ratio of correctly labelled positive cases over the total positive predictions made. |
| Sensitivity (Recall; TPR) | $\frac{tp}{tp + fn}$ | The ability of the algorithm to correctly identify true positive cases. |
| Specificity | $\frac{tn}{tn + fp}$ | The ability of the algorithm to correctly identify true negative cases. |

AUC, area under the ROC curve; BCA, balanced classification accuracy; fn, false negatives; fp, false positives; FPR, false positive ratio; MAE, mean absolute error; NPV, negative predictive value; PPV, positive predictive value, tn, true negatives; tp, true positives; TPR, true positive ratio; ROC, receiver operating characteristics.

**Risk of Bias**

Assessment of risk of bias (ROB) was performed by using the prediction model risk of bias assessment tool (PROBAST) [103], which contains 20 signalling questions to assess ROB across four domains: participants, predictors, outcomes and analysis [62,103]. Each domain is assessed as low, high or unclear ROB. An overall ROB is calculated for each study, taking all domains into consideration. Studies are deemed low ROB providing all individual domains were scored as low ROB. If one or more of the domains were scored as unclear ROB, but all other domains were low ROB, the study should be labelled as unclear ROB. Finally, if one or more of the domains is scored as high ROB, then the overall ROB would be deemed as high, regardless of the scores on the other domains [103]. Additionally, PROBAST allows assessment of the applicability of each study to the review, which is assessed and scored in a similar way as the ROB analysis, with studies being scored as low, high or unclear regarding applicability issues. PROBAST does not evaluate the applicability of the analysis, so the applicability assessment only consists of the participants, predictors and outcome domains. The applicability assessment evaluates whether there are any concerns regarding the relevance of an individual study to the review question [103]. For example, if a model was developed on participants in a different setting to the one specified in the review the question, then the model may not be applicable to the originally defined setting, and therefore, the study would be deemed as having high concerns regarding applicability. No studies were excluded based on the ROB or applicability assessments. PROBAST assessment was performed by one reviewer (TM), and a random sample of articles (≈ 20%) was checked for agreement by a second reviewer (NF).

**Reporting Standards**

The reporting standards of ML studies were assessed using the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) guidelines [17]. TRIPOD consists of 22 items assessing the reporting standards of research studies developing or validating a multivariable prediction model. Items that are not relevant for all review outcomes (e.g., treatment details) were denoted as not applicable (NA). Additionally, TRIPOD items 4b and 5a were omitted due to lack of relevance. Many studies in this review were lab-based, and therefore reporting key dates and study setting is uncommon. Items 11, 14b and 17 were removed as they are optional and were not relevant to this review. Item 15a was omitted as it was relevant to traditional prediction studies but did not apply to ML. Item 15b was removed as it was not fully applicable to ML without altering the item. Additionally, all non-development items were excluded as they were not applicable to the studies in this review. As the reporting standards of medical ML studies have shown low adherence to recommended guidelines [109], we aimed to assess the quality of reporting throughout the literature. Assessment of reporting standards was performed by one reviewer (TM), and approximately 20% of articles were randomly sampled and checked for consistency by a second reviewer (NF).

**Data Synthesis**

The heterogeneity of the literature was assessed by the similarity of study populations and methods (ML and Neuroimaging). A meta-analysis was not considered appropriate for this review due to the absence of consistent reporting standards (see Reporting Standards

Assessment), differences in study designs, methods, classification definitions and, in some cases, inadequate numerical data presented within the publications. Consequently, we performed a narrative synthesis, adhering to the synthesis without meta-analysis (SWiM) guidelines [12]. The included studies are aligned with one of the three review outcomes, pain intensity, pain phenotype or response to treatment. The data has been narratively synthesised by presenting the range of the performance metrics reported in each section (e.g., accuracy, sensitivity or specificity) for each review outcome. However, inconsistent reporting means that the performance metrics reflect a subset of the sample.

## Results

The searches resulted in the identification of a total of 1384 results, comprised of 1380 citations from the searches and four studies from manual identification. Following the removal of 165 duplicate results, the title and abstracts of 1219 records were screened for relevance, resulting in 92 potentially relevant articles retrieved for full-text review. A total of 48 studies were excluded at the full-text review stage. Reasons for exclusion can be identified in the PRISMA flow chart in Figure 1. Subsequently, a total of 44 results were included in this review, with 22 evaluating the prediction of pain intensity [4,5,60,64,65,68,73,78,87,88,93,94,7,107,108,13,25,29,37,46,47,51], 15 of pain phenotypes [2,3,84,86,91,97,105,14,28,33,50,66,74,77,83] and seven of response to treatment [31,32,34–36,43,99].

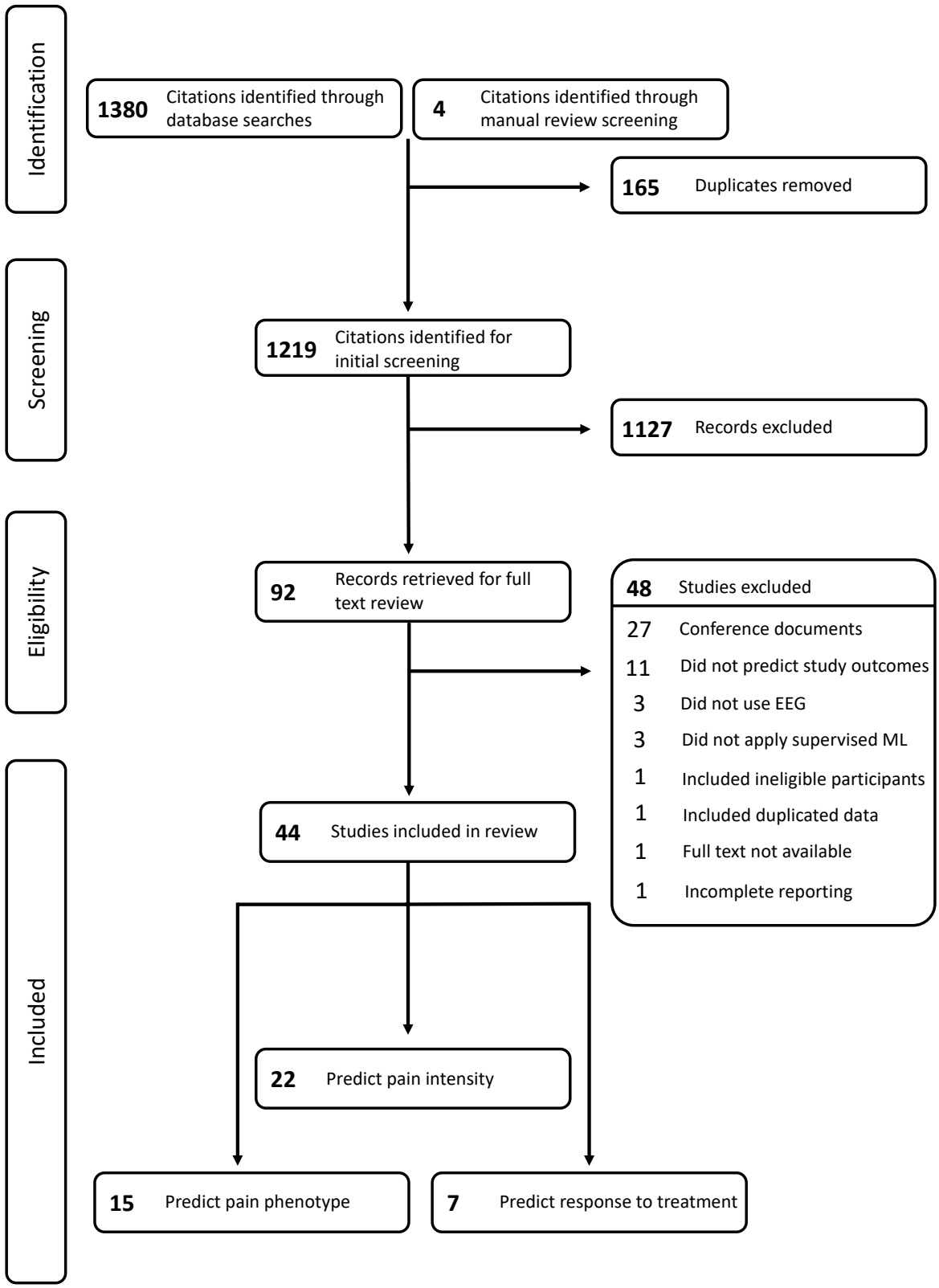


*Figure 1.* PRISMA flow chart.

**PROBAST Assessment**

The ROB assessment demonstrated that 42 of the 44 studies in this review were categorised as high ROB, as summarised in Figure 2. The full assessment is presented in supplementary material 2. Concerning the participant domain, the most significant concern for bias resulted from sample issues, such as small sample sizes (typically $\leq$ 20 participants) or insufficient sample diversity (e.g., only male participants), with 12 of the 44 studies being scored high ROB. Additionally, five studies were deemed unclear for the participant domain as the inclusion and exclusion criteria were not clearly defined. The studies deemed at either high or unclear ROB for the outcomes domain were labelled as such due to missing or unclear outcome definitions (e.g., grouping justifications). Here, three studies were scored as high ROB, whilst one was deemed unclear ROB. The majority of the studies in this review were deemed as having high ROB in the analysis domain. The most common reason for high ROB arises from insufficient external validation, in-line with the PROBAST expectations (e.g., temporal or geographical validation) [62], with 42 of the 44 studies being scored as high ROB on the analysis domain. Overall, the results presented in the synthesis should be interpreted with caution. Many of the studies synthesised are at a high ROB, and therefore, it is unclear to what extent the results generalise.

