UNIVERSITY OF LIVERPOOL

Decoding Experimental Pain Intensity and Observation of Pain Using Electroencephalography and Machine Learning Techniques

Thesis submitted in accordance with the requirements of the University of Liverpool for

the degree of Doctor in Philosophy by Tyler Mari

October 2023

Contents

Declaration

No part of this work was submitted in support of any other applications for degree or qualification at this or any other university or institute of learning.

Acknowledgements

Firstly, I would like to thank my supervisor, Dr Nick Fallon, for their invaluable knowledge, expertise, and support throughout this PhD. I am extremely privileged to have had such an excellent supervisor, who enabled me to succeed despite the many challenges over the last 4 years. I am also very grateful to Dr Christopher Brown and Dr Andrej Stancak for their insight, knowledge, and support. I could not have asked for a better supervisory team. I extend my gratitude to the Liverpool Reviews and implementation group for their guidance and contribution to our systematic review, with a special mention to Dr Rui Duarte for their knowledge and supervision.

Thank you to all the Doctoral Academic Teachers for their support and often, much needed, distraction. I would like to give a special mention to my officemates Silvia Ajao and Kerry Lewis. Thank you for your support, friendship, and guidance. Moreover, I am grateful to all current and past members of the lab for their support and encouragement. I especially appreciate the many hours spent collecting data and proofreading this work.

I am grateful to my family and friends for their unwavering support, with a special thanks to Mum and Dad for their continued encouragement, support, and genuine enthusiasm for my research.

To my partner Alex, thank you for everything that you are.

This is for you.

List of Abbreviations

ACC	Anterior Cingulate Cortex
ALE	Activation-likelihood Estimation
ANFIS	Adaptive Network Fuzzy Inference System
ANN	Artificial Neural Network
AP	Action Potential
AUC	Area Under the Curve
BCA	Balanced Classification Accuracy
BO	Bayesian Optimization
CI	Confidence Intervals
CNN	Convolutional Neural Network
CV	Cross Validation
DA	Discriminant Analysis
DBSCAN	Density-based Spatial Clustering of Applications with Noise
DRG	Dorsal Root Ganglia
DT	Decision Tree
ECG	Electrocardiographic
EDA	Electrodermal Activity
EEG	Electroencephalogram
ELM	Extreme Learning Machine
EMG	Electromyographic
EOG	Electrooculographic
ERD	Event-related Desynchronisation
ERP	Event-related Potentials
ERS	Event-related Synchronisation
ET	Extra Trees
FASTER	Fully Automated Statistical Thresholding for EEG Artefact Rejection
FDR	False Discovery Rate
FFT	Fast Fourier Transform

FN	False Negative
FP	False Positive
FPR	False Positive Rate
GA	Genetic Algorithm
GP	Gaussian Process
GS	Grid Search
HAPPE	Harvard Automated Processing Pipeline for Electroencephalography
HB	HyperBand
HP	Hyperparameter
HPO	Hyperparameter Optimisation
HRV	Heart Rate Variability
IASP	International Association for the Study of Pain
ICA	Independent Component Analysis
ISI	Interstimulus Interval
KNN	K-Nearest Neighbours
LCD	Liquid Crystal Display
LDA	Linear Discriminant Analysis
LPP	Late Positive Potential
LR	Logistic Regression
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MEG	Magnetencephalography
ML	Machine Learning
MLPNN	Multilayer Perceptron Neural Network
MWA	Migraine with Aura
MWoA	Migraine without Aura
NB	Naïve Bayes
NPV	Negative Predictive Value
NRS	Numerical Rating scale
OFC	Orbitofrontal Cortex
OR	Odds Ratio

PCA	Principal Component Analysis
PCS	Pain Catastrophizing Scale
PNP	Paraplegic without Neuropathic Pain
PPV	Positive Predictive Value
PREP	Standardised Early-stage EEG Processing Pipeline
PRISMA	Preferred Reporting Items for Systematic Review and Meta-Analysis
PROBAST	Prediction Model Risk of Bias Assessment Tool
PSD	Power Spectral Density
PSO	Particle Swarm Optimisation
PWP	Paraplegic with Neuropathic ain
RBF	Radial Basis Function
RF	Random Forest
RMSE	Root Mean Square Error
ROB	Risk of Bias
ROC	Receiver Operating Characteristic
SBELM	Sparse Bayesian Extreme Learning Machine
SCS	Spinal Cord Stimulator
SD	Standard Deviation
SE	Standard Error
SFFN	Single-hidden-layer Feed-forward Neural Network
SI	Primary Somatosensory Cortex
SII	Secondary Somatosensory Cortex
SWiM	Synthesis without Meta-analysis
SKOPT	Scikit-optimize
SMAC	Sequential Model-based Algorithm Configuration
SVM	Support Vector Machine
SVR	Support Vector Regression
TMS	Transcranial Magnetic Stimulation
TN	True Negative
TP	True Positive
TPE	Tree-structured Parzen Estimators

TRIPOD	Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis
VAS	Visual Analogue Scale
VPP	Vertex Positive Potential

List of Figures

List of Tables

Abstract

Pain is a complex sensation comprised of biological, psychological, and social components. Accurate pain assessment is imperative for effective pain management in both acute and chronic pain. Self-reported measures are currently the gold standard for pain assessment but are not suitable for numerous populations due to a reliance on social and linguistic skills. Consequently, vulnerable populations such as individuals with dementia, cognitive impairments, and disorders of consciousness often receive sub-optimal pain management due to the challenges associated with assessing their pain. Moreover, research demonstrates that healthcare practitioners typically underestimate patient pain intensity in such scenarios, reducing the likelihood of effective pain management. Therefore, techniques enabling objective pain assessment, which negate the use of self-report and alleviate observational bias, are urgently needed to improve pain management in these populations.

The combination of Electroencephalogram (EEG) and Machine Learning (ML) has demonstrated promise for decoding neural responses to infer an individual's internal state. Previous research suggests that subjective pain intensity can be reasonably predicted by training ML models on EEG activity. However, methodological limitations including small sample sizes and a paucity of recommended practices such as external validation, hinder the interpretability of previous research, which limits the translational clinical potential of the approach. Moreover, to our knowledge, no research has considered decoding neural responses during the observation of visual pain stimuli, which could enhance the understanding of empathic responses, e.g., in a patient-clinician interaction, or medical education.

This thesis aimed to conduct the first external validation paradigms for the prediction of both subjective pain intensity and observation of visual pain stimuli to provide realistic estimates of the potential of ML and EEG for decoding pain-related neural responses, overcoming the limitations of the field. The findings of this thesis demonstrated that subjective pain intensity can be predicted with above-chance levels using features derived from EEG. Features predominantly from frontal, central, and parietal scalp regions in theta, alpha, beta, and gamma frequency bands enabled accurate pain prediction. Specifically, our results demonstrated that subjective pain intensity could be predicted in novel samples with accuracies up to 69%. In addition, our results demonstrated that pain observation could not be reliably decoded using EEG and ML, providing evidence of the current limitations of the approach.

For the first time, the effectiveness of ML and EEG for the prediction of pain intensity and pain observation has been evaluated using gold-standard external validation procedures. Our results suggest that the existing literature has overestimated the potential of the method, with highly promising performance metrics possibly due to methodological issues. Further developments and improvements in methodological rigour are imperative to provide sufficient evidence for the effectiveness of ML and EEG for the prediction of pain intensity. Overall, this thesis provides the most robust estimates of the potential of EEG and ML for pain intensity and pain observation decoding.

Chapter 1:

General Introduction

1.1 Pain Perception, Impact, and Measurement

1.1.1 Pain Perception

Pain is an imperative function of the nervous system that provides information about a potential injury threat or the occurrence of an injury to the body (Julius & Basbaum, 2001; Raja et al., 2020). In 2020, the International Association for the Study of Pain (IASP) redefined pain as "an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" (Raja et al., 2020, p. 1977). The term pain encompasses the subjective experience of the sensation, whilst nociception refers to the encoding and transmission of noxious stimuli from the peripheral nervous system to the central and autonomic nervous systems and not the subjective percept (Q. Chen & Heinricher, 2022; Dubin & Patapoutian, 2010; Garland, 2012; Mischkowski et al., 2018; Sneddon, 2018; W. D. Tracey, 2017).

Three broad categories of pain have been proposed: nociceptive, neuropathic, and nociplastic (Fitzcharles et al., 2021; IASP, 2017; Kosek et al., 2016; Woolf, 2010, 2011). Nociceptive pain occurs due to threatened or actual damage to tissue and is initiated through nociceptor activation (Fitzcharles et al., 2021; IASP, 2017; Kosek et al., 2016; Woolf, 2010, 2011). Nociceptive pain offers a protective function, detecting and minimising the impact of noxious or potentially noxious stimuli (Fitzcharles et al., 2021; Kosek et al., 2016; Woolf, 2010). Neuropathic pain can occur due to a disease or lesion of either the peripheral or central somatosensory nervous systems, resulting in nerve damage (Fitzcharles et al., 2021; IASP, 2017; Kosek et al., 2016). Finally, nociplastic pain arises from alterations in peripheral and central nervous system function, leading to enhanced sensitivity (Fitzcharles et al., 2021; IASP, 2017). Nociplastic pain stems from altered nociceptive processing which does not have clear evidence of threatened or actual tissue damage or evidence of a lesion or disease of the somatosensory system (IASP, 2017). Nociplastic pain differs mechanistically from both nociceptive and neuropathic pain and relates to chronic pain conditions (e.g., fibromyalgia; Fitzcharles et al., 2021; Kosek et al., 2016).

The experience of pain arises through a complex interplay of biological, psychological, and social components (Fillingim, 2017; Garland, 2012; Gatchel et al., 2007). Pain is modulated through both top-down (e.g., cognitive influences) and bottom-up factors (e.g., sensory inputs; Chen & Heinricher, 2022; Hauck et al., 2015; Legrain et al., 2009; I. Tracey & Mantyh, 2007). For example, diverting attention (top-down process) away from nociceptive stimuli has been shown to reduce both subjective pain intensity and pain-related neural responses, whilst increased attention is associated with enhanced subjective pain intensity (Hauck et al., 2015; Legrain et al., 2005, 2009; Wiech et al., 2008). Furthermore, bottom-up processes are exogenous and can be manipulated via alterations to sensory inputs such as stimulus intensity (Hauck et al., 2015; Tiemann et al., 2015; Torta et al., 2017). Due to the complexity and subjectivity of pain, significant variability in pain experience is observed both across and within individuals (Fillingim, 2005, 2017; Nielsen et al., 2009; Quiton & Greenspan, 2008).