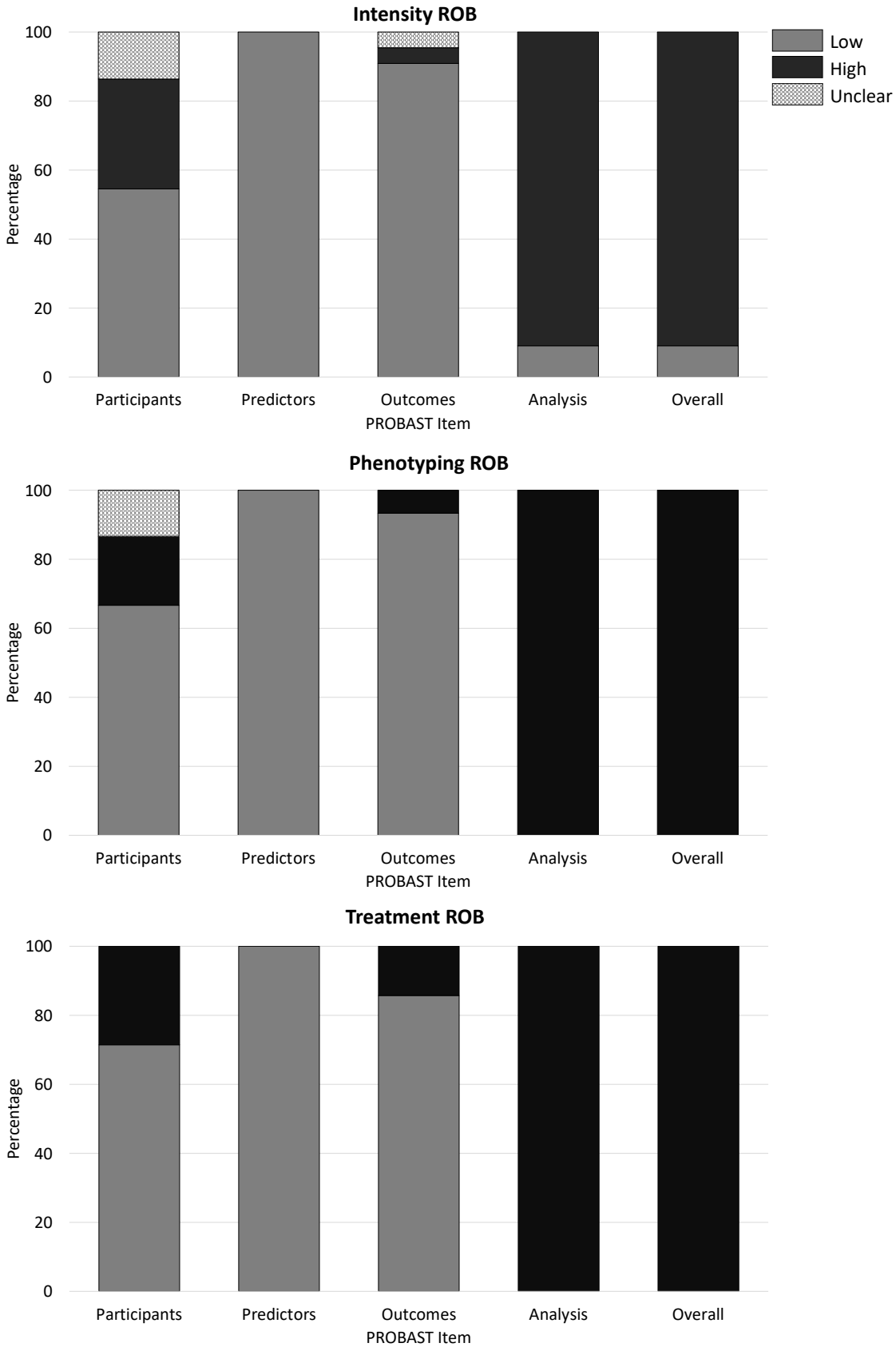


*Figure 2. PROBAST assessments for pain intensity, pain phenotyping and response to treatment studies.*

The applicability assessment demonstrated that only one of the 44 included studies was deemed as having applicability concerns [65]. The study was deemed as having high applicability concerns on the outcome domain [65]. Here, the study predicted stimuli intensity rather than directly predicting pain intensity. All other domains had low concerns regarding applicability. No other studies were deemed high or unclear regarding applicability to the review question. The full applicability assessment is presented in supplementary material 3.

**Reporting Standards Assessment**

The assessment of reporting standards demonstrated relatively low adherence to reporting guidelines. The areas with the lowest adherence across studies included the title and abstract, where none of the articles met TRIPOD expectations. Regarding the title, none of the studies were entitled as developing a prediction model. The abstract adherence was more varied, but generally studies did not report model discrimination or calibration in line with TRIPOD expectations. Additionally, the majority did not report the number of outcome events in the abstract. Many of the studies included also had low adherence throughout the methods. For example, only two of the studies reported their justification for the sample size and only around half of the intensity and phenotyping studies reported the presence and handling of missing data. Concerning model performance, many of the studies did not sufficiently define or report all metrics following the guidance of TRIPOD. Moreover, the majority of the studies in the intensity and phenotype clusters did not sufficiently discuss the clinical or research implications of the prediction model. Other domains also had relatively low adherence and can be seen in the TRIPOD summary in Table 3. However, the low adherence to reporting

guidelines could be partially explained by the compatibility of the tools used (see review limitations).

*Table 3. TRIPOD summary for all of the review outcomes*

| | Number Reported, n (%) | | |
|---|---|---|---|
| Tripod Item | Pain Intensity (N = 22) | Pain Phenotype (N = 15) | Treatment Response (N = 7) |
| **Title** | | | |
| Identify the study as developing and/or validating a multivariable prediction model, the target population, and the outcome to be predicted. | 0 (0%) | 0 (0%) | 0 (0%) |
| **Abstract** | | | |
| Provide a summary of objectives, study design, setting, participants, sample size, predictors, outcome, statistical analysis, results, and conclusions. | 0 (0%) | 0 (0%) | 0 (0%) |
| **Background and Objectives** | | | |
| Explain the medical context (including whether diagnostic or prognostic) and rationale for developing or validating the multivariable prediction model, including references to existing models. | 19 (86%) | 8 (53%) | 6 (86%) |
| Specify the objectives, including whether the study describes the development or validation of the model or both. | 4 (18%) | 1 (7%) | 0 (0%) |
| **Method** | | | |
| Describe the study design or source of data (e.g., randomized trial, cohort, or registry data), separately for the development and validation data sets, if applicable. | 21 (95%) | 15 (100%) | 7 (100%) |
| **Participants** | | | |
| Describe eligibility criteria for participants. | 17 (77%) | 15 (100%) | 7 (100%) |
| Give details of treatments received, if relevant. | NA | NA | 7 (100%) |
| **Outcome** | | | |
| Clearly define the outcome that is predicted by the prediction model, including how and when assessed. | 21 (95%) | 14 (93%) | 7 (100%) |
| Report any actions to blind assessment of the outcome to be predicted. | 21 (95%) | 15 (100%) | 7 (100%) |
| **Predictors** | | | |
| Clearly define all predictors used in developing or validating the multivariable | 22 (100%) | 14 (93%) | 7 (100%) |

| | | | |
|---|---|---|---|
| prediction model, including how and when they were measured. | | | |
| Report any actions to blind assessment of predictors for the outcome and other predictors. | 22 (100%) | 15 (100%) | 7 (100%) |
| **Sample Size** | | | |
| Explain how the study size was arrived at. | 0 (0%) | 1 (7%) | 1 (14%) |
| **Missing Data** | | | |
| Describe how missing data were handled (e.g., complete-case analysis, single imputation, multiple imputation) with details of any imputation method. | 11 (50%) | 7 (47%) | 7 (100%) |
| **Statistical Analysis** | | | |
| Describe how predictors were handled in the analyses. | 22 (100%) | 15 (100%) | 7 (100%) |
| Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation. | 19 (86%) | 14 (93%) | 5 (71%) |
| Specify all measures used to assess model performance and, if relevant, to compare multiple models. | 0 (0%) | 0 (0%) | 1 (14%) |
| **Results: Participants** | | | |
| Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful. | 8 (36%) | 12 (80%) | 5 (71%) |
| Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome. | 5 (23%) | 7 (47%) | 6 (86%) |
| **Model Development** | | | |
| Specify the number of participants and outcome events in each analysis. | 12 (55%) | 12 (80%) | 6 (86%) |
| **Model Performance** | | | |
| Report performance measures (with confidence intervals) for the prediction model. These should be described in results section of the paper. | 0 (0%) | 0 (0%) | 0 (0%) |
| **Discussion: Limitations** | | | |
| Discuss any limitations of the study. | 14 (64%) | 8 (53%) | 7 (100%) |
| **Interpretation** | | | |
| Give an overall interpretation of the results considering objectives, limitations, results from similar studies and other relevant evidence. | 22 (100%) | 15 (100%) | 7 (100%) |
| **Implication** | | | |

| | | | |
|---|---|---|---|
| Discuss the potential clinical use of the model and implications for future research. | 4 (18%) | 6 (40%) | 5 (71%) |
| **Other Information: Supplementary Information** | | | |
| Provide information about the availability of supplementary resources, such as study protocol, web calculator, and data sets. | 5 (23%) | 3 (20%) | 1 (14%) |
| **Funding** | | | |
| Give the source of funding and the role of the funders for the present study. | 3 (14%) | 0 (0%) | 2 (29%) |

## Pain Intensity

The characteristics of the 22 included studies which investigated pain intensity are reported in Table 4. The articles attempt binary classification, multiclass classification, continuous score prediction or a combination of any of these methods. All of the studies in the intensity section predict differing levels of pain intensity. For example, binary classification may discriminate classes such as no pain versus pain or low pain versus high pain. In contrast, multiclass classification occurs when *n* (*n>2)* different levels of pain are used as classes for prediction. These classes typically reflect broad pain classes (e.g., low, medium or high pain). In some instances, the continuous pain rating scale is converted to classes for classification, such that the number of classes reflects the responses on the rating scale. Here the number is treated as a label rather than a numerical value. Finally, continuous prediction attempts to identify the numerical value of reported pain intensity on a numerical rating scale. Continuous prediction differs from the previous example as the prediction is a numerical value rather than a discrete label.

Of the 22 included studies, a total of 13 perform binary classification [4,5,88,93,94,7,13,37,46,60,65,73,78], eight implement multiclass classification [5,25,47,64,87,94,107,108] and five conduct continuous prediction [7,29,51,68,88]. The algorithms used within these studies are Support Vector Machines (SVM) [4,5,47,60,64,65,73,78,88,93], with one study using a Support Vector Regression (SVR) [88], regression models, including linear and logistic [7,29,51,68], Artificial Neural Networks (ANN), which includes Convolutional Neural Networks (CNN), Multilayer Perceptrons, and other feed-forward neural networks (e.g. Sparse Bayesian Extreme Learning Machine; SBELM) [13,25,46,87,107,108], Linear Discriminant Analysis (LDA) [7], Random Forest models (RF) [94] and one study used a Mahalanobis classifier [37].

*Table 4. Summary of pain intensity studies.*

| Authors | Classification Type | Sample Demographics (Mean age ± Standard Deviation) | EEG Montage | Feature Category | Best Algorithm | Outcome | Performance Metrics | |
|---|---|---|---|---|---|---|---|---|
| Alazrai et al. (2019) | Binary | 24 Healthy Subjects (12 F, 22.5 ± 3.2) | 14 EEG Electrodes | Time Frequency | SVM (RBF Kernel) | No Pain vs Pain | Accuracy | 89.2% ± 3.2% |
| | | | | | | | F1 (No Pain) | 87.4% ± 4.1% |
| | | | | | | | F1 (Pain) | 89.5% ± 3.3% |
| Alazrai et al. (2019) | Binary | 24 Healthy Subjects (13 M, 23.5 ± 2.3) | 14 EEG Electrodes | Time Frequency | SVM (RBF Kernel) | No pain vs Pain | Accuracy | 93.86% |
| | | | | | | | Precision | 94.02% |
| | | | | | | | Specificity | 93.92% |
| | | | | | | | Sensitivity | 88.88% |
| | | | | | | | F1 | 90.58% |
| | Multiclass | | | | SVM (RBF Kernel) | No Pain vs No Pain-to-pain vs Pain | Accuracy | 90.18% |
| | | | | | | | Precision | 91.34% |
| | | | | | | | Specificity | 95.10% |
| | | | | | | | Sensitivity | 86.99% |
| | | | | | | | F1 | 88.75% |
| Bai et al. (2016) | Binary | 34 Healthy Subjects (17 F, 21.6 ± 1.7) | 64 EEG Electrodes | Event Related Potentials | LDA | Low Pain vs High Pain | Accuracy | 70.36% ± 14.18% |
| | Continuous | | | | Linear Regression | Pain Rating (4-10; High Pain Trials) | MAE | 1.173 ± 0.278 |
| Cao et al. (2020) | Binary | 18 Healthy Subjects (10 M, 25 ± 3.5) | 16 EEG Electrodes | Time Frequency | SBELM | No Pain vs Pain | Train Accuracy | 89.3% ± 3.4% |
| | | | | | | | Accuracy | 90.1% ± 2.8% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | AUC | 0.95 |
| Elsayed et al. (2020) | Multiclass | 30 Healthy Subjects (17 M, 24 ± 3) | 8 EEG Electrodes | Time Frequency | ANN (Three hidden layers) | No Pain vs Low Pain vs Moderate Pain vs High Pain | Accuracy<br>Precision<br>Recall<br>F1 | 94.83%<br>93.92%<br>95.14%<br>94.17% |
| Furman et al. (2018) | Continuous | 44 Healthy Subjects* (22 M, 28.4) | 64 EEG Electrodes | Time Frequency | Leave one out Regression | Pain Intensity (0-100) | r | 0.55 |
| Hadjileontiadis (2015) | Binary | 17 Healthy Subjects (9 M, 23.22 ± 1.72) | 14 EEG Electrodes | Time Frequency | Mahalanobis classifier | No Pain vs Pain | Accuracy | 90.25% ± 2.08% |
| Kaur et al. (2019) | Binary | 39 Healthy Subjects (34 M, 24.59 ± 3.03) | 4 EEG Electrodes | Time Frequency | MLPNN (One hidden layer with 9 Neurons) | No Pain vs Pain | Train Accuracy<br>Test Accuracy<br>CV Accuracy | 97.29%<br>90%<br>82.73% |
| Kimura et al. (2021) | Multiclass | 23 Subjects with hip Osteoarthritis or Osteonecrosis who underwent total hip arthroplasty (18 F, 64.6 ± 11.9) | 1 EEG Electrode | Time Frequency | SVM (RBF Kernel) | No Pain vs Mild Pain vs Moderate Pain vs Severe Pain | Accuracy<br>Precision$^{+}$<br>Recall$^{+}$<br>F1$^{+}$ | 79.6%$^{++}$<br>78.28% ± 6.03%$^{++}$<br>77.03% ± 9.05%$^{++}$<br>77.67% ± 7.41%$^{++}$ |
| Li et al. (2018) | Continuous | 34 Healthy Subjects* (17 F, 21.6 ± 1.7) | 64 EEG Electrodes | Event Related Potentials | Linear Regression | Continuous Pain Ratings | MAE | 1.19 ± 0.35 |
| Misra et al. (2017) | Binary | 30 Healthy Subjects (16 F, 20 ± 2) | 128 EEG Electrodes | Time Frequency | SVM (RBF Kernel) | Low Pain vs High Pain | Accuracy<br>Misclassification | 89.58%<br>10.42% |
| Nezam et al. (2021) | Multiclass | 24 Healthy Subjects (15 M, 25) | 30 EEG Electrodes | Time Frequency | SVM (RBF Kernel) | No Pain vs Low Pain vs High Pain<br><br>No Pain vs Low Pain vs Moderate Pain vs High Pain vs Intolerable Pain | Accuracy<br>Specificity<br>Sensitivity<br>Accuracy<br>Specificity | 83% ± 5%<br>91% ± 4%<br>93% ± 5%<br>62% ± 6%<br>78% ± 3% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Sensitivity | 87% ± 4% |
| Okolo & Omurtag (2018) | Binary | 9 Healthy Subjects (7 M, Age Range 20 - ≥ 40) | 19 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | No Pain vs Low Stimulus | Accuracy[+] | 89.78% ± 5.97% |
| | | | | | | No Pain vs Max Stimulus | Accuracy[+] | 89.51% ± 8.36% |
| | | | | | | Low vs Max Stimulus | Accuracy[+] | 69.2% ± 12.02% |
| Prichep et al. (2018) | Continuous | 77 Chronic Pain Subjects* (53% F, 49.3 ± 15.8) | 19 EEG Electrodes | Time Frequency | Stepwise Logistic Regression | Continuous Pain Rating (0 - 10) | r | 0.907[++] |
| Sai et al. (2019) | Binary | 10 Parturient Women (29.6 ± 4.9) | 16 EEG Electrodes | Time Frequency | SVM (RBF Kernel) | No Pain vs Pain | Accuracy | 84% |
| | | | | | | | Sensitivity | 87.20% |
| | | | | | | | Specificity | 81.10% |
| Schulz et al. (2012) | Binary | 23 Healthy Subjects (14 F, 26) | 64 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Low Pain vs High Pain | Accuracy | 62% |
| | | | | | | Pain Sensitive vs Pain Insensitive | Accuracy | 83% |
| | | | | | | | Sensitivity | 50% |
| | | | | | | | Specificity | 100% |
| Tripanpitak et al. (2020) | Multiclass | 13 Healthy Subjects (8 M, 33.2 ± 7.9) | 16 EEG Electrodes | Event Related Potentials | ANN (One hidden Layer with 10 neurons) | No Pain vs Pain vs Max Pain | Train Accuracy | 100% |
| | | | | | | | Accuracy | 100% |
| | | | | | | No Pain vs Sensation vs Pain vs Max Pain | Train Accuracy | 87.50% |
| | | | | | | | Accuracy | 94.40% |
| Tu et al. (2016) | Binary | 96 Healthy Subjects* (51 F, 21.6 ± 1.7) | 64 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Low Pain vs High Pain | Accuracy | 83.5% ± 6.8% |
| | | | | | | | Sensitivity | 79.2% ± 14.6% |
| | | | | | | | Specificity | 72.2% ± 14.2% |
| | Continuous | | | | SVR | Continuous Pain Rating (0 - 10) | MAE | 1.15 ± 0.32 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Vatankhah et al. (2013) | Binary | 15 Healthy Subjects* (8 F, 28) | 12 EEG Electrodes | Time Frequency | SVM (ANFIS adapted RBF) | No Pain vs Pain | Accuracy | 95% |
| | | | | | | Pain vs Intolerable Pain | Accuracy | 75% |
| Vijayakumar et al. (2017) | Binary | 25 Healthy Subjects (11 F, Median Age 24) | 64 EEG Electrodes | Time Frequency | RF Model | No Pain vs Pain | BCA | 95.33% ± 0.6% |
| | Multiclass | | | | RF Model | Categorised Pain Rating (1-10) | BCA | 89.45% ± 1.05% |
| Yu et al. (2020) | Multiclass | 32 Healthy Subjects (20 F, Age Range 19-35) | 32 EEG Electrodes | Time Frequency | CNN (Adam Optimiser) | No Pain vs Moderate Pain vs Severe Pain | Accuracy | 97.37% ± 0.26% |
| | | | | | | | Precision | 96.05% |
| | | | | | | | Specificity | 98.03% |
| | | | | | | | Sensitivity | 96.06% |
| | | | | | | | F1 | 96.05% |
| Yu et al. (2020) | Multiclass | 20 Healthy Subjects* (11 M, Age Range 23-42) | 32 EEG Electrodes | Time Frequency | SFNN (ELM) | No Pain vs Minor Pain vs Moderate Pain vs Severe Pain | Accuracy | 68.9% ± 3.12% |

Key: * Number of participants used in the final model is different from the overall reported sample size. ✢ Manually averaged performance metrics. The values here represents the average across participants or conditions, which is not reported in the original paper. ✢✢ Cross-validation method unclear or not reported.

ANFIS, adaptive network fuzzy inference system; ANN, artificial neural network; AUC, area under the ROC curve; BCA, balanced classification accuracy; CNN, convolutional neural network; CV, cross-validation; EEG, electroencephalogram; ELM, extreme learning machine; F, females; LDA, linear discriminant analysis; M, males; MAE, mean absolute error; MLPNN, multilayer perceptron neural network; RBF, radial basis function; RF, random forest; ROC, receiver operating characteristics; SBELM, sparse bayesian extreme learning machine; SFFN, single-hidden-layer feed-forward neural network; SVM, support vector machine; SVR, support vector regression.

Regarding the prediction of no pain conditions relative to pain conditions, the studies in this review have yielded accuracies between 82.73% and 95.33% [4,5,13,37,46,65,73,93,94], with eight of nine studies obtaining an accuracy greater than 85% [4,5,13,37,46,65,93,94]. Additionally, five of the studies included in this review attempt to discern low pain and high pain classes [7,60,65,78,88]. The performance of these studies is more varied than the no pain and pain classification, with a range of accuracies between 62% and 89.58%. Here, only two of five studies achieved an accuracy of over 80% [60,88]. Taken together, the ability to discern binary pain intensity classes appears to be greater than chance levels. Here, detecting the presence of pain is achievable, with accuracies surpassing 80%, whilst discriminating low pain from high pain can be achieved with accuracies greater than 60%, with one study demonstrating an accuracy close to 90% [60].

Despite the promise of binary classification, the clinical utility of merely identifying the presence of pain or broad pain categories (low pain vs high pain) may be limited. As such, other studies included in this review attempt multiclass or continuous prediction, which increases the resolution of pain intensity that can be determined and thus improves the potential clinical relevance. For example, differentiating between just three classes of pain intensity (no pain, low pain and high pain) allows the inference of the presence of pain but also provides some indication regarding the intensity in the same classification, which would not be possible in a single binary classification. Summarising the multiclass performance is challenging, as the number of classes differs across studies (range 3 - 10 classes). Therefore, individual results should be referred to Table 4. Nevertheless, the accuracy range for the classification of three or more pain classes is between 62% and 100% [5,25,47,64,87,94,107,108]. These results suggest that pain classification at a finer resolution is achievable, with half of the eight studies achieving accuracies between 90% and 100% [5,25,87,107].