1.1.2 Pain Impact

Whilst pain is often advantageous to protect the body from tissue damage, chronic pain is not protective and may result from abnormal nervous system function (Woolf, 2010). Chronic pain is defined as persistent pain exceeding three months in duration (Crofford, 2015; Treede et al., 2015), which persists after removal of the injurious stimulus, or after the tissue has healed (Hylands-White et al., 2017). Approximately 20% of adults globally are affected by chronic pain (Breivik et al., 2006; Dahlhamer et al., 2018), with estimates suggesting that up to 43% of British adults are impacted (Fayaz et al., 2016). Moreover, the estimates suggest that between 10.4% and 14.3% of the UK population suffer from moderately or severely debilitating pain (Fayaz et al., 2016), which demonstrates the profound impact of chronic pain. On an individual level, chronic pain impairs the quality of life, professional prospects, and personal life, leading to an increased risk of psychopathologies and suicidal ideation (Ataoğlu et al., 2013; Demyttenaere et al., 2007; Fishbain et al., 1997; Hadi et al., 2019; Ratcliffe et al., 2008). Whilst it is challenging to quantify the economic impact, estimates in the US from 2010 suggest that the total costs associated with chronic pain exceed that of heart disease, cancer, and diabetes, ranging from \$560 - \$635 Billion (Gaskin & Richard, 2012). To provide effective pain management for individuals with chronic pain, improved pain assessment remains imperative (Breivik et al., 2008; Dansie & Turk, 2013). The assessment of pain in chronic pain conditions is often complex, resulting in underestimated pain and ineffective treatment recommendations (Dansie & Turk, 2013; Zanocchi et al., 2008). Consequently, improving pain assessment through personalised medicine and improved clinical tools may significantly improve pain management and treatment outcomes (Zanocchi et al., 2008).

1.1.3 Pain Measurement

Accurate pain assessment is imperative for effective pain management (Breivik et al., 2008). Self-report approaches are the current gold standard and should be used where possible (Breivik et al., 2008). Rating scales, such as a visual analogue scale (VAS) or numerical rating scale (NRS), are effective for quantifying pain intensity (Breivik et al., 2008; Dansie & Turk, 2013; Melzack & Katz, 2013). Despite being the preferred approach, self-report methods necessitate the capacity to accurately communicate pain, requiring both linguistic and social skills (Hadjistavropoulos et al., 2001; Schiavenato & Craig, 2010). Consequently, these measures are not suitable for populations who cannot accurately communicate their internal state, such as individuals with dementia (Breivik et al., 2008; Hadjistavropoulos et al., 2001; Herr et al., 2011; Kunz et al., 2009), cognitive impairments (Hadjistavropoulos et al., 2001; Herr et al., 2011; Voepel-Lewis et al., 2002), traumatic brain injury (Arbour & Gélinas, 2014), disorders of consciousness (Herr et al., 2011; Schnakers & Zasler, 2007), non-verbal individuals (Herr et al., 2011; D. Li et al., 2008; McGuire et al., 2016), and young children or infants (Hadjistavropoulos et al., 2001; Herr et al., 2011; Witt et al., 2016).

Alternatively, observational methods, which assess non-verbal indicators such as grimacing, can be used to approximate pain (e.g., Hadjistavropoulos et al., 2001; Malviya et al., 2006). However, observational approaches are also imperfect (Hadjistavropoulos et al., 2001). Healthcare professionals often underestimate pain intensity from patient interactions when compared to self-report ratings (Kappesser et al., 2006; Seers et al., 2018), which may lead to ineffective pain management (e.g., undertreatment or overtreatment; Kelley et al., 2008; King & Fraser, 2013). Observational methods are also at risk of bias due to several reasons including economic, individual, and social factors (Atkins & Mukhida, 2022; Hoffman et al., 2016; Pierson et al., 2021). In one study, researchers examined the impact of inaccurate beliefs about biological differences between black and white individuals on pain estimation and treatment (Hoffman et al., 2016). They found that approximately 50% of the sample endorsed inaccurate beliefs and were more likely to underestimate the pain of a black individual (Hoffman et al., 2016). Consequently, improved pain assessment techniques, and understanding of pain observational and evaluation, are also desirable to improve clinical training.

Given the complexity of pain assessment, the advent of objective measures is desirable to improve pain management. However, it has been argued that objective pain assessment is impossible due to the subjectivity of pain (Breivik et al., 2008). Despite the difficulty associated with "objectively" assessing pain, proxy measures for use in populations where current methods fail, would demonstrate significant clinical utility and are a target of considerable research. In particular, biological markers, or biomarkers, may prove effective for pain assessment, with neuroimaging-based methods demonstrating promise (van der Miesen et al., 2019). Therefore, in this thesis, we assess the feasibility of neuroimaging-based pain assessment techniques in healthy individuals. Whilst an objective pain assessment technique is desirable for clinical populations; the current thesis aims to develop a proof of concept for the approach in healthy individuals. Consequently, one of the aims of this thesis is to assess the effectiveness of ML and EEG for pain intensity decoding, which would provide insight into the clinical potential of the measure. Another aim is to consider decoding pain observation processes, which could help to improve clinical training regarding bias and empathy for patient-practitioner interactions.

Throughout the remainder of this introduction, we discuss the biological mechanisms of pain, ranging from peripheral mechanisms of cutaneous nociception to neural correlates. Subsequently, we overview the current approaches for predicting pain intensity from electroencephalogram (EEG) data. In addition, in this thesis, we also assess the predictive capability of neural responses during pain observation. Therefore, we also provide an overview of the empathic processing of pain, discussing the neural mechanisms before describing the existing literature.

1.2 Pain Processing in Healthy Individuals

1.2.1 Peripheral Mechanisms of Cutaneous Nociception

The human body contains three primary classes of neurons: sensory/afferent, motor/efferent, and interneurons (Yam et al., 2018). Cutaneous afferents that innervate the skin are responsible for sensing a range of tactile, thermal, pain, and itch stimuli (Abraira & Ginty, 2013; McGlone & Reilly, 2010). Cutaneous sensory afferents are classed as either Aβ, Aδ, or C-fibres, depending on their axon diameter, myelination, conduction velocity, and cell body sizes (Abraira & Ginty, 2013; Yam et al., 2018). Most Aβ fibres respond to innocuous mechanical stimulation such as touch (Abraira & Ginty, 2013; Julius & Basbaum, 2001). Whereas Aδ and C-fibres are nociceptors responsible for first- and second-pain, respectively (Bishop & Landau, 1958; Julius & Basbaum, 2001; Schaible et al., 2011). Nociceptors are free nerve endings of primary sensory neurons which are responsible for the transduction of numerous environmental stimuli including cold, heat, chemical, and mechanical (Dubin & Patapoutian, 2010; R. Z. Hill & Bautista, 2020; Kandel et al., 2012). Figure 1.1 illustrates the organisation of cutaneous receptors in the skin.

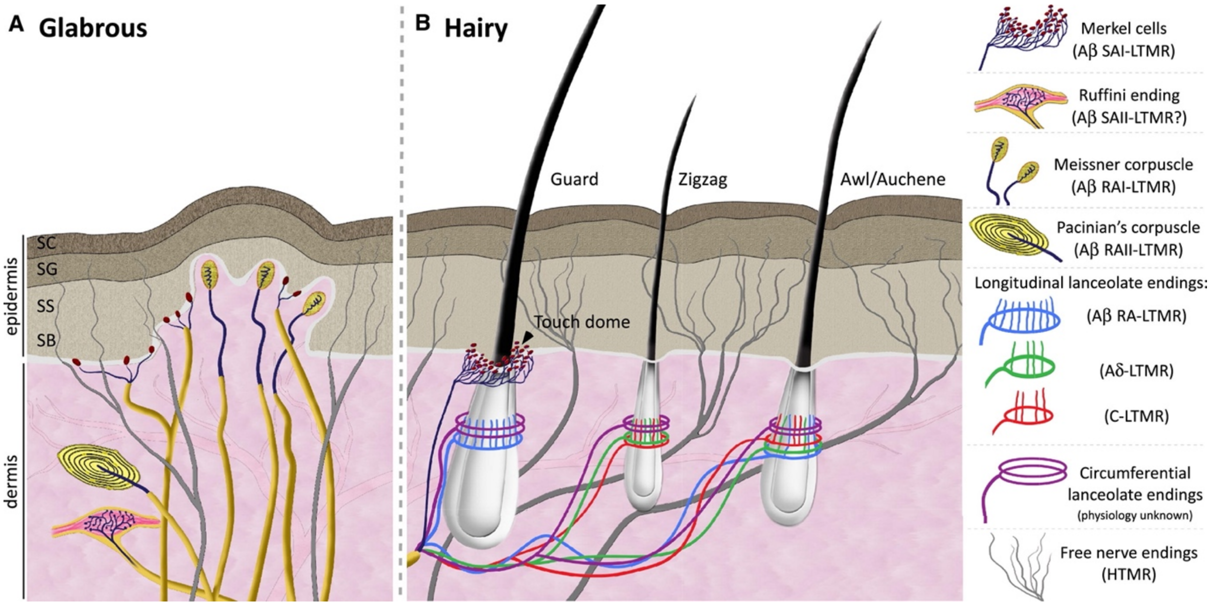

A Glabrous
B Hairy
epidermis
dermis
SC
SG
SS
SB
Guard
Zigzag
Awl/Auchene
Touch dome
Merkel cells (Aβ SAI-LTMR)
Ruffini ending (Aβ SAII-LTMR?)
Meissner corpuscle (Aβ RAI-LTMR)
Pacinian's corpuscle (Aβ RAII-LTMR)
Longitudinal lanceolate endings:
(Aβ RA-LTMR)
(Aδ-LTMR)
(C-LTMR)
Circumferential lanceolate endings (physiology unknown)
Free nerve endings (HTMR)

Figure 1.1 The organisation of cutaneous mechanoreceptors in the skin. Reprinted from Neuron, 79(4), by V.E. Abraira & D.D. Ginty. "The sensory neurons of touch", 618-639, copyright (2013), with permission from Elsevier.