Finally, the ultimate goal of pain intensity prediction is to predict the actual pain intensity reported on a rating scale. The majority of the studies that perform a continuous prediction attempt to identify the pain rating reported on a 10- or 11-point scale [7,51,68,88], whilst one study attempted pain prediction using a 0 to 100 scale [29]. The performance of these algorithms is either evaluated using a correlation coefficient or their mean absolute error (MAE). The studies that evaluate their model's performance using MAE achieved an error between 1.15 and 1.19 [7,51,88]. Regarding the studies that evaluate their model using a correlation coefficient, the two studies achieved a positive correlation between predicted pain intensity and actual pain intensity between 0.55 and 0.907 [29,68].

**Pain Phenotypes**

The characteristics of the 15 phenotyping studies are reported in Table 5. To achieve consistency within the reporting of this narrative review, the phenotyping studies can be further divided into subgroups. Since all of the phenotyping studies utilised binary classification, the studies were divided based on the types of groups or conditions predicted. One study attempted multiclass classification in addition to binary classification [50]. We do not synthesise the multiclass results, as they are only comprised of a single study. However, the performance metrics for the multiclass classification are reported in Table 5.

Six of the 15 phenotyping studies attempt to predict migraine phenotypes [2,3,14,28,83,86]. Within these six studies, four classified migraine versus healthy controls [2,3,83,86], one classified migraine with aura versus migraine without aura [28] and one classified the interictal phase

versus the preictal phase of migraine [14]. Furthermore, five of the 15 studies predicted neuropathic or neurogenic pain [74,77,91,97,105]. Four of the five studies above predicted the presence of neuropathic pain or neurogenic pain versus healthy controls [74,77,91,97], and one study classified neuropathic patients into two groups: pain below the lesion versus without pain below the lesion [105]. Furthermore, one study classified a broad group of chronic pain patients versus healthy controls [84]. Here, the chronic pain group consisted of various conditions, including chronic back pain, chronic widespread pain, joint pain, unspecific neuropathic pain, postherpetic neuralgia and polyneuropathic pain. Additionally, one study classified fibromyalgia patients versus healthy controls [66]. Moreover, one study classified radiculopathy versus healthy controls [50]. Here, the authors also perform multiclass classification of radiculopathy subjects, individuals with chronic lumbar pain scheduled to receive an implanted spinal cord stimulator and healthy subjects. Finally, one study predicted experimentally induced visceral hypersensitivity versus a placebo condition [33]. SVM was the most common algorithm [3,14,28,50,66,74,84,91] including SVR [33], whilst ANN [2,86], discriminant analysis [77,97,105] and RF models [83] were also used.

*Table 5. Summary of pain phenotyping studies.*

| Authors | Classification Type | Sample Demographics (Mean age ± Standard Deviation) | EEG Montage | Feature Category | Best Algorithm | Outcome | Performance Metrics | |
|---|---|---|---|---|---|---|---|---|
| Akben et al. (2012) | Binary | 30 participants; 15 Migraine (13 F), 15 Healthy Controls (10 F). Age Range 20 - 35 | 18 EEG Electrodes | Time Frequency | MLPNN (One hidden layer with 50 neurons) | Healthy Control vs Migraine | Accuracy<br>Sensitivity<br>Specificity | 93.33%<br>93.33%<br>93.33% |
| Akben et al. (2016) | Binary | 60 Participants; 30 Migraine (21 F), 30 Healthy Controls (19 F). Age Range 20 - 40 | 18 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Healthy Control vs Migraine | Accuracy<br>Sensitivity<br>Specificity | 88.40%<br>90%<br>86.70% |
| Cao et al. (2018) | Binary | 80 Participants; 40 Migraine (30 F, 38.1 ± 8.2). 40 Healthy Controls (32 F, 36.1 ± 9.8) | 4 EEG Electrodes | EEG Entropy | SVM (RBF Kernel) | Interictal Phase vs Preictal phase | Accuracy<br>Sensitivity (Recall)<br>Precision (PPV)<br>F1 | 76% ± 4%<br>75% ± 5%<br>75% ± 5%<br>74% ± 6% |
| De Tommaso et al. (1999) | Binary | 120 Migraine (80 F, 36.7 ± 4.5), 51 Healthy Controls (36 F). Age Range 25-46 | 12 EEG Electrodes | Time Frequency | ANN (Two hidden neurons) | Healthy Control vs Migraine | Sensitivity<br>FPR | 95.83%<br>4.16% |
| Frid et al. (2020) | Binary | 53 Participants* (All with episodic migraine). Age Range 18 - 75 | 32/64** EEG Electrodes | Time Frequency | SVM (RBF Kernel) | MWA vs MWoA | Accuracy<br>AUC | 84.62%<br>0.8813 |
| Graversen et al. (2011) | Binary | 15* Healthy Participants (11 M, 32.9) | 3 EEG Electrodes | Time Frequency | SVR (Linear Kernel) | Visceral Hypersensitivity Sensitisation vs placebo Condition | Accuracy | 91.70% |
| Levitt et al. (2020) | Binary | 57 Participants; 20 Radiculopathy (11 F, 54.25), 20 Healthy Controls (11 F, 54.15), 17 Chronic Lumbar scheduled to receive implanted SCS (10 F, 56.88) | 16 EEG Electrodes | Time Frequency | SVM (RBF Kernel) | Healthy Control vs Radiculopathy | Accuracy<br>AUC<br>Accuracy | 82.50%<br>0.8225<br>71.90% |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | Multiclass | | | | | Healthy Control vs Radiculopathy vs Pre-SCS | AUC (Radiculopathy)<br>AUC (Healthy)<br>AUC (Pre-SCS) | 0.828<br>0.842<br>0.962 |
| Paul et al. (2019) | Binary | 32 Participants; 16 Fibromyalgia (12 F, 46.81 ± 4.28), 16 Healthy Controls (12 F, 45.19 ± 4.48) | 8 EEG Electrodes | Time Frequency | SVM (Polynomial Kernel) | Healthy Control vs Fibromyalgia | Accuracy<br>Sensitivity<br>Specificity<br>Precision (PPV) | 96.15%<br>96.88%<br>95.65%<br>93.94% |
| Saif et al. (2021) | Binary | 30 Participants; 10 Healthy Controls (7 M, 39.6 ± 10.2), 10 PNP (8 M, 43.8 ± 9.1), 10 PWP (7 M, 46.2 ± 9.4) | 61 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Healthy Control vs PWP<br>Healthy Control vs PNP<br>PWP vs PNP | Accuracy<br>Accuracy<br>Accuracy | 99% ± 0.49%<br>97% ± 0.6%<br>91% ± 1% |
| Sarnthein et al. (2006) | Binary | 30 Participants; 15 Neurogenic Pain (9 M, Median Age 64), 15 Healthy Controls (8 F, Median Age 60) | 60 EEG Electrodes | Time Frequency | LDA | Healthy Control vs Neurogenic Pain | Accuracy<br>CI | 87%[+,+]<br>69% - 96% |
| Subasi et al. (2019) | Binary | 30 Participants; 15 Migraine (13 F, 27 ± 4.4), 15 Healthy Controls (10 F, 26 ± 5.3) | 18 EEG Electrodes | Time Frequency | RF Model | Healthy Control vs Migraine | Accuracy<br>Sensitivity<br>Specificity | 85.95%<br>85.20%<br>86.70% |
| Ta Dinh et al. (2019) | Binary | 185 Participants; 101 Chronic Pain*[+] (69 F, 58.2 ± 13.5), 84 Healthy Controls (55 F, 57.8 ± 14.6) | 64 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Healthy Control vs Chronic Pain | Accuracy<br>Sensitivity<br>Specificity | 57% ± 4%<br>60% ± 5%<br>57% ± 5% |
| Vanneste et al. (2018) | Binary | 342 Participants; 78 Neuropathic Pain (43 M, 47.39 ± 10.26), 264 Healthy Controls (152 M, 49.51 ± 12.54) | 19 EEG Electrodes | Time Frequency | SVM | Healthy Control vs Neuropathic Pain | Accuracy<br>Sensitivity (TPR)<br>FPR<br>AUC | 92.53% ± 1.59%<br>93% ± 2%<br>21% ± 2%<br>0.95 ± 0.01 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Vuckovic et al. (2018) | Binary | 21 Participants*; 11 Neuropathic Pain (7 M, 44.9 ± 16.9), 10 Healthy Controls (7 M, 35.2 ± 7.2) | 48 EEG Electrodes | Time Frequency | LDA | Healthy Control vs Neuropathic Pain | Accuracy [95 CI] | 88% ± 10% [86%-89%] |
| | | | | | | | Sensitivity [95 CI] | 89% ± 7% [88%-90%] |
| | | | | | | | Specificity [95 CI] | 86% ± 12% [84%- 88%] |
| Wydenkeller et al. (2009) | Binary | 26 Participants* with Spinal cord injury (20 M,47 ± 15) | 32 EEG Electrodes | Time Frequency | DA | Participant with pain below the lesion vs Participant without pain below the lesion | Accuracy | 84.2%✢✢ |

Key: * Number of participants used in the final model is different from the overall reported sample size, ** 3 different EEG caps were used during this study, ✢ Various chronic pain conditions including: 47 with chronic back pain, 30 chronic widespread pain, 6 joint pain, 5 unspecific neuropathic pain, 7 postherpetic neuralgia, 6 polyneuropathic pain. ✢✢Cross-validation method unclear or not reported.

ANN, artificial neural network; AUC, area under the ROC curve; CI, confidence interval; DA, discriminant analysis; EEG, electroencephalogram; F, females; FPR, false positive ratio; LDA, linear discriminant analysis; M, males; MLPNN, multilayer perceptron neural network; MWA, migraine with aura; MWoA, migraine without aura; PNP, paraplegic without neuropathic pain; PWP, paraplegic with neuropathic pain; PPV, positive predictive value; RBF, radial basis function; RF, random forest; ROC, receiver operating characteristics; SCS, spinal cord stimulator; SVM, support vector machine; SVR, support vector regression; TPR, true positive ratio.

The majority of the studies included in the pain phenotyping section of this review attempt to phenotype different aspects of migraine. To summarise the performance of phenotyping migraine, we report the ranges of values obtained for accuracy, sensitivity and specificity across these studies [2,3,14,28,83,86]. However, not all of the studies reported include all three metrics and, therefore, each range reflects a proportion of the whole data set. Out of the six studies, five report accuracy [2,3,14,28,83], five report sensitivity [2,3,14,83,86] and three report specificity [2,3,83]. The ability to discriminate different characteristics of migraine ranges between 76% and 93.33%, 75% and 95.83%, 86.7% and 93.33% for accuracy, sensitivity and specificity, respectively.