Aδ fibres are lightly myelinated, medium diameter (approx. 2 - 5 µm), with conduction velocities between 5 and 30 m/s (Abraira & Ginty, 2013; Crawford & Caterina, 2020; McGlone & Reilly, 2010; Yam et al., 2018). Aδ fibres have small receptive fields, resulting in well-localised acute pain sensations (Basbaum et al., 2009; Ploner et al., 2002). C-fibres are unmyelinated and thin (< 2 µm in diameter), resulting in slower conduction velocities (< 2 m/s; Abraira & Ginty, 2013; Smith & Lewin, 2009; Yam et al., 2018). C-fibres are responsible for second pain, a duller, longer-lasting sensation that persists long after the injury (Dubin & Patapoutian, 2010). C-fibres have relatively large receptive fields (100 mm^2 in humans; Schmidt et al., 1997) and poor stimuli localisation (Basbaum et al., 2009; Voscopoulos & Lema, 2010; Yam et al., 2018).

Type I nociceptors, located in hairy and Glabrous skin, are polymodal and respond to mechanical and chemical stimulation, but have high heat activation thresholds (> 50^{o}C; Abraira & Ginty, 2013; Basbaum et al., 2009; Djouhri & Lawson, 2004). The heat threshold of type 1 receptors is reduced during sustained heat stimulation or tissue injury (Basbaum et al., 2009; Treede et al., 1998). As these cells have low mechanical and chemical thresholds, they likely account for the first pain evoked by the noxious mechanical stimulation (Basbaum et al., 2009; Granovsky et al., 2005). Type II nociceptors, located in hairy skin, have much lower heat and higher mechanical thresholds compared to type I receptors, which mediates the acute pain response to noxious heat stimulation (Basbaum et al., 2009; Djouhri & Lawson, 2004; Dubin & Patapoutian, 2010). In the present thesis, we deliver mechanical stimulation to the finger-nail bed of the left-hand index finger.

C-fibres account for between 60 to 70% of skin afferents (Basbaum et al., 2009; Lewin & Moshourab, 2004). C-fibres are heterogeneous and polymodal, responding to mechanical, thermal, and chemical stimulation in a slowly adapting manner (Basbaum et al., 2009; Lewin & Moshourab, 2004; Smith & Lewin, 2009; Wooten et al., 2014). However, not all C-fibres are polymodal (Smith & Lewin, 2009). For example, a subset of silent C-fibres is unresponsive to thermal and mechanical stimulation (Handwerker et al., 1991; Schmidt et al., 1995). However, silent C-fibres can become responsive to both heat and mechanical stimulation following sensitisation (Kress et al., 1992). For example, sensitisation can be induced using capsaicin or repeated mechanical or heat stimulation (Banik & Brennan, 2008; Torebjörk et al., 1992). Finally, a class of low threshold C-fibres are responsive to innocuous touch, which may contribute to the encoding of pleasant touch (Löken et al., 2009; Olausson et al., 2007).

1.2.2 Spinal Cord Projections

From initial peripheral processing, nociceptive information is relayed to the spinal cord for further projection (D'Mello & Dickenson, 2008). The cell bodies of primary sensory neurons reside in the dorsal root ganglia (DRG; Abraira & Ginty, 2013; Dubin & Patapoutian, 2010; Smith & Lewin, 2009). The fibres that carry somatosensory information are combined into peripheral nerve fibre bundles as they enter the DRG (Kandel et al., 2012). Consequently, the spinal cord is the first relay station for nociceptive information, with the terminals of primary afferents terminating in the dorsal horn before projecting to higher-order brain regions (Brooks & Tracey, 2005; D'Mello & Dickenson, 2008). DRG neurons have two axonal projections, one to peripheral sites and the other to the central nervous system (Kandel et al., 2012). The axons of second-order neurons cross the midline and ascend before synapsing with a third-order neuron located in the thalamus, which projects to sensory regions of the cerebral cortex (R. S. Snell, 2009).

The spinal cord comprises 31 pairs of spinal nerves, consisting of both white (e.g., axons) and grey matter (e.g., cell bodies, dendrites, glial cells; Diaz & Morales, 2016; Henmar et al., 2020; Purves et al., 2017). Spinal cord grey matter contains several tissue layers which have been divided into 10 laminae (lamina I to X) based on variations in neuronal size and compactness, known as cytoarchitectonics (Diaz & Morales, 2016; Rexed, 1952). Lamina I and II, the nucleus marginalis, and substantia gelatinosa transduce pain and temperature signals (D'Mello & Dickenson, 2008; Diaz & Morales, 2016; Purves et al., 2017). Lamina III and IV process pressure touch and vibration, whilst neurons in lamina V encode stimuli from muscle, cutaneous, joint mechanical, and visceral nociceptors, whilst lamina VI contributes to the flexion reflex

(Basbaum et al., 2009; Diaz & Morales, 2016; Purves et al., 2017). Lamina VII, VIII, IX, and X contribute to proprioception and autonomic functions (Diaz & Morales, 2016; Purves et al., 2017).

The dorsal horn comprises lamina I to VI, with most Aδ and C-fibres terminating in superficial layers (lamina I/II), with some terminating in deeper layers, whilst most Aβ-fibres terminate in laminae III to VI (Basbaum et al., 2009; Craig, 2002; D'Mello & Dickenson, 2008; Diaz & Morales, 2016; Kandel et al., 2012; Todd, 2010). Inputs to lamina V receive both innocuous and noxious input from monosynaptic A-fibre afferents directly, and C-fibres indirectly, which are polysynaptic (Basbaum et al., 2009). Wide dynamic range neurons receive input from several sensory fibres and respond to noxious, visceral, and innocuous mechanical stimuli (Basbaum et al., 2009; D'Mello & Dickenson, 2008). Wide dynamic range neurons exhibit wind-up, a form of synaptic plasticity, which through repeated stimulation increases the magnitude and frequency of the evoked response (D'Mello & Dickenson, 2008; Herrero et al., 2000).

The axons that enter the spinal cord from the dorsal root ganglion progress to the posterior grey column (R. S. Snell, 2009). Here, they bifurcate into both ascending and descending branches that traverse one to two spinal cord segments (relative to the segment of origin), forming the posterolateral tract of Lissauer (R. S. Snell, 2009; Steeds, 2009). The fibres of first-order neurons synapse with cells in the posterior grey column, which includes neurons in the substantia gelatinosa (R. S. Snell, 2009). Subsequently, most second-order neuronal axons cross the anterior grey and white commissures and ascend contralaterally in the white column, forming the spinothalamic tract (Craig, 2002; Kandel et al., 2012; R. S. Snell, 2009).

The spinothalamic tract is located in the anterolateral white matter of the spinal cord and consists of the lateral and anterior spinothalamic tracts, which relay somatosensory information from the dorsal horn to the thalamus and cortex (Craig, 2002; Kandel et al., 2012; Steeds, 2009; Yam et al., 2018). Specifically, the spinothalamic tract ascends through the spinal cord, with collateral branches to the reticular formation of the medulla oblongata, pons and brainstem, which includes the gigantocellularis and paragigantocellularis nuclei and the periaqueductal grey (Almeida et al., 2004; Kandel et al., 2012; R. S. Snell, 2009; Steeds, 2009). Both lateral and anterior tracts ascend alongside each other and form the anterolateral system. Here, the lateral spinothalamic tract relays pain and temperature information, whilst the anterior spinothalamic tract relays touch and firm pressure signals (Kandel et al., 2012; Yam et al., 2018). Aδ and C-fibres ascend via the spinothalamic tract, with Aδ travelling through the neospinothalamic tract and C-fibres through the paleospinothalamic tract (Bussone & Grazzi, 2013; Steeds, 2009). The neospinothalamic tract experiences minimal modulation before reaching the cortex, and is responsible for the sensory-discriminative aspects of pain (Bussone & Grazzi, 2013; Steeds, 2009). Whereas the paleospinothalamic tract receives modulation throughout and contributes to the affective nature of pain (Bussone & Grazzi, 2013; Steeds, 2009). Figure 1.2 illustrates the spinothalamic tract (Betts et al., 2013).

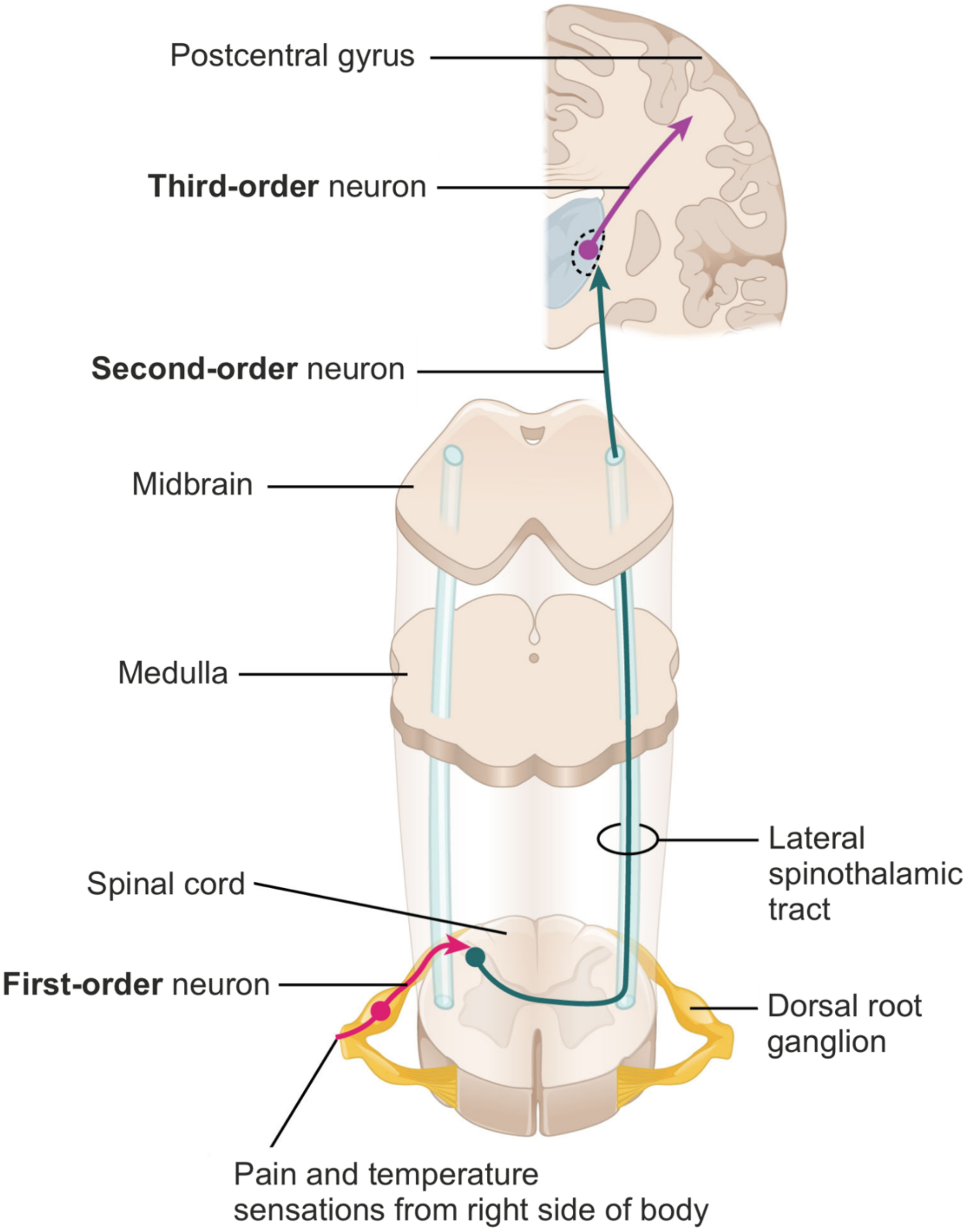


Figure 1.2 Illustration of the spinothalamic tract. Adapted with permission from Betts et al. (2013).

After ascending through the midbrain, many spinothalamic tract fibres synapse with third-order neurons that project to the ventral posterolateral thalamus (Craig, 2002; Kandel et al., 2012; R. S. Snell, 2009). The axons of the spinothalamic tract terminate in either the medial or lateral nuclei (Steeds, 2009; Yen & Lu, 2013). The lateral thalamic nuclei consist of ventral posterior and posterior nuclei, whilst the medial thalamic nuclei comprise intralaminar, dorsal and midline nuclei (Yen & Lu, 2013). Subsequently, axons ascend and traverse the posterior limb of the internal capsule and the corona radiata, reaching the somatosensory area of the postcentral gyrus; located in the cortex (Craig, 2002; Kandel et al., 2012; R. S. Snell, 2009).

Nociceptive information is transmitted to cortical and subcortical regions including the primary (SI) and secondary (SII) somatosensory cortices, insular cortex, anterior cingulate cortex, prefrontal cortex, amygdala, hypothalamus, periaqueductal grey, cerebellum, and basal ganglia (Apkarian et al., 2005; Garland, 2012; Kandel et al., 2012; Steeds, 2009). SI and SII are involved in processing temporal, spatial, and intensity characteristics of pain (Bornhövd et al., 2002; Coghill et al., 1999). Whereas a neural circuit comprised of frontal regions, the periaqueductal grey, and the brainstem contributes to the emotional modulation of pain (Bushnell et al., 2013). The core regions involved in pain processing were originally described using the term pain neuromatrix (Melzack, 1999, 2001), before progressing to the pain matrix (Su et al., 2019; I. Tracey & Mantyh, 2007). More recently, the use of the term pain matrix has declined due to perceived issues regarding the specificity of the observed neural responses (Iannetti & Mouraux, 2010; Mouraux & Iannetti, 2018). Consequently, concepts such as the neurologic pain signature have been proposed to encompass the neural mechanisms of pain (Wager et al., 2013). Figure 1.3 shows the core brain regions involved in pain processing (reprinted from Bushnell et al., 2013).

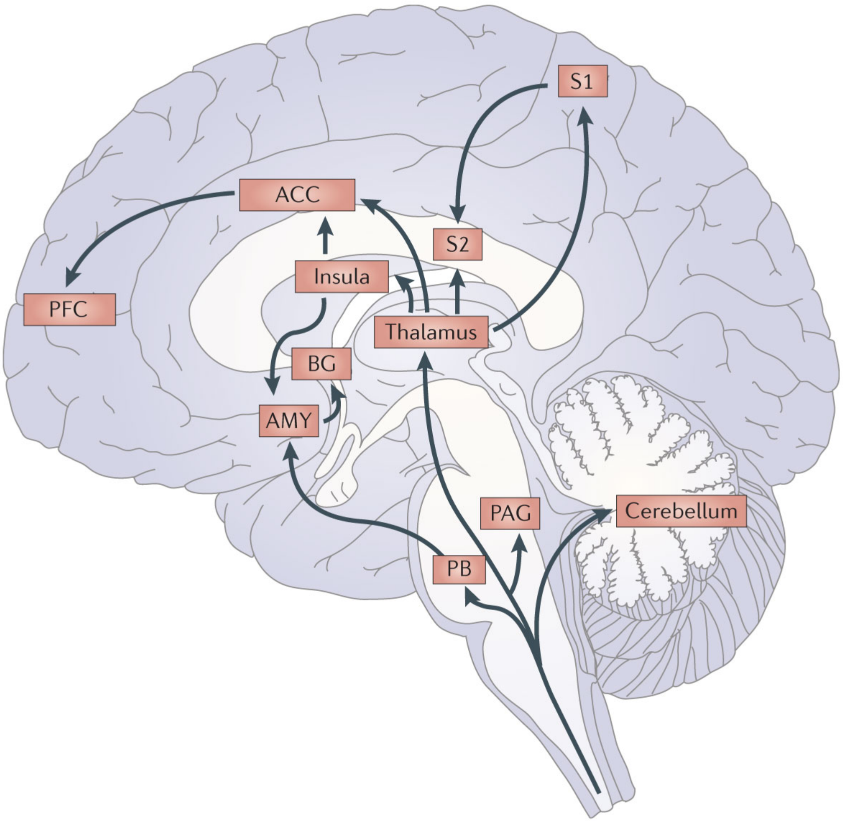


Figure 1.3 Illustration of the brain regions involved in pain processing. Arrows represent projection directions. ACC, anterior cingulate cortex; AMY, amygdala; BG, basal ganglia; PAG, periaqueductal grey; PB, parabrachial nucleus; PFC, prefrontal cortex, SI, primary somatosensory cortex; SII, secondary somatosensory cortex. Reproduced from Cognitive and emotional control of pain and its disruption in chronic pain, M.C. Bushnell, M. Čeko, and L.A. Low, Nature Reviews Neuroscience, 13, Springer Nature, 2013, with permission from SNCSC.

1.3 Neuroimaging of Pain in Humans

Human neuroimaging techniques have allowed researchers to explore neural representations of pain. Pain processing relies on a distributed network of brain regions including the somatosensory cortex, insular cortex, and cingulate cortex (Duerden & Albanese, 2013;

Jensen et al., 2016; Tanasescu et al., 2016; A. Xu et al., 2020). A recent coordinate-based activation-likelihood estimation (ALE) meta-analysis of 222 fMRI experiments identified pain-related activation in bilateral SII, thalamus, brainstem, amygdala, left insula and midcingulate cortex and right middle frontal gyrus (A. Xu et al., 2020). Further ALE meta-analyses of 138 (Jensen et al., 2016), 266 (Tanasescu et al., 2016), and 140 fMRI studies (Duerden & Albanese, 2013) support the importance of the bilateral insula, SII, prefrontal cortex, SI, anterior cingulate cortex, thalamus, and cerebellum in pain processing. In one study, the thalamus, SII, midcingulate cortex, and insula were the most consistently activated regions across experimental paradigms and were reliably activated regardless of stimulation type, location, and gender (A. Xu et al., 2020). Moreover, similar research has demonstrated the importance of the bilateral insular cortices for pain processing, with 66% of studies included in the review reporting pain-related activations (Tanasescu et al., 2016). Moreover, the authors investigated the differences in pain-related activations between healthy controls and individuals with chronic pain, demonstrating that pain-related activations were comparable between the groups, with no significant spatial differences observed during nociceptive processing (Tanasescu et al., 2016). However, increased activity in several clusters (ACC, bilateral insular, right SII, left striatum, right middle frontal gyrus) was shown to be correlated with central sensitisation (Tanasescu et al., 2016).

Xu and colleagues (2020) also investigated whether different experimental stimuli resulted in varied pain-related neural activation. Mechanical stimulation resulted in peak activation magnitudes in the bilateral insular, supramarginal gyrus, thalamus, and right midcingulate cortex. Electrical stimulation paradigms demonstrated peak activations in the bilateral thalamus, right midcingulate cortex, right Rolandic operculum, and left postcentral gyrus.

Contrasts demonstrated that electrical stimulation exhibited larger activation in the right Rolandic operculum, thalamus, and superior temporal gyrus. Moreover, chemical pain stimulation activated the brainstem, bilateral insular, thalamus, left Rolandic operculum, midcingulate cortex, supplementary motor area, and the right postcentral gyrus. Finally, thermal stimulation was associated with activation in the right Rolandic operculum, midcingulate cortex, precentral gyrus, and left cerebellum. Thermal pain stimulation resulted in a larger convergence of activation in the bilateral midcingulate cortex, whilst non-thermal experiments resulted in stronger activation in the right insular and left Rolandic operculum (A. Xu et al., 2020).

A recent mega-analysis of 11 fMRI studies has supported the notion of distributed pain processing by investigating the neural correlates of evoked pain intensity (Petre et al., 2022). Several areas were predictive of evoked pain intensity including regions involved in processing sensory stimuli and nociception including the dorsal posterior insular, ventral posterior medial thalamus, and the periaqueductal grey, areas associated with motor control including the cerebellum, supplementary motor area, and red nucleus, and regions associated with attention, including the lateral prefrontal cortex, frontal operculum, and anterior cingulate cortex. Additionally, the findings demonstrated that the somatomotor, ventral attention, dorsal attention, and visual resting state networks accurately predicted pain intensity. Here, the somatomotor and ventral attention networks predicted pain intensity with reasonable accuracy, with a positive association between predicted pain and true pain intensity being observed. Therefore, this research provides evidence for the existence of a core pain network, which may be important to brain-based decoding of pain.

1.3.1 Electrophysiology of Pain

Spreng and Ichioka ((1964); cited by Apkarian et al., 2005) were the first to investigate evoked potentials elicited by transient painful stimuli. Early studies investigating neural responses during experimental pain stimulation have measured both event-related potentials (ERPs; Carmon et al., 1976; Chatrian et al., 1975) and neural oscillations (Backonja et al., 1991; Ferracuti et al., 1994), providing insight into the electrophysiological correlates of pain experience. For example, one study identified alpha desynchronisation over contralateral parietal regions during noxious cold stimulation (Ferracuti et al., 1994). Subsequently, a plethora of research articles have been published exploring electrophysiological measures of pain (See J. A. Kim & Davis, 2021; Ploner et al., 2017; Zis et al., 2022 for reviews). Throughout the remainder of this section, we provide an overview of the relationship between cortical oscillations and pain which will form the basis of EEG-based pain decoding approaches later in this thesis.