The remaining studies in the phenotyping sections are more heterogeneous and are therefore inherently more challenging to group. However, the remaining studies are grouped based on the notion that they attempt to predict one or more chronic pain conditions (inclusive of experimentally induced hypersensitivity) compared with a group of healthy controls or predict the presence of pain relating to a lesion [33,50,66,74,77,84,91,97,105]. Again, not all of the studies report all of the required metrics. Consequently, synthesised results are reported from a subset of the final sample size of nine. All nine studies reported accuracy, whilst sensitivity and specificity were reported from four [66,84,91,97] and three studies [66,84,97], respectively. The accuracy range across these studies is 57% and 99%. Here, the sensitivity is between 60% and 96.88%, and the specificity is between 57% and 95.65%. Therefore, the results demonstrate that various chronic pain conditions can be identified with at least above chance level, with six studies surpassing 85% accuracy [33,66,74,77,91,97].

**Response to Treatment**

The characteristics of the seven treatment response studies are reported in Table 6. Two of the six studies classified active treatment or placebo conditions [34,35], whilst a further four predicted whether treatment was successful [31,32,36,99]. The final study for the response to treatment conducted a continuous prediction to assess the change in the brief pain inventory score after medication [43]. The models used within the response to treatment studies include SVMs [31,32,34,35], regression models, including linear and logistic [36,43] and a k-nearest neighbours algorithm [99].

*Table 6. Summary of response to treatment studies*

| Authors | Classification Type | Sample Demographics (Mean age ± Standard Deviation) | EEG Montage | Feature Category | Best Algorithm | Outcome | Performance Metrics | |
|---|---|---|---|---|---|---|---|---|
| Gram et al. (2015) | Binary | 32 Healthy Participants (17 M, 27.2 ± 7.1) | 62 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Responders vs Non-Responders (Response to Opioid; Morphine day) | Accuracy | 71.90% |
| | | | | | | | PPV | 70% |
| | | | | | | | NPV | 75% |
| | | | | | | Responders vs Non-Responders (Response to Opioid; Placebo day) | Accuracy | 71.90% |
| | | | | | | | PPV | 75% |
| | | | | | | | NPV | 68.80% |
| Gram et al. (2017) | Binary | 81 Participants (45 F); 51 Responders (26 F, 64.2 ± 10.4), 30 Non-Responders (19 F, 64.9 ± 15.7) | 34 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Responders vs Non-Responders (Response to Opioid) | Accuracy | 65% |
| | | | | | | | PPV | 76% |
| | | | | | | | NPV | 53% |
| Graversen et al. (2012) | Binary | 28 Participants with chronic pancreatitis; 14 Pregabalin group (8 F, 50), 14 Placebo group (11 M, 53) | 62 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Pregabalin Group vs Placebo Group | Accuracy | 85.70% |
| Graversen et al. (2015) | Binary | 21 Healthy Male Participants (20.35) | 62 EEG Electrodes | Time Frequency | SVM (Linear Kernel) | Remifentanil Group vs Placebo Group | Accuracy | 95.24% |
| Grosen et al. (2017) | Binary | 59 Patients with Chronic Pain (41 F, 55 ± 16) | 9 EEG Electrodes | Time Frequency | Logistic Regression | Successful vs Unsuccessful Clinical Treatment | OR | $1.18^{++}$ |
| | | | | | | | SE | 0.09 |
| | | | | | | | CI | 1.01 - 1.37 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Hunter et al. (2009) | Continuous | 12 Participants* with Fibromyalgia (9 F, 50.1 ± 8.2), 6 in treatment group, 6 in placebo group. | 35 EEG Electrodes | Time Frequency | Linear Regression | Brief Pain Inventory Change at Week 12 (Duloxetine Treatment) | Coefficient | 2.9[++] |
| | | | | | | | $R^2$ | 0.93 |
| Wei et al. (2020) | Binary | 70 Participants with Herpes Zoster; 45 Responders (25 M, 61 ± 11.8), 25 Non-Responders (14 F, 65.5 ± 8.7) | 32 EEG Electrodes | Time Frequency | KNN (K=5) | Responders vs Non-Responders | Accuracy | 80% ± 11.7% |
| | | | | | | | Sensitivity | 82.5 ± 14.7% |
| | | | | | | | Specificity | 77.7 ± 27.3% |
| | | | | | | | AUC | 0.85 |

Key: * Number of participants used in the final model is different from the overall reported sample size. ++Cross-validation method unclear or not reported.

AUC, Area under the ROC curve; CI, confidence interval; EEG, electroencephalogram; F, females; KNN, k-nearest neighbours; M, males; NPV, negative predictive value; OR, odds ratio; PPV, positive predictive value; ROC, receiver operating characteristics; SE, standard error; SVM, support vector machine.

The three studies that classified whether participants were responders or non-responders to treatment achieved accuracies between 65% and 80% [31,32,99]. Here, two studies achieved a positive predictive value (PPV) and negative predictive value (NPV) between 70% and 76% and 53% and 75% [31,32], respectively. Moreover, the final study that classified responders and non-responders to medication achieved a sensitivity of 82.5% and a specificity of 77.7% [99]. Regarding the classification of active treatment versus placebo groups, the two studies achieved accuracies between 85.7% and 95.24% [34,35]. The remaining two studies used regression models to predict treatment response [36,43].

## Discussion

This review investigated the effectiveness of ML for predicting pain-related outcomes, pain intensity, pain phenotypes and treatment response. Here, we focus on the potential usefulness of ML and EEG for pain outcome identification, rather than exploring the individual patterns of neural activation that constitute a biomarker. Other studies present overviews of the excellent utility of biomarkers in pain science [21,57,59]. Nevertheless, pain intensity reflects self-reported pain ratings resulting from naturalistic or experimentally induced pain. This review demonstrates that the presence of pain can be predicted, with all applicable studies demonstrating accuracies greater than 80% [4,5,13,37,46,65,73,93,94]. Regarding multiclass prediction, five out of eight studies demonstrated an accuracy of over 85% [5,25,87,94,107], with two surpassing 97% [87,107]. Furthermore, continuous pain ratings can be predicted with an error of approximately 10% on a 10- or 11-point rating scale [7,51,88]. The ability to detect pain intensity with an error of approximately one point on a rating scale demonstrates the potential of ML for pain prediction.

Concerning pain phenotyping, which reflects characteristics of pain conditions and may assist with diagnosis, our results show that pain conditions, such as migraine or neuropathic pain, can be discriminated above chance level (50%), with the majority of studies achieving an accuracy greater than 85% [2,3,33,66,74,77,83,91,97]. Regarding migraine, all relevant studies achieved over 75% accuracy, sensitivity and specificity, with four and three studies surpassing 85% sensitivity [2,3,83,86] and specificity [2,3,83], respectively. Moreover, regarding the prediction of pain conditions relative to controls, six of nine studies achieved accuracies of over 85% [33,66,74,77,91,97]. Additionally, three studies demonstrated a sensitivity of over 85% [66,91,97], with one demonstrating a sensitivity of almost 97% for detecting individuals with fibromyalgia relative to healthy controls [66]. However, the heterogeneity of the literature makes identifying specific use cases challenging currently. The scope of this review was to assess various phenotypes (as defined by the original authors), with no limit on inclusions, allowing for a diverse synthesis. As the field develops, we anticipate that narrower reviews will be conducted, which include alternative information such as the instruments used, providing a specific reference to researchers and clinicians in the field. However, this was beyond the scope of our review, as we believe that a broad synthesis is currently the most appropriate approach. Nevertheless, the results demonstrate the potential of EEG and ML to identify pain phenotypes that may eventually assist diagnostic assessments.

The results show that responders and non-responders to pain treatments can be classified with accuracies above 65% [31,32,99], whilst treatment and placebo groups can be predicted with accuracies greater than 85% [34,35]. However, the evidence suggests treatment response requires additional investigation, as it is currently under-researched. Additionally, the clinical

utility of predicting treatment response by classifying participants into responders and non-responders is unclear, whilst the demarcation can be heterogeneous and arbitrary [79,81]. The field might benefit more from parametric outcomes, such as predicting the reduction in subjective pain reported on a rating scale. Moreover, two studies that classified participants into responder status did so during tonic pain stimulation [31,32]. The clinical relevance is therefore currently questionable.

Should future research improve the current limitations and performance , ML may eventually be clinically advantageous by reducing trial-and-error treatment [30–32]. Indeed, the results across all three domains remain promising, but considering that 42 of the 44 studies in this review were deemed high ROB, there is a possibility that the synthesised results are inflated or are not fully generalisable. Therefore, we suggest that the results and, more importantly, the current clinical relevance of ML and EEG are tentatively interpreted.

Despite the concerns, predicting pain outcomes from EEG using ML may demonstrate clinical utility, should further research validate the technique. Detecting pain-related outcomes remains challenging [3,19,69], with many tools failing in those who cannot accurately communicate their pain, such as individuals with dementia [11,39]. Many of the studies in this review used ML to classify pain intensity in healthy individuals. However, a recent study demonstrated promising performance for identifying pain intensity in those with chronic pain [47], which demonstrates the potential for both pain identification in those with and without chronic pain conditions. Moreover, there are limited objective methods to ascertain clinical interventions effectiveness for a given patient. Should ML be eventually clinically validated, it could automate pain intensity or phenotype detection, benefiting patients and clinicians.

These tools could enable screening before a clinical assessment or facilitate improved diagnosis or prognosis [20,22]. For example, ML may allow clinicians to identify information in brief appointments, which currently cannot be achieved [21]; reducing patient visits. However, recording EEG from all patients is unnecessary and challenging. The most appropriate use case being for individuals who cannot communicate their pain accurately or at all. Indeed, improvements may be significant in conditions that are challenging to diagnose, such as migraine [3,69], where ML could assist both pain specialist and non-pain specialist clinicians. Eventually, ML could guide treatment. An algorithm that predicts treatment response would decrease ineffective treatments and patient suffering [31]. Should these algorithms be clinically validated, they could be applied throughout the clinical process, providing this is implemented ethically [22,23]. ML should not replace clinicians but instead, be used as an additional tool; automating routine tasks and increasing time with patients [1].

The promise of ML is exciting but not without challenges. Evidence demonstrating that ML significantly improves patient care is sparse [58]. Consequently, substantial clinical implementation is unlikely until the end of the decade [20]. Perhaps this is optimistic, as different algorithms and features are used, with little indication whether models can be effectively trained using similar features and methods [59]. It is unknown whether models trained on lab-based samples are ecologically valid and generalise to other samples or clinical settings [22,59]. The current lack of sufficient external validation is the primary ROB across the studies in this review and severely limits the clinical applicability of ML. Most of the studies in this review performed internal validation; mostly through cross-validation (e.g., k-fold). However, issues arise when using certain internal validation methods on small samples. For example, research has shown that k-fold validation likely overestimates performance in small

sample sizes, resulting in overfitting and ungeneralisable models [90]. Therefore, as many of the studies in this review had small samples and several performed k-fold validation, the generalisability of the prediction model is unclear. The authors also note that splitting the dataset into training and test sets provide robust estimates in small samples [90]. However, the PROBAST guidelines suggest that splitting the data into training and test sets is an insufficient form of internal validation, which is often erroneously referred to as external validation [103]. Developing and validating a model on the same participants is not appropriate evidence for potential clinical applications [70]. Therefore, to demonstrate sufficient evidence for clinical translation, extensive external validation is imperative [10]. In line with PROBAST recommendations, future prediction models should include either external temporal validation, whereby the testing data is collected at a later time period than the training data, or geographical validation, whereby data is collected by other investigators in a different location [103]. The latter, however, may require increased international collaboration and data sharing, which we strongly encourage. Alternatively, researchers can evaluate the model performance using data from a different study [17]. Nevertheless, external validation is essential for future research to thoroughly assess the clinical utility and generalisability of ML and EEG, whilst also reducing bias.

Many ML algorithms require specialist knowledge to implement, whilst EEG signals require pre-processing. Currently, ML is too user-dependent, and it is unlikely that clinicians will have the time to complete ML training. Convolutional neural networks (CNN), that can learn features directly from medical imaging; removing handcrafted feature selection [56,107], could be a potential solution. Only one study reviewed implemented CNNs, achieving 97% accuracy in a three-class paradigm [107]. In other medical fields such as skin cancer detection, CNNs

demonstrate comparable accuracy to experts [26]. However, CNNs are complex to interpret; hindering clinical applications [72]. Nevertheless, CNNs are worth exploring due to their potential for superior performance and the current lack of lab-based research.

The lack of standardisation in reporting across studies makes interpretation and replication difficult, whilst also increasing bias. This problem appears to be pervasive, and several studies have demonstrated that adherence to reporting standards are deficient across medical ML research [40,63,98,109]. A recent systematic review exploring the reporting quality of ML for medical diagnosis demonstrated that many studies lack sufficient details; hindering interpretation and replication [109]. They found that all 28 studies in their review did not follow reporting guidelines. Poor reporting makes it difficult for the end-user to assess the utility of ML [58]; providing a barrier for clinical uptake. Future research should adhere to reporting standards, to improve research clarity and allow for replication, which is imperative for clinical ML applications. Recently developed tools such as transparency, reproducibility, ethics and effectiveness (TREE) may improve reporting standards [95]. Additionally, the recent extensions to CONSORT and SPIRIT guidelines to include AI studies [18,53,54] are welcome and could lead to improved research quality with reduced bias.

The goal of this review was to explore the effectiveness of ML for predicting pain-related outcomes. Consequently, we reported the best performing algorithms in the respective studies identified by the systematic review. Whilst this highlights the potential of ML, it also poses the risk of inflating the current capability of ML for predicting pain-related outcomes from EEG data. This issue arises as many of the studies perform multiple classifications, using various algorithms. Consequently, several studies report models that have worse

performance metrics than those presented here. Therefore, our results do not represent the full state-of-play regarding ML and EEG for pain prediction, but instead presents the current state-of-the-art methods that may hold the potential for clinical translation.

Whilst the use of PROBAST and TRIPOD tools are appropriate for this review, and of an excellent standard in traditional prediction model studies, we found that they did not fully apply to ML studies. Therefore, the ROB and reporting standards assessments should be interpreted with caution. Altering the tools to fit certain studies increases the risk of arbitrary, non-replicable decisions, which does not present itself as a systematic process. Additionally, many of the TRIPOD items are highly stringent and even slight deviations result in the criteria not being met. For example, none of the studies met the title expectations as they were not titled as developing (or synonyms) a prediction model. As the tools are not fully applicable to ML, these slight differences may explain why many of the studies have low adherence to reporting standards. Therefore, more appropriate tools for assessment of ML and neuroimaging studies may be needed. Ongoing development of the TRIPOD ML [16], which is intended for ML will be a welcome addition to the tools available and will also be useful for researchers to use as a checklist to ensure that reporting standards have been sufficiently adhered to. Researchers may wish to use the current version of TRIPOD as an approximate guideline, until TRIPOD ML is available. Nevertheless, we strongly recommend that new tools are developed for ML and neuroimaging with clinical outcomes, that are not diagnostic or prognostic. Alternatively, standardised alterations to PROBAST allowing it to be applied to non-clinical and ML research, would also be welcomed. For example, altering the participant domain, such that the appropriateness of the sample size is assessed, rather than the sources of data would improve the applicability of this tool to lab-based research. Additionally, the

alteration or development of items to fit ML would also benefit the field. For example, an item assessing whether the classes of ML are approximately equal, or whether imbalanced classes have been handled appropriately, would be advantageous. Many ML algorithms struggle with imbalanced classes, as they typically focus on the dominant class, as the minor class does not hold much discriminatory significance, which can affect performance [9,41,44]. The development or alteration of such tools would improve scientific rigour; subsequently increasing clinical translation feasibility.

A formal assessment of certainty of the evidence could not be performed due to limitations in applicability of the ROB tools available but also assessment of GRADE domains such as inconsistency, and imprecision was hindered by a lack of reporting in the included studies of precision estimates such as 95% confidence intervals.

**Conclusion**

The results demonstrate that ML of EEG is an emerging area of research for pain prediction. Through further research and external validation, it may become feasible to adopt ML for clinical applications, with potential to individualise and improve the management of clinical pain. However, our systematic review demonstrates several limitations within the field which should be addressed in future research. Firstly, improved reporting standards are imperative to allow for thorough model evaluation. This would increase the transparency across studies and enable clearer interpretation of the clinical potential of ML. Secondly, future studies should be carefully designed, with a particular emphasis on the analysis protocol (e.g., external validation), to reduce the ROB. Additionally, we suggest that current ROB and

reporting standards tools are adapted, or new tools are developed, to enable a comprehensive assessment of quality for ML and neuroimaging studies. The lack of appropriate tools limits the current interpretation of the assessments and impacts the evaluation of results. Through the development of more appropriate tools and standardised processes, the research quality will improve, providing stronger evidence to develop the clinical potential of ML.

**References**


1. Ahuja AS: The impact of artificial intelligence in medicine on the future role of the physician. PeerJ 7:e7702, 2019. Available from: 10.7717/peerj.7702
2. Akben SB, Subasi A, Tuncel D: Analysis of Repetitive Flash Stimulation Frequencies and Record Periods to Detect Migraine Using Artificial Neural Network. J Med Syst 36:925–31, 2012. Available from: 10.1007/s10916-010-9556-2
3. Akben SB, Tuncel D, Alkan A: Classification of multi-channel EEG signals for migraine detection. Biomed Res 27:743–8, 2016.
4. Alazrai R, AL-Rawi S, Alwanni H, Daoud MI: Tonic Cold Pain Detection Using Choi–Williams Time-Frequency Distribution Analysis of EEG Signals: A Feasibility Study. Appl Sci 9:3433, 2019. Available from: 10.3390/app9163433
5. Alazrai R, Momani M, Khudair HA, Daoud MI: EEG-based tonic cold pain recognition system using wavelet transform. Neural Comput Appl 31:3187–200, 2019. Available from: 10.1007/s00521-017-3263-6
6. Alloghani M, Al-Jumeily D, Mustafina J, Hussain A, Aljaaf AJ: A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. Unsupervised Semi-Supervised Learn page 3–212020. Available from: 10.1007/978-3-030-22475-2_1
7. Bai Y, Huang G, Tu Y, Tan A, Hung YS, Zhang Z: Normalization of pain-evoked neural responses using spontaneous EEG improves the performance of EEG-based cross-individual pain prediction. Front Comput Neurosci Frontiers Media S.A.; 10:, 2016.
8. Bargshady G, Zhou X, Deo RC, Soar J, Whittaker F, Wang H: Enhanced deep learning algorithm development to detect pain intensity from facial expression images. Expert Syst Appl 149:113305, 2020. Available from: 10.1016/j.eswa.2020.113305

9. Bauder RA, Khoshgoftaar TM: The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. Heal Inf Sci Syst 6:9, 2018. Available from: 10.1007/s13755-018-0051-3

10. Bleeker S., Moll H., Steyerberg E., Donders AR., Derksen-Lubsen G, Grobbee D., Moons KG.: External validation is necessary in prediction research: J Clin Epidemiol 56:826–32, 2003. Available from: 10.1016/S0895-4356(03)00207-5

11. Breivik H, Borchgrevink PC, Allen SM, Rosseland LA, Romundstad L, Breivik Hals EK, Kvarstein G, Stubhaug A: Assessment of pain. Br J Anaesth 101:17–24, 2008. Available from: 10.1093/bja/aen103

12. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, Hartmann-Boyce J, Ryan R, Shepperd S, Thomas J, Welch V, Thomson H: Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ :l6890, 2020. Available from: 10.1136/bmj.l6890

13. Cao T, Wang Q, Liu D, Sun J, Bai O: Resting state EEG-based sudden pain recognition method and experimental study. Biomed Signal Process Control 59:101925, 2020. Available from: 10.1016/j.bspc.2020.101925

14. Cao Z, Lai K-L, Lin C-T, Chuang C-H, Chou C-C, Wang S-J: Exploring resting-state EEG complexity before migraine attacks. Cephalalgia 38:1296–306, 2018. Available from: 10.1177/0333102417733953

15. Chai T, Draxler RR: Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. Geosci Model Dev 7:1247–50, 2014.

16. Collins GS, Moons KGM: Reporting of artificial intelligence prediction models. Lancet 393:1577–9, 2019. Available from: 10.1016/S0140-6736(19)30037-6

17. Collins GS, Reitsma JB, Altman DG, Moons KGM: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): The TRIPOD Statement. Ann Intern Med 162:55, 2015. Available from: 10.7326/M14-0697

18. Cruz Rivera S, Liu X, Chan A-W, Denniston AK, Calvert MJ: Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nat Med 26:1351–63, 2020. Available from: 10.1038/s41591-020-1037-7

19. Dansie EJ, Turk DC: Assessment of patients with chronic pain. Br J Anaesth 111:19–25, 2013. Available from: 10.1093/bja/aet124

20. Davenport T, Kalakota R: The potential for artificial intelligence in healthcare. Futur Healthc J 6:94–8, 2019. Available from: 10.7861/futurehosp.6-2-94

21. Davis KD, Aghaeepour N, Ahn AH, Angst MS, Borsook D, Brenton A, Burczynski ME, Crean C, Edwards R, Gaudilliere B, Hergenroeder GW, Iadarola MJ, Iyengar S, Jiang Y, Kong J-T, Mackey S, Saab CY, Sang CN, Scholz J, Segerdahl M, Tracey I, Veasley C, Wang J, Wager TD, Wasan AD, Pelleymounter MA: Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. Nat Rev Neurol 16:381–400, 2020.