Before discussing the existing literature, it is important to note that much of the research conducting time-frequency analysis of EEG transforms individual frequencies into canonical frequency bands through averaging (Keil et al., 2022; Schomer & Lopes, 2010). Demarcations of such frequencies are usually determined by the speed and power, with traditional frequency bands comprised of delta (< 3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (> 30 Hz; Keil et al., 2022; Schomer & Lopes, 2010). The different frequency bands also differ exponentially in terms of their power, with the slower frequency bands typically exhibiting greater power (Keil et al., 2022; Pritchard, 1992; Schomer & Lopes, 2010).

Existing research has demonstrated that pain elicits changes in neural oscillations in delta, theta, alpha, beta, and gamma bands across the scalp (J. A. Kim & Davis, 2021; Ploner et al., 2017; Zis et al., 2022). Research investigating the association between delta oscillations and pain provides conflicting results (Zis et al., 2022). Several studies have reported augmented delta oscillations associated with pain stimulation (Ferracuti et al., 1994; Giehl et al., 2014; Gram et al., 2015; Huber et al., 2006; Stevens et al., 2000). For example, research has observed increased bilateral frontal delta activity during noxious stimulation (Chang et al., 2001; Ferracuti et al., 1994). Additionally, tonic cold pain stimulation leads to increased delta power when averaged across EEG electrodes, when compared to resting state (Gram et al., 2015). However, several studies have shown no association between delta and painful stimulation (Bunk et al., 2018; Chang et al., 2003; Dowman et al., 2008; Huishi Zhang et al., 2016; Shao et al., 2012). Although, one study demonstrated that delta power was associated with stimulus intensity, but not subjective pain intensity (Bunk et al., 2018). Therefore, delta oscillations may be associated with pain, but the evidence is contentious.

Similarly to delta activity, the role of augmented theta oscillations is contradictory, with research demonstrating increased theta power during experimental pain stimulation (Babiloni et al., 2002; Ferracuti et al., 1994; Michail et al., 2016; Misra, Wang, et al., 2017), whilst other studies report opposite or null findings (Bunk et al., 2018; Chang et al., 2001; Huishi Zhang et al., 2016). For example, Michail and colleagues (2016) observed increased theta amplitudes over central and parietal regions during touch and pain stimulation. Painful stimuli elicited larger increases in theta power than tactile stimulation (Michail et al., 2016). Bunk et al. (2018) report contradictory findings, demonstrating that increased subjective pain intensity during tonic heat stimulation was associated with decreased theta power (Bunk et al., 2018). They also found no association between stimulus intensity and theta activity (Bunk et al., 2018).

Despite conflicting evidence regarding the association between theta oscillations and experimental pain stimulation, augmented theta activity is associated with several chronic pain disorders including neurogenic pain (Sarnthein et al., 2006; Stern et al., 2006), neuropathic pain (Boord et al., 2008; Vuckovic et al., 2014), and fibromyalgia (Fallon et al., 2018). Stern and colleagues (2006) identified enhanced theta power in the insular cortex, anterior cingulate cortex, prefrontal regions, and SI and II in neurogenic pain patients compared to controls. Therefore, augmented theta oscillations are often observed in individuals with chronic pain.

Changes in alpha and beta bands are regularly associated with subjective pain perception during phasic and tonic pain stimulation, with parietal alpha suppression and temporal beta enhancement often observed (Bunk et al., 2018; Chang et al., 2001; Gram et al., 2015; L. Hu et al., 2013; Huishi Zhang et al., 2016; Mouraux et al., 2003; Nickel et al., 2017; Nir et al., 2012; Ploner et al., 2006; Shao et al., 2012). However, altered alpha and beta oscillations occur across scalp regions during pain stimulation, with research demonstrating global changes (Mouraux et al., 2003). Furthermore, research has reported lower global alpha and higher beta power during cold pain stimulation relative to control (Shao et al., 2012). Source analysis revealed decreased alpha activity in bilateral frontal and parietal cortices, and mid to posterior cingulate regions and decreased beta oscillations in bilateral posterior cingulate areas during noxious cold stimulation relative to controls. Moreover, increased beta activity was observed in frontal, parietal, temporal, insular, anterior cingulate, occipital, and parahippocampal regions. Finally, decreased alpha power over posterior parietal-occipital regions and increased beta power across bilateral frontal-temporal areas have been recorded during noxious cold stimulation (Chang et al., 2002).

Recently, peak alpha frequency has been shown to be an effective predictor of subjective pain sensitivity (Furman et al., 2018, 2019, 2020; McLain et al., 2022; Millard et al., 2022). Recent research has demonstrated an association between peak alpha frequency and subjective pain intensity, with slower peak frequency correlated with increased pain during prolonged experimental pain stimulation (Furman et al., 2018). Further research identified that peak alpha frequency was negatively associated with sensitivity to prolonged painful stimulation and could be used to predict high-pain sensitivity individuals in a novel sample (Furman et al., 2020). Finally, resting peak alpha frequency has also been shown to be able to predict musculoskeletal pain intensity (modelled using nerve growth factor injections), with lower peak frequency associated with greater pain (Furman et al., 2019). Consequently, the evidence suggests that pain experience is associated with EEG features, supporting the feasibility of an EEG-based pain decoding tool.

Finally, increased gamma band activity has been observed during experimental pain stimulation and is correlated with both pain and stimulus intensity (Babiloni et al., 2002; Gross et al., 2007; Z. Li et al., 2023; Nickel et al., 2017; Schulz et al., 2015; Zhang et al., 2012), making it a potential candidate for brain-based pain decoding approaches. For example, research has shown that gamma-band oscillations over SI reliably predict subjective pain intensity (Zhang et al., 2012). Interestingly, the authors concluded that the association remained even during reduced stimulus saliency, providing evidence for the involvement of the gamma band in the perception of pain. Moreover, further research has shown that gamma amplitudes were positively associated with both subjective pain intensity and stimulus intensity (Gross et al., 2007). Here, SI gamma oscillations were more closely related to pain intensity compared to stimulus intensity.

In addition to SI, prefrontal regions are also associated with the subjective experience of pain (L. Li et al., 2016; Nickel et al., 2017; Schulz et al., 2015). Research has shown that oscillations over the medial prefrontal cortex were positively correlated with subjective pain ratings (Nickel et al., 2017). Here, gamma oscillations were more correlated with subjective pain intensity than stimulus intensity. Moreover, further research has shown that gamma oscillations predicted pain intensity independent of the stimulus delivery site (Nickel et al., 2017). Schulz et al. (2015) report similar findings, with prefrontal gamma oscillations associated with subjective pain intensity, but not stimulus intensity during noxious tonic heat stimulation (Schulz et al., 2015). A recent review found that phasic stimulation resulted in gamma oscillations over central regions, likely originating from sensorimotor regions, whereas tonic and chronic pain resulted in gamma oscillations over frontal regions (Z. Li et al., 2023). Taken together, gamma oscillations may be directly related to subjective pain intensity. Consequently, gamma features could enable accurate pain prediction using ML.

1.4 The Decoding of Pain-related EEG

Throughout the previous section, we have outlined the observed changes in different frequency bands during pain stimulation. The differences in EEG activity can be used to predict subjective pain intensity, providing a proxy measure. Given that the focus of this thesis is to validate algorithms for decoding neural responses to pain stimulation (inclusive of subjective pain intensity and pain observation), we now provide a brief overview of the research predicting subjective pain intensity using EEG data and ML. In Chapter 3, we conducted a systematic review of the effectiveness of ML and EEG for the classification of pain intensity, pain phenotypes, and response to treatment, which provides a more comprehensive overview of the current state-of-the-art approaches.

Several studies have attempted to predict subjective pain intensity using event-related potentials (Bai et al., 2016; G. Huang et al., 2013; L. Li et al., 2018; Tripanpitak et al., 2020). Huang et al. (2013) aimed to predict pain intensity using single-trial laser-evoked potentials at both binary and continuous levels using a naïve Bayes classifier and linear regression, respectively. They found that the naïve Bayes classifier could accurately classify low and high pain trials with accuracies of above 80% both across and within subjects (G. Huang et al., 2013). Moreover, the linear regression model was able to predict subjective pain intensity on a continuous scale (0-10), achieving a mean absolute error of below 2 for cross and within-subject predictions. Similar research has reported comparable findings to Huang et al. (2013) achieving accurate predictions using ML and event-related potentials (see; Bai et al., 2016; L. Li et al., 2018; Tripanpitak et al., 2020).

Additionally, previous research has used ML and time-frequency transformed EEG data to predict subjective pain intensity (Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; T. Cao et al., 2020; Elsayed et al., 2020; Furman et al., 2018, 2019, 2020; Kimura et al., 2021; Misra, Wang, et al., 2017; Okolo & Omurtag, 2018; Vatankhah et al., 2013; Vijayakumar et al., 2017; M. Yu, Sun, et al., 2020). Misra and colleagues (2017) assessed spectral power changes during low and high-pain thermal stimulation. Their results demonstrated increased theta and gamma power and decreased alpha and beta power over frontal regions, with high pain stimulation resulting in larger theta and gamma band power increases. Additionally, decreased alpha and beta power in both conditions were observed over sensorimotor regions, with a significant reduction in beta during high-pain stimulation, and increased theta power during low pain. Using gamma and theta band activity from the medial prefrontal domain, and lower beta band activity from the sensorimotor regions, the authors used a Gaussian support vector machine (SVM) to classify the data into low and high pain classes, achieving a cross-validation accuracy of 89.58%. Furthermore, the gamma and lower beta bands were most important for classification performance, yielding a classification accuracy of 87.5%.

In addition, Schulz et al. (2012) used ML and EEG to predict a subject's pain sensitivity during painful laser stimulation. The results demonstrated increased theta and gamma oscillations and reduced alpha waves predominantly over central regions. Subsequently, the authors developed a SVM to predict pain sensitivity across subjects, achieving a maximum accuracy of 83% (Schulz et al., 2012). Finally, recent research has developed a convolutional neural network (CNN) for the classification of subjective pain intensity during tonic noxious cold stimulation (M. Yu, Sun, et al., 2020). Using alpha, beta, and gamma oscillations, the authors successfully classified EEG data into no pain, moderate pain, and severe pain, achieving an accuracy of 97.37%. Overall, these findings suggest that ML and EEG may provide an effective tool for pain assessment.

1.4.1 Limitations and Knowledge Gaps

Firstly, most previous research delivers either laser, noxious cold, or heat stimuli. Currently, there is limited research exploring mechanical stimulation (e.g., pressure). Pressure stimulation exhibits greater clinical relevance, resulting in achy, somatic pain that is comparable to muscle soreness and musculoskeletal pain (Birnie et al., 2014). To our knowledge, only one study explored experimental pressure stimulation, EEG, and ML (Okolo & Omurtag, 2018). However, the sample size consisted of only 9 subjects, questioning the generalisability of the findings, especially as the results exhibited significant variability in the individual subjects' classification accuracies (e.g., range > 20% for rest and maximum stimuli classification). Small samples increase variability leading to inflated metrics (Varoquaux, 2018), which reduces confidence in the results. Research has shown that across several domains (e.g., prediction models for psychiatric diagnosis), prediction model accuracy decreases as a function of sample size, that is, larger samples are associated with reduced model performance (Varoquaux, 2018). However, from our systematic review in Chapter 3, we identified that the median sample size for EEG and ML-pain studies was only 24. Given that small samples are susceptible to overfitting, resulting in exaggerated performance (Vabalas et al., 2019), larger samples are required to ascertain improved estimates of ML performance for pain intensity prediction.

Further methodological improvements such as external validation are imperative to achieve clinical translation (Mechelli & Vieira, 2020). Briefly speaking, external validation is the process of evaluating model performance on data independent of the training set (Cabitza et al., 2021; Collins et al., 2015; Lever et al., 2016). The novel data should be obtained from different cohorts, facilities or repositories, or collected at a different location, time, or using a different experimental paradigm (Cabitza et al., 2021; Collins et al., 2015). External validation is imperative, as internal validation methods often fail to control overfitting, leading to inflated, un-generalisable performance (Bleeker et al., 2003; Ramspek et al., 2021; Steyerberg & Harrell, 2016; Vabalas et al., 2019). Additionally, research has shown that ML models often exhibit reduced performance when evaluated on external data (X. Li et al., 2019; Mari et al., 2023; Siontis et al., 2015). In Chapter 4, we show that ML can predict low or high pain intensity with an internal validation accuracy of 73.18%, but performance reduces to 68.32% when tested on a new cohort, and to 60.42% when the experimental stimuli are altered in the new cohort (Mari et al., 2023). Therefore, improved validation procedures are needed to sufficiently evaluate ML performance to prevent a new replication crisis (Hutson, 2018). In Chapter 2, we provide a more comprehensive overview of ML evaluation (e.g., external validation).

In the present thesis, we aimed to conduct the first external validation paradigms of EEG and ML for pain intensity prediction. Here, we prioritised improving methodological rigour by including multistage validation procedures and increasing the overall sample size, to provide the most robust estimates of model performance for pain intensity prediction, to date. In Chapter 4, we aimed to externally validate ML and EEG for the prediction of low and high pain trials using a multistage validation procedure. Here, the models were evaluated on both a novel sample and using novel experimental pain stimuli. Moreover, in Chapter 6, we externally validated ML and EEG for both continuous pain prediction (e.g., 0 – 100) and binary classification (low, high) using both cross-subject and within-subject external validation procedures.

1.5 Empathic Processing of Pain Observation

In this thesis, we also aimed to classify neural response during pain observation, or pain empathy. Empathy is a vital human concept which is challenging to define (Batson, 2009; Cuff et al., 2016; Decety et al., 2012). However, empathy can be considered:

> an emotional response (affective), dependent upon the interaction between trait capacities and state influences. Empathic processes are automatically elicited but are also shaped by top-down control processes. The resulting emotion is similar to one's perception (directly experienced or imagined) and understanding (cognitive empathy) of the stimulus emotion, with recognition that the source of the emotion is not one's own. (Cuff et al., 2016, p.150)

Pain empathy, which refers to the ability to share and resonate with another individual's pain, is imperative for avoiding dangerous scenarios and promoting prosocial behaviour (Decety et al., 2016; Hein et al., 2010; Zhou et al., 2020). Empathy for observed pain serves an aversive function, resulting in negative cognitive or affective states (Fallon et al., 2020). Pain empathy can be elicited using images depicting painful scenarios (e.g., physical injury), or through facial expressions (e.g., grimacing; Coll, 2018; Jauniaux et al., 2019). Moreover, both bottom-up and top-down processes contribute to empathic processing. Prior experience with a specific type of pain leads to increased empathic responses and exemplifies top-down processing, whilst bottom-up cues arise from both verbal and non-verbal components such as facial expressions (Goubert et al., 2005). Consequently, pain empathy in clinical settings can be impacted by several factors, which could influence clinical decision-making. Consequently, objective neural measures of both pain intensity and pain observation could enable improved treatment practice and medical education (Preusche & Lamm, 2016).

1.6 Neuroimaging Investigations of Empathy

Empathic processing results in the activation of numerous brain areas. Fan and colleagues (2011) conducted a meta-analysis of 40 fMRI studies and identified that the dorsal anterior cingulate cortex, anterior midcingulate cortex, supplementary motor area, and bilateral anterior insular were consistently activated during empathy regardless of task and stimuli (Y. Fan et al., 2011). They also observed that the right anterior insular was predominantly involved in affective-perceptual empathy paradigms, whilst the left anterior insular was involved in both affective-perceptual and cognitive-evaluative paradigms (Y. Fan et al., 2011). During cognitive-evaluative types of empathy, the dorsal anterior midcingulate cortex was more commonly activated.

The brain regions involved in empathic processing share significant neural representations with pain empathy. Timmers et al. (2018) investigated the overlap between empathy and pain empathy by conducting a coordinate-based activation likelihood estimation meta-analysis of 128 fMRI studies. They identified a core neural network for empathy (regardless of pain component), which included bilateral anterior insular, bilateral midcingulate cortex, supplementary motor area, SI, inferior parietal lobe, thalamus, amygdala, and brainstem. In addition, their conjunction analysis demonstrated significant overlap in the core brain regions for empathy and pain empathy, with the inferior and superior frontal areas, thalamus, globus pallidus, amygdala, left midcingulate cortex, and left anterior insular being activated during pain empathy tasks (Timmers et al., 2018). Additionally, a recent meta-analysis demonstrated empathy-specific activations in supramarginal, occipitotemporal, and inferior frontal regions which were distinct from pain processing (Fallon et al., 2020). Further meta-analyses support the importance of the insular and cingulate cortex for pain empathy (Fallon et al., 2020; Lamm et al., 2011).

Pain empathy also leads to differences in EEG activity. Numerous research studies have demonstrated that pain observation suppresses mu/alpha and beta oscillations in the sensorimotor system when compared to no-pain conditions (Cheng et al., 2014; Fabi & Leuthold, 2016; Perry et al., 2010; Riečanský et al., 2015; Whitmarsh et al., 2011). Perry et al. (2010) showed participants images depicting hands being stimulated by either a needle (pain) or cotton bud (no pain). Their results demonstrated that pain observation increased suppression of mu/alpha oscillations over frontal-central regions when compared to the neutral condition. Furthermore, similar research further demonstrated sensorimotor alpha suppression during pain observation (Whitmarsh et al., 2011). Using magnetencephalography (MEG), Whitmarsh and colleagues (2011) found greater alpha suppression during pain observation over sensorimotor regions when compared to non-painful images. However, the authors found no significant differences in beta power between the two conditions. Taken together, pain observation elicits observable differences in neural responses, predominantly in the mu/alpha frequency band.

1.6.1 Event-related Potentials in Pain Empathy Research

Differences in event-related potentials (ERPs; See Chapter 2) have also been observed during pain observation. The P3 and late positive potential (LPP) components are associated with pain observation (Coll, 2018). A meta-analysis of 36 studies found that viewing pain images increased P3 amplitudes, with the maximal effect observed at central-parietal electrodes (Coll, 2018). Moreover, a similar effect was identified for the LPP, with pain observation increasing LPP amplitudes over central-parietal regions. Furthermore, some studies report significant differences over frontal regions for the N1 and N2 components (Coll, 2018). However, the meta-analysis demonstrated that the overall effect of pain observation on N1 and N2 amplitudes is non-significant. The authors highlight the heterogeneity of the direction of the effect in the individual studies (e.g., both increased and decreased N1 amplitudes are reported; Coll, 2018). Therefore, the evidence suggests that the observation of pain reliably enhances P3 and LPP amplitudes over central-parietal electrodes.

Fan and Han's (2008) seminal study provided insight into the electrophysiological responses during pain observation. They presented participants with images depicting neutral and pain conditions for both human and cartoon conditions. Subjects were required to perform a pain judgement task, which involved rating the perceived pain intensity of the image, or a counting task to divert attention. They found that pain observation positively shifted early negative components over frontal-central electrodes. Moreover, larger P3 amplitudes during pain observation over central-parietal regions were observed. Moreover, the amplitudes were altered by the attention, with larger amplitudes recorded in the pain judgement task.

Consequently, the enhanced P3 amplitudes over central-parietal regions may reflect pain observation, which could be used to decode pain empathy neural responses.

Further evidence of P3 enhancement during pain observation has been published since the work of Fan and Han (2008). Numerous studies have reported larger P3 amplitudes during pain observation when compared to neutral conditions (Cheng et al., 2012; Cui et al., 2016; Decety et al., 2010; Y. Fan & Han, 2008; Galang et al., 2020; Han et al., 2008; Ibáñez et al., 2011; Liao et al., 2021; Suzuki et al., 2015). Research has demonstrated larger P3 amplitudes during pain observation across electrodes, with maximum amplitudes observed over Pz and Cz, respectively (Decety et al., 2010). Moreover, Suzuki and colleagues (2015) showed participants photographs of human or robot hands in either a neutral or painful condition. They found that the pain condition resulted in larger P3 amplitudes over frontal electrodes in both the ascending and descending aspects of the P3 component (Suzuki et al., 2015).

Moreover, differences in the LPP component have been reported across several studies (C. Chen et al., 2012; Cheng et al., 2012; Cui et al., 2016; Fallon, Li, Chiu, et al., 2015; Y. Fan & Han, 2008; Galang et al., 2020). For example, previous research by our group has demonstrated that images depicting pain resulted in enhanced LPP amplitudes over central-parietal electrodes when compared to situation-matched non-painful images in both healthy participants and a chronic pain population (Fallon, Li, Chiu, et al., 2015). Moreover, research has demonstrated an enhanced LPP over several scalp regions: frontal, central, temporal, parietal, and occipital during pain observation when compared to neutral-matched stimuli (C. Chen et al., 2012). Interestingly, research has demonstrated that medical professionals exhibit reduced ERP (e.g., P3) and behavioural (e.g., pain ratings) responses during pain observation (Decety et al., 2010). Therefore, elucidating the neural mechanisms of empathy has important applications in clinical contexts, such as medical education (Preusche & Lamm, 2016). Overall, ERP components may enable effective classification of pain observation.