22. Davis KD, Flor H, Greely HT, Iannetti GD, Mackey S, Ploner M, Pustilnik A, Tracey I, Treede R-D, Wager TD: Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. Nat Rev Neurol 13:624–38, 2017. Available from: 10.1038/nrneurol.2017.122

23. Davis KD, Racine E, Collett B: Neuroethical issues related to the use of brain imaging: Can we and should we use brain imaging as a biomarker to diagnose chronic pain? Pain 153:1555–9, 2012. Available from: 10.1016/j.pain.2012.02.037

24. Dey A: Machine learning algorithms: a review. Int J Comput Sci Inf Technol 7:1174–9,

2016.

25. Elsayed M, Sim KS, Tan SC: A Novel Approach to Objectively Quantify the Subjective Perception of Pain Through Electroencephalogram Signal Analysis. IEEE Access 8:199920–30, 2020. Available from: 10.1109/ACCESS.2020.3032153

26. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S: Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–8, 2017. Available from: 10.1038/nature21056

27. Fine PG: Long-Term Consequences of Chronic Pain: Mounting Evidence for Pain as a Neurological Disease and Parallels with Other Chronic Disease States. Pain Med 12:996–1004, 2011. Available from: 10.1111/j.1526-4637.2011.01187.x

28. Frid A, Shor M, Shifrin A, Yarnitsky D, Granovsky Y: A Biomarker for Discriminating Between Migraine With and Without Aura: Machine Learning on Functional Connectivity on Resting-State EEGs. Ann Biomed Eng 48:403–12, 2020.

29. Furman AJ, Meeker TJ, Rietschel JC, Yoo S, Muthulingam J, Prokhorenko M, Keaser ML, Goodman RN, Mazaheri A, Seminowicz DA: Cerebral peak alpha frequency predicts individual differences in pain sensitivity. Neuroimage 167:203–10, 2018. Available from: 10.1016/j.neuroimage.2017.11.042

30. Ginsburg G, McCarthy JJ: Personalized medicine: revolutionizing drug discovery and patient care. Trends Biotechnol 19:491–6, 2001.

31. Gram M, Erlenwein J, Petzke F, Falla D, Przemeck M, Emons MI, Reuster M, Olesen SS, Drewes AM: Prediction of postoperative opioid analgesia using clinical-experimental parameters and electroencephalography. Eur J Pain (United Kingdom) Blackwell Publishing Ltd; 21:264–77, 2017.

32. Gram M, Graversen C, Olesen AE, Drewes AM: Machine learning on encephalographic

activity may predict opioid analgesia. Eur J Pain 19:1552–61, 2015. Available from: 10.1002/ejp.734

33. Graversen C, Brock C, Drewes AM, Farina D: Biomarkers for visceral hypersensitivity identified by classification of electroencephalographic frequency alterations. J Neural Eng 8:056014, 2011. Available from: 10.1088/1741-2560/8/5/056014

34. Graversen C, Malver LP, Kurita GP, Staahl C, Christrup LL, Sjøgren P, Drewes AM: Altered Frequency Distribution in the Electroencephalogram is Correlated to the Analgesic Effect of Remifentanil. Basic Clin Pharmacol Toxicol 116:414–22, 2015.

35. Graversen C, Olesen SS, Olesen AE, Steimle K, Farina D, Wilder-Smith OHG, Bouwense SAW, van Goor H, Drewes AM: The analgesic effect of pregabalin in patients with chronic pain is reflected by changes in pharmaco-EEG spectral indices. Br J Clin Pharmacol 73:363–72, 2012. Available from: 10.1111/j.1365-2125.2011.04104.x

36. Grosen K, Olesen AE, Gram M, Jonsson T, Kamp-Jensen M, Andresen T, Nielsen C, Pozlep G, Pfeiffer-Jensen M, Morlion B, Drewes AM: Predictors of opioid efficacy in patients with chronic pain: A prospective multicenter observational cohort study. PLoS One 12:, 2017. Available from: 10.1371/journal.pone.0171723

37. Hadjileontiadis LJ: EEG-Based Tonic Cold Pain Characterization Using Wavelet Higher Order Spectral Features. IEEE Trans Biomed Eng 62:1981–91, 2015. Available from: 10.1109/TBME.2015.2409133

38. Haefeli M, Elfering A: Pain assessment. Eur Spine J 15:S17–24, 2006. Available from: 10.1007/s00586-005-1044-x

39. Herr K, Coyne PJ, McCaffery M, Manworren R, Merkel S: Pain Assessment in the Patient Unable to Self-Report: Position Statement with Clinical Practice Recommendations. Pain Manag Nurs 12:230–50, 2011. Available from:

10.1016/j.pmn.2011.10.002

40. Heus P, Damen JAAG, Pajouheshnia R, Scholten RJPM, Reitsma JB, Collins GS, Altman DG, Moons KGM, Hooft L: Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. BMC Med 16:120, 2018. Available from: 10.1186/s12916-018-1099-2

41. Holder LB, Haque MM, Skinner MK: Machine learning for epigenetics and future medical applications. Epigenetics 12:505–14, 2017. Available from: 10.1080/15592294.2017.1329068

42. Hossin M, Sulaiman M.: A Review on Evaluation Metrics for Data Classification Evaluations. Int J Data Min Knowl Manag Process 5:01–11, 2015.

43. Hunter AM, Leuchter AF, Cook IA, Abrams M, Siegman BE, Furst DE, Chappell AS: Brain Functional Changes and Duloxetine Treatment Response in Fibromyalgia: A Pilot Study. Pain Med 10:730–8, 2009. Available from: 10.1111/j.1526-4637.2009.00614.x

44. Johnson JM, Khoshgoftaar TM: Survey on deep learning with class imbalance. J Big Data 6:27, 2019. Available from: 10.1186/s40537-019-0192-5

45. Jordan MI, Mitchell TM: Machine learning: Trends, perspectives, and prospects. Science (80- ) 349:255–60, 2015. Available from: 10.1126/science.aaa8415

46. Kaur M, Prakash NR, Kalra P, Puri GD: Electroencephalogram-Based Pain Classification Using Artificial Neural Networks. IETE J Res :1–14, 2019. Available from: 10.1080/03772063.2019.1702903

47. Kimura A, Mitsukura Y, Oya A, Matsumoto M, Nakamura M, Kanaji A, Miyamoto T: Objective characterization of hip pain levels during walking by combining quantitative electroencephalography with machine learning. Sci Rep 11:3192, 2021. Available from: 10.1038/s41598-021-82696-1

48. Kotsiantis SB: Supervised Machine Learning: A Review of Classification Techniques. Informatica 31:249–68, 2007. Available from

49. LeCun Y, Bengio Y, Hinton G: Deep learning. Nature 521:436–44, 2015. Available from: 10.1038/nature14539

50. Levitt J, Edhi MM, Thorpe R V., Leung JW, Michishita M, Koyama S, Yoshikawa S, Scarfo KA, Carayannopoulos AG, Gu W, Srivastava KH, Clark BA, Esteller R, Borton DA, Jones SR, Saab CY: Pain phenotypes classified by machine learning using electroencephalography features. Neuroimage 223:117256, 2020. Available from: 10.1016/j.neuroimage.2020.117256

51. Li L, Huang G, Lin Q, Liu J, Zhang S, Zhang Z: Magnitude and Temporal Variability of Inter-stimulus EEG Modulate the Linear Relationship Between Laser-Evoked Potentials and Fast-Pain Perception. Front Neurosci 12:, 2018. Available from: 10.3389/fnins.2018.00340

52. Littlewort GC, Bartlett MS, Lee K: Automatic coding of facial expressions displayed during posed and genuine pain. Image Vis Comput 27:1797–803, 2009. Available from: 10.1016/j.imavis.2008.12.010

53. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK: Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med 26:1364–74, 2020. Available from: 10.1038/s41591-020-1034-x

54. Liu X, Faes L, Calvert MJ, Denniston AK: Extension of the CONSORT and SPIRIT statements. Lancet 394:1225, 2019. Available from: 10.1016/S0140-6736(19)31819-7

55. Lötsch J, Ultsch A: Machine learning in pain research. Pain 159:623–30, 2018. Available from: 10.1097/j.pain.0000000000001118

56. Lundervold AS, Lundervold A: An overview of deep learning in medical imaging

focusing on MRI. Z Med Phys 29:102–27, 2019. Available from: 10.1016/j.zemedi.2018.11.002

57. Mackey S, Greely HT, Martucci KT: Neuroimaging-based pain biomarkers: definitions, clinical and research applications, and evaluation frameworks to achieve personalized pain medicine. PAIN Reports 4:e762, 2019. Available from: 10.1097/PR9.0000000000000762

58. Mateen BA, Liley J, Denniston AK, Holmes CC, Vollmer SJ: Improving the quality of machine learning in health applications and clinical research. Nat Mach Intell 2:554–6, 2020. Available from: 10.1038/s42256-020-00239-1

59. van der Miesen MM, Lindquist MA, Wager TD: Neuroimaging-based biomarkers for pain: state of the field and current directions. Pain reports Wolters Kluwer; 4:e751–e751, 2019. Available from: 10.1097/PR9.0000000000000751

60. Misra G, Wang W, Archer DB, Roy A, Coombes SA: Automated classification of pain perception using high-density electroencephalography data. J Neurophysiol 117:786–95, 2017. Available from: 10.1152/jn.00650.2016

61. Moher D, Liberati A, Tetzlaff J, Altman DG: Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6:e1000097, 2009. Available from: 10.1371/journal.pmed.1000097

62. Moons KGM, Wolff RF, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S: PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. Ann Intern Med 170:W1, 2019. Available from: 10.7326/M18-1377

63. Nagendran M, Chen Y, Lovejoy CA, Gordon AC, Komorowski M, Harvey H, Topol EJ, Ioannidis JPA, Collins GS, Maruthappu M: Artificial intelligence versus clinicians:

systematic review of design, reporting standards, and claims of deep learning studies. BMJ :m689, 2020. Available from: 10.1136/bmj.m689

64. Nezam T, Boostani R, Abootalebi V, Rastegar K: A Novel Classification Strategy to Distinguish Five Levels of Pain Using the EEG Signal Features. IEEE Trans Affect Comput 12:131–40, 2021. Available from: 10.1109/TAFFC.2018.2851236

65. Okolo C, Omurtag A: Research : Use of Dry Electroencephalogram and Support Vector for Objective Pain Assessment. Biomed Instrum Technol 52:372–8, 2018. Available from: 10.2345/0899-8205-52.5.372

66. Paul JK, Iype T, R D, Hagiwara Y, Koh JEW, Acharya UR: Characterization of fibromyalgia using sleep EEG signals with nonlinear dynamical features. Comput Biol Med Elsevier Ltd; 111:, 2019.