1.7 Decoding Neural Responses

1.7.1 Visual Stimuli

To our knowledge, EEG and ML classification of pain empathy has yet to be attempted. Despite this, previous research has classified EEG responses during the observation of discrete image categories (Bagchi & Bathula, 2022; Cudlenco et al., 2020; Ghosh et al., 2021; Kaneshiro et al., 2015; Stewart et al., 2014; Yavandhasani & Ghaderi, 2022; Zheng et al., 2020). Stewart et al. (2014) presented colour photographs of common objects such as lightbulbs and aimed to use EEG data to predict the presence or absence of a visual object during a given EEG segment. They developed a SVM for each of the 7 subjects which could classify the presence or absence of an object with an average accuracy of 87%. For most subjects, EEG components over occipital (visual processing) areas were imperative to classification performance. Furthermore, recent studies suggest that ML and EEG can be combined to decode the observation of discrete image classes. Using neural network classifiers, Zheng et al. (2020) used EEG responses from 6 subjects to classify 40 image classes from the ImageNet database. The results demonstrated that the neural network achieved an accuracy greater than 90% on the classification task. Comparable findings have also been reported for discrete categories of visual stimuli including scenes, objects, humans (including facial expressions), and animals (Bagchi & Bathula, 2022; Cudlenco et al., 2020; Ghosh et al., 2021; Kaneshiro et al., 2015; Yavandhasani & Ghaderi, 2022).

1.7.2 Empathic Stimuli

Despite a paucity of EEG-based empathy prediction models, previous research has shown that empathic responses can be predicted using facial mimicry (Drimalla et al., 2019) and fMRI (Christov-Moore et al., 2020; Vaughn et al., 2018; Zhou et al., 2020). For example, Drimalla et al. (2019) aimed to classify electromyography (EMG) responses during the observation of images depicting either cognitive or emotional empathy. They found that ML models could discriminate the two conditions, achieving accuracies of up to 72% (Drimalla et al., 2019). Furthermore, recent research has classified empathic responses using resting-state fMRI connectivity (Christov-Moore et al., 2020). They found that empathic concern could be predicted using the resting-state connectivity of the sensorimotor network (Christov-Moore et al., 2020).

Empathic neural responses can also accurately predict group allegiance using fMRI and ML (Vaughn et al., 2018). The previous research trained ML models to predict ingroup and outgroup allegiance using neural responses from the empathy network, which consisted of the insular cortex, anterior cingulate cortex (affective), lateral occipital cortex, and the fusiform supramarginal gyrus (sensorimotor), the relief network comprised of the left inferior frontal gyrus, right middle frontal gyrus, right posterior insular, precentral gyrus, precuneus, bilateral posterior superior temporal sulci, and the bilateral angular gyri. Using these regions, the ML model was able to successfully discriminate group allegiance, achieving an accuracy of 72%. The ML model generalisability was then assessed in two further experiments where group allegiance was arbitrarily assigned. The results demonstrated that neural responses could predict group allegiance in arbitrarily assigned groups with accuracies of 64% and 71% for experiments two and three, respectively.

Previous research has used a linear SVM and fMRI to classify pain empathy in response to vicarious pain scene images and facial expressions and matched neural control images (Zhou et al., 2020). The results showed that neural responses during the observation of pain scene images (Figure 1.4A) could accurately discriminate the neutral and pain images, achieving a cross-validation accuracy of 88%. Similar results were found for the facial expression images, with neural activation (Figure 1.4B) accurately discriminating neutral and painful classes with an accuracy of 80%. Additionally, the authors investigated whether the neural responses observed during vicarious pain scene images could classify empathic neural responses elicited through facial expression stimuli and vice versa. The cross-modality prediction demonstrated that neural responses during the observation of scene images could discriminate painful and neutral expressions with an accuracy of 69%, whilst neural responses during the observation of facial expressions could decode neutral and painful scenes with an accuracy of 78% (Zhou et al., 2020). Taken together, the results demonstrate that neural responses recorded during pain empathy tasks can be accurately classified using ML, providing evidence for the feasibility of the empathy classification task in Chapter 5.

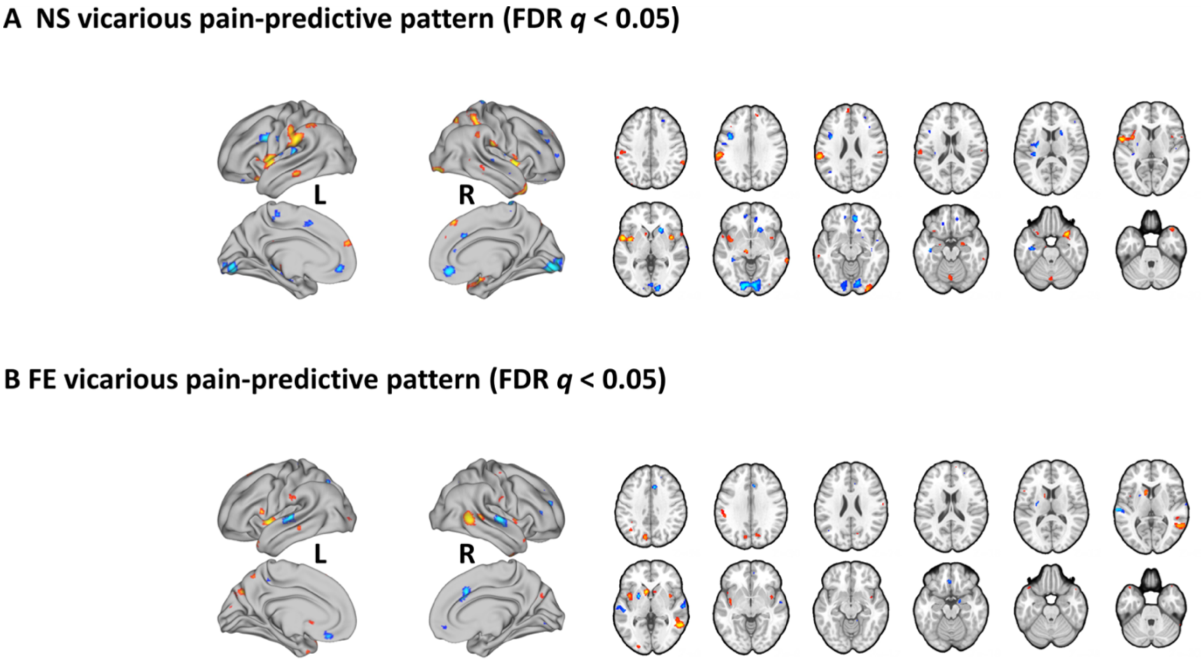

A NS vicarious pain-predictive pattern (FDR q < 0.05)
L
R
B FE vicarious pain-predictive pattern (FDR q < 0.05)
L
R

*Figure 1.4 Neural responses that contribute to the decoding of vicarious pain. Neural activity associated with pain scene images (**A**) and facial expressions (**B**). Adapted from Elife 9, e56929 by F. Zhou et al., "Empathic pain evoked by sensory and emotional-communicative cues share common and process-specific neural representations", copyright (2020) by Zhou et al., Adapted with permission.*

1.7.3 Limitations and Knowledge Gaps

Whilst the previous research provides evidence supporting the notion of empathic response decoding using biological measures such as fMRI and EMG, to our knowledge, no research has attempted to classify pain empathy using EEG features. Empathic processing of pain has clear electrophysiological correlates, particularly alpha suppression, and the enhancement of ERP components such as the P3 or LPP, providing support for the feasibility of decoding EEG responses during pain empathy. We also aim to improve upon the established research for visual imagery decoding. Much of the existing research that decodes neural responses during visual stimuli has small sample sizes, often consisting of less than 10 subjects (e.g., Bagchi &

Bathula, 2022; Cudlenco et al., 2020; Kaneshiro et al., 2015; Stewart et al., 2014; Yavandhasani & Ghaderi, 2022; Zheng et al., 2020). As previously discussed, small samples are at greater risk of overfitting and are associated with inflated and un-generalisable ML performance metrics (Vabalas et al., 2019; Varoquaux, 2018). Therefore, we aimed to conduct the first pain empathy decoding study using EEG and ML, with a specific focus on increasing sample size and thus the generalisability of the models. Moreover, we aimed to externally validate the ML models to provide robust, initial estimates of ML and EEG for decoding pain empathy neural responses.

In this thesis, we aimed to classify EEG responses associated with empathic processing, elicited through the observation of images depicting inflicted pain, painful expressions, or neutral-matched control images. Much of the previous research was conducted in small sample sizes without sufficient external validation procedures. Therefore, we aimed to significantly increase the sample size to provide more robust estimates of model performance. Moreover, we also externally validate the models both across and within subjects. To our knowledge, our study is the first to attempt to classify EEG activity during pain observation.

1.8 Research Problems and Hypotheses

Given the prevalent limitations of research at the intersection of ML and neuroscience, the core aim of this thesis is to robustly develop and evaluate ML models and EEG for the prediction of both subjective pain intensity and pain observation. To achieve this, we aimed to follow recommended standards and improve methodological rigour. For example, we assess model calibration, which is rarely reported but is imperative to the successful development of models with clinical potential (Christodoulou et al., 2019; Mari et al., 2022; Van Calster et al., 2019). Based on the existing literature regarding pain prediction using ML-EEG approaches, it is currently unknown whether model performance generalises to novel samples. Moreover, we aim to investigate whether pain observation can be decoded using EEG and ML, which, to our knowledge, has not yet been attempted. Consequently, this thesis aims to evaluate the effectiveness of ML for decoding pain-related neural responses and to provide evidence for the potential for the method to progress towards practical applications.

This thesis aimed to evaluate the effectiveness of ML for decoding pain-related neural responses. Firstly, we aimed to systematically review the existing literature that implements ML and EEG for the prediction of subjective pain intensity, pain phenotypes, and response to treatment. Here, we aimed to identify knowledge gaps and areas for development. Subsequently, we aimed to externally validate ML and EEG for binary pain intensity prediction in response to experimental pain stimuli delivered using a mechanical pressure stimulator. Furthermore, we also aimed to develop and externally validate ML and EEG for the decoding of pain empathy during passive viewing of visual pain stimuli. Finally, we aimed to increase the predictive capability of ML and EEG by predicting continuous subjective pain intensity during graded levels of mechanical stimulation.