67. Powers DM: Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. J Mach Learn Technol 2:37–63, 2011.

68. Prichep LS, Shah J, Merkin H, Hiesiger EM: Exploration of the Pathophysiology of Chronic Pain Using Quantitative EEG Source Localization. Clin EEG Neurosci 49:103–13, 2018. Available from: 10.1177/1550059417736444

69. Pryse-Phillips WEM, Dodick DW, Edmeads JG, Gawel MJ, Nelson RF, Allan Purdy R, Robinson G, Stirling D, Worthington I: Guidelines for the diagnosis and management of migraine in clinical practice. Cmaj 156:1273–87, 1997.

70. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M: External validation of prognostic models: what, why, how, when and where? Clin Kidney J 14:49–58, 2021. Available from: 10.1093/ckj/sfaa188

71. Roy SD, Bhowmik MK, Saha P, Ghosh AK: An Approach for Automatic Pain Detection through Facial Expression. Procedia Comput Sci 84:99–106, 2016. Available from:

10.1016/j.procs.2016.04.072

72. Rudin C: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–15, 2019. Available from: 10.1038/s42256-019-0048-x

73. Sai CY, Mokhtar N, Yip HW, Bak LLM, Hasan MS, Arof H, Cumming P, Mat Adenan NA: Objective identification of pain due to uterine contraction during the first stage of labour using continuous EEG signals and SVM. Sādhanā 44:87, 2019. Available from: 10.1007/s12046-019-1058-4

74. Saif MGM, Hassan MA, Vuckovic A: Efficacy evaluation of neurofeedback applied for treatment of central neuropathic pain using machine learning. SN Appl Sci 3:58, 2021. Available from: 10.1007/s42452-020-04035-9

75. Samuel AL: Some Studies in Machine Learning Using the Game of Checkers. IBM J Res Dev 3:210–29, 1959. Available from: 10.1147/rd.33.0210

76. Sarker IH: Machine Learning: Algorithms, Real-World Applications and Research Directions. SN Comput Sci 2:160, 2021. Available from: 10.1007/s42979-021-00592-x

77. Sarnthein J, Stern J, Aufenberg C, Rousson V, Jeanmonod D: Increased EEG power and slowed dominant frequency in patients with neurogenic pain. Brain 129:55–64, 2006. Available from: 10.1093/brain/awh631

78. Schulz E, Zherdin A, Tiemann L, Plant C, Ploner M: Decoding an Individual's Sensitivity to Pain from the Multivariate Analysis of EEG Data. Cereb Cortex 22:1118–23, 2012. Available from: 10.1093/cercor/bhr186

79. Senn S: Disappointing dichotomies. Pharm Stat 2:239–40, 2003. Available from: 10.1002/pst.90

80. Simons LE, Elman I, Borsook D: Psychological processing in chronic pain: A neural

systems approach. Neurosci Biobehav Rev 39:61–78, 2014. Available from: 10.1016/j.neubiorev.2013.12.006

81. Snapinn SM, Jiang Q: Responder analyses and the assessment of a clinically relevant treatment effect. Trials 8:31, 2007. Available from: 10.1186/1745-6215-8-31

82. Sokolova M, Lapalme G: A systematic analysis of performance measures for classification tasks. Inf Process Manag 45:427–37, 2009. Available from: 10.1016/j.ipm.2009.03.002

83. Subasi A, Ahmed A, Aličković E, Rashik Hassan A: Effect of photic stimulation for migraine detection using random forest and discrete wavelet transform. Biomed Signal Process Control 49:231–9, 2019. Available from: 10.1016/j.bspc.2018.12.011

84. Ta Dinh S, Nickel MM, Tiemann L, May ES, Heitmann H, Hohn VD, Edenharter G, Utpadel-Fischler D, Tölle TR, Sauseng P, Gross J, Ploner M: Brain dysfunction in chronic pain patients assessed by resting-state electroencephalography. Pain 160:2751–65, 2019. Available from: 10.1097/j.pain.0000000000001666

85. Tharwat A: Classification assessment methods. New Engl J Entrep 17:168–92, 2020.

86. De Tommaso M, Sciruicchio V, Bellotti R, Guido M, Sasanelli G, Specchio LM, Puca F: Photic driving response in primary headache: diagnostic value tested by discriminant analysis and artificial neural network classifiers. Ital J Neurol Sci 20:23–8, 1999. Available from: 10.1007/s100720050006

87. Tripanpitak K, Viriyavit W, Huang SY, Yu W: Classification of Pain Event Related Potential for Evaluation of Pain Perception Induced by Electrical Stimulation. Sensors 20:1491, 2020. Available from: 10.3390/s20051491

88. Tu Y, Tan A, Bai Y, Sam Hung Y, Zhang Z: Decoding subjective intensity of nociceptive pain from pre-stimulus and post-stimulus brain activities. Front Comput Neurosci 10:,

2016.

89. Uddin S, Khan A, Hossain ME, Moni MA: Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19:281, 2019. Available from: 10.1186/s12911-019-1004-8

90. Vabalas A, Gowen E, Poliakoff E, Casson AJ: Machine learning algorithm validation with a limited sample size. Hernandez-Lemus E, editor. PLoS One 14:e0224365, 2019. Available from: 10.1371/journal.pone.0224365

91. Vanneste S, Song J-J, De Ridder D: Thalamocortical dysrhythmia detected by machine learning. Nat Commun 9:1103, 2018. Available from: 10.1038/s41467-018-02820-0

92. Varrassi G, Müller-Schwefe G, Pergolizzi J, Orónska A, Morlion B, Mavrocordatos P, Margarit C, Mangas C, Jaksch W, Huygen F, Collett B, Berti M, Aldington D, Ahlbeck K: Pharmacological treatment of chronic pain – the need for CHANGE. Curr Med Res Opin 26:1231–45, 2010. Available from: 10.1185/03007991003689175

93. Vatankhah M, Asadpour V, Fazel-Rezai R: Perceptual pain classification using ANFIS adapted RBF kernel support vector machine for therapeutic usage. Appl Soft Comput 13:2537–46, 2013. Available from: 10.1016/j.asoc.2012.11.032

94. Vijayakumar V, Case M, Shirinpour S, He B: Quantifying and Characterizing Tonic Thermal Pain Across Subjects From EEG Data Using Random Forest Models. IEEE Trans Biomed Eng 64:2988–96, 2017.

95. Vollmer S, Mateen BA, Bohner G, Király FJ, Ghani R, Jonsson P, Cumbers S, Jonas A, McAllister KSL, Myles P, Grainger D, Birse M, Branson R, Moons KGM, Collins GS, Ioannidis JPA, Holmes C, Hemingway H: Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. BMJ :l6927, 2020. Available from: 10.1136/bmj.l6927

96. Vu M-AT, Adalı T, Ba D, Buzsáki G, Carlson D, Heller K, Liston C, Rudin C, Sohal VS, Widge AS, Mayberg HS, Sapiro G, Dzirasa K: A Shared Vision for Machine Learning in Neuroscience. J Neurosci 38:1601–7, 2018.

97. Vuckovic A, Gallardo VJF, Jarjees M, Fraser M, Purcell M: Prediction of central neuropathic pain in spinal cord injury based on EEG classifier. Clin Neurophysiol 129:1605–17, 2018. Available from: 10.1016/j.clinph.2018.04.750

98. Wang W, Kiik M, Peek N, Curcin V, Marshall IJ, Rudd AG, Wang Y, Douiri A, Wolfe CD, Bray B: A systematic review of machine learning models for predicting outcomes of stroke with structured data. Beiki O, editor. PLoS One 15:e0234722, 2020. Available from: 10.1371/journal.pone.0234722

99. Wei M, Liao Y, Liu J, Li L, Huang G, Huang J, Li D, Xiao L, Zhang Z: EEG Beta-Band Spectral Entropy Can Predict the Effect of Drug Treatment on Pain in Patients With Herpes Zoster. J Clin Neurophysiol Publish Ah:, 2020. Available from: 10.1097/WNP.0000000000000758

100. Whittington JCR, Bogacz R: Theories of Error Back-Propagation in the Brain. Trends Cogn Sci 23:235–50, 2019. Available from: 10.1016/j.tics.2018.12.005

101. Williamson A, Hoggart B: Pain: a review of three commonly used pain rating scales. J Clin Nurs 14:798–804, 2005. Available from: 10.1111/j.1365-2702.2005.01121.x

102. Willmott C, Matsuura K: Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Clim Res 30:79–82, 2005. Available from: 10.3354/cr030079

103. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S: PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 170:51, 2019. Available from:

10.7326/M18-1376

104. Woo C-W, Chang LJ, Lindquist MA, Wager TD: Building better biomarkers: brain models in translational neuroimaging. Nat Neurosci 20:365–77, 2017.

105. Wydenkeller S, Maurizio S, Dietz V, Halder P: Neuropathic pain in spinal cord injury: significance of clinical and electrophysiological measures. Eur J Neurosci 30:91–9, 2009. Available from: 10.1111/j.1460-9568.2009.06801.x

106. Younger J, McCue R, Mackey S: Pain outcomes: A brief review of instruments and techniques. Curr Pain Headache Rep 13:39–43, 2009. Available from: 10.1007/s11916-009-0009-x

107. Yu M, Sun Y, Zhu B, Zhu L, Lin Y, Tang X, Guo Y, Sun G, Dong M: Diverse frequency band-based convolutional neural networks for tonic cold pain assessment using EEG. Neurocomputing 378:270–82, 2020. Available from: 10.1016/j.neucom.2019.10.023

108. Yu M, Yan H, Han J, Lin Y, Zhu L, Tang X, Sun G, He Y, Guo Y: EEG-based tonic cold pain assessment using extreme learning machine. Intell Data Anal 24:163–82, 2020. Available from: 10.3233/IDA-184388

109. Yusuf M, Atal I, Li J, Smith P, Ravaud P, Fergie M, Callaghan M, Selfe J: Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. BMJ Open 10:e034568, 2020. Available from: 10.1136/bmjopen-2019-034568