1.8.1 Thesis Hypotheses

H1) The combination of EEG and ML will be able to classify low and high pain levels with above-chance performance (>50%) on external validation.

H2) ML and EEG will successfully classify the observation of pain, compared to neutral stimuli, with accuracies greater than chance levels.

H3) ML models will predict subjective pain intensity (0 – 100) in novel samples more accurately (lower error) than simple heuristic models.

1.8.2 Thesis Chapters

Chapter 2 provides an overview of the methods used in this thesis. Specifically, this chapter describes the principles of EEG, covering the physiological mechanisms, acquisition, pre-processing, and analysis, whilst also highlighting the strengths and limitations of the method. Moreover, this chapter also introduces ML and the underpinning principles. We provide an overview of supervised learning, before progressing to model development (e.g., feature selection, hyperparameter optimisation) and evaluation (e.g., internal validation, model discrimination and calibration). Finally in this section, we introduce systematic review methodologies.

Chapter 3 describes a systematic review of research that has investigated the use of ML and EEG for the prediction of pain intensity, pain phenotypes, or response to treatment. Here, studies applying ML to EEG data to predict pain-related outcomes were reviewed and summarised. Moreover, we conducted reporting standards and risk of bias assessments to evaluate the current state of the field and to identify knowledge gaps and areas of development. The review aimed to identify the effectiveness of ML for predicting pain-related outcomes from EEG and to critically appraise the previous research.

Chapter 4 explored the effect of differing levels of experimental pain stimuli (low, high) on single-trial effects on cortical oscillations (H1). Two experimental studies were conducted, with study one being used for model development and study two for external validation. The second experimental study consisted of new subjects and alternative experimental pain stimuli. In both studies, a custom pneumatic pressure stimulator was used to deliver differing levels of pain intensity, whilst EEG was used to record changes in cortical oscillations during the stimulation. Subsequently, seven popular ML models were trained on single-trial EEG to classify data into either low or high-intensity trials.

Chapter 5 investigates the neural correlates of pain empathy and develops ML models to classify the observation of neutral or pain images (H2). Again, we aimed to externally validate the ML models by recruiting three different samples. Here, the model was evaluated for both cross and within-subject predictions. During the experimental paradigm, participants were shown either neutral or pain expressions or scenes, whilst EEG was recorded. Features calculated from single-trial ERP waveforms were used to train a Random Forest (RF) model to predict the observation of the different classes. Here, three classifications (face – scene, scenes: neutral – pain, faces: neutral – pain) were attempted.

Chapter 6 expands on the work of Chapter 4 by attempting to predict continuous subjective pain intensity using ML and EEG (H3). Here, 10 levels of stimulus intensity, ranging from light touch to moderately-strongly painful, were delivered using the pneumatic pressure stimulator, whilst EEG responses were recorded. Three samples were recruited to perform both cross-subject and within-subject external validation. A RF model and neural network were developed for pain intensity prediction and compared to simple heuristic models.

Moreover, we aimed to replicate our findings from Chapter 4, by externally validating ML and EEG for the classification of low and high pain trials.

Chapter 7 provides a general discussion of the results from the experimental studies of this thesis. Here, we identify and discuss the core themes and implications of this thesis. Moreover, we provide several recommendations for future research which can help to advance the field towards clinically meaningful results and offer support and suggestions for the development of translational tools.

Chapter 2:

General Methods

2.1 Principles of EEG

2.1.1 Physiological Mechanisms of EEG

The human brain consists of approximately 86 billion neurons, which communicate through a combination of chemical and electrical activity (e.g., action potentials; APs; Azevedo et al., 2009; Kandel et al., 2012; Lovinger, 2008; Stuart et al., 1997). The biophysical mechanisms of APs have been well-established since Hodgkin and Huxley's investigation of the giant squid axon (Hodgkin & Huxley, 1952; Schwiening, 2012). APs are discrete voltage spikes generated in axon cell bodies, which propagate through the axon, reaching inhibitory or excitatory synapses (Kandel et al., 2012; Luck, 2014; Yam et al., 2018). APs are transient events, with a duration of approximately one to two milliseconds (usually < 10ms), and occur with a limited potential (Kandel et al., 2012; Luck, 2014).

Selective permeability of the cell membrane to specific cations enables changes in membrane potential and APs (Kandel et al., 2012; F. H. Yu & Catterall, 2003). Incoming signals initiate depolarisation via an influx and efflux of sodium and potassium ions through voltage-gated ion channels (Kandel et al., 2012; Kirschstein & Köhling, 2009; Tivadar & Murray, 2019; F. H. Yu & Catterall, 2003). Following sufficient depolarisation, an AP is generated and propagates along the axon by depolarising the adjacent membrane (Barnett & Larkman, 2007; Tivadar & Murray, 2019). Subsequently, voltage-gated sodium and potassium channels are automatically inactivated and activated respectively, leading to repolarisation (Kandel et al., 2012; Kirschstein & Köhling, 2009).

APs are usually not synchronous, and any electrical field is cancelled out due to the biphasic properties of APs, meaning that they are not detectable at the scalp (Buzsáki et al., 2012; Jackson & Bolger, 2014; Luck, 2014). Therefore, EEG does not directly measure APs. Rather, postsynaptic potentials, which occur for tens of milliseconds (typically between 15-20ms), enable synchronous activity, resulting in potential changes that are observable at the scalp due to temporal overlap and signal summation (Buzsáki et al., 2012; Kirschstein & Köhling, 2009; Luck, 2014; Olejniczak, 2006).

Both excitatory and inhibitory postsynaptic potentials contribute to the EEG signal (Olejniczak, 2006). Excitatory currents (Na^+ and Ca^{2+}), known as passive return currents, flow from intracellular to extracellular space, whilst inhibitory currents (Cl^+ and K^+) flow in opposite directions (Olejniczak, 2006). Consequently, scalp electrodes record the potential differences arising from the postsynaptic potentials (Olejniczak, 2006). Volume conduction enables observable EEG, as the current can flow through biological tissue between the source and an electrode (Olejniczak, 2006).

The spatial organisation and morphology of cortical neurons is essential for the summation of postsynaptic potentials (Jackson & Bolger, 2014). Pyramidal neurons located in layers III, V, and VI are the main excitatory and most numerous cortical cells which have the optimal configuration for signal summation (DeFelipe & Fariñas, 1992; Elston, 2011; Spruston, 2008). Pyramidal cells are often oriented perpendicular to the scalp, contributing to the generation of an open field (Jackson & Bolger, 2014; Luck, 2014; Nunez & Srinivasan, 2006; Olejniczak, 2006). Pyramidal cells consist of several basal dendrites and one apical dendrite, with the apical dendrite organised perpendicular to the cortical surface (Luck, 2014). The release of an excitatory neurotransmitter at the apical dendrite initiates the flow of positively charged ions into the cell, resulting in a negative charge in the extracellular space and positive polarity (an inhibitory current will have the opposite effect; Luck, 2014). This process results in the formation of a dipole. Through the summation of numerous dipoles, electrical activity is observable at scalp electrodes (Luck, 2014). A single EEG electrode records the synchronised synaptic activity of over 1 million cortical synapses arranged over a small surface area (Nunez & Srinivasan, 2006).

2.1.2 EEG Acquisition

EEG acquisition relies on metal electrodes placed across the surface of the scalp which are connected to an amplifier (Górecka & Makiewicz, 2019; Teplan, 2002). Electrodes' contacts on the scalp have electrical impedance, which is usually measured in kΩ, and must be monitored and maintained below a given level (e.g., <10 kΩ; Górecka & Makiewicz, 2019). To reduce electrical impedance, a conductive liquid such as a gel or saline solution is applied to the electrode site (Tallgren et al., 2005).

EEG electrodes are positioned according to derivatives of the international 10-20 system. The 10-20 system places electrodes at 10% and 20% points along both latitude and longitude planes of the head using standard anatomical landmarks including the inion, nasion and the left and right preauricular points (Jasper, 1958; Klem et al., 1999). Extensions of the 10-20 system exist (e.g., dense electrode arrays). Throughout this thesis, we utilised a Geodesic EEG system (Electrical Geodesic Inc., EGI, now Magstim EGI, Eugene, Oregon, USA), which consists of 129 electrodes that are equidistant from each other. Although the Geodesic layout differs from the 10-20 international system, many of the electrode locations have a direct correspondence to the International 10-20 system (see Figure 2.1 for a schematic of the electrode array; Luu & Ferree, 2005).

In healthy adults, the raw EEG amplitude is usually < 100 µV and is amplified by a factor of 1000-50000 known as the gain, before digitisation, for data visualisation and analysis (Luck, 2014). To construct the signal for a single EEG channel, three different electrode types are required, namely active, reference and ground electrodes (Luck, 2014). The EEG signal recorded represents the potential for current to move between the active and ground electrodes (Luck, 2014). EEG systems also use a differential amplifier, which utilises a reference electrode to subtract electrical noise from the true signal, due to potential electrical interference from the ground circuit (Luck, 2014). The subtraction is computed as the difference between the reference and ground electrodes and the active and ground electrodes (Luck, 2014). As the electrical noise is equivalent in both electrode pairs, the interference will be removed during subtraction (Lei & Liao, 2017; Luck, 2014).

Theoretically, reference systems should have either constant or zero potential (Lei & Liao, 2017; Yao et al., 2019). However, no electrode sites conform to this assumption (Lei & Liao, 2017; Yao et al., 2019). Average referencing overcomes this limitation. The average reference is based on the assumption that the integral of potentials across the scalp is zero, making it an ideal reference system (Bertrand et al., 1985; Yao, 2017). As the coverage of the scalp increases (e.g., dense electrode arrays), the average potential across all electrode sites tends to zero, providing a suitable reference signal (Lei & Liao, 2017; Yao et al., 2019). Throughout this thesis we use the average reference approach.

In the present thesis, we conduct several EEG studies using a 129-channel sponge-based geodesic sensor net (Electrical Geodesic Inc., EGI, now Magstim EGI, Eugene, Oregon, USA). A schematic of the net array is provided in Figure 2.1, demonstrating both the electrode layout and the equivalent 10-20 system electrodes' locations. The geodesic sensor net covers the entire head and suborbital areas of the face. Due to the composition of the system, the net application is quick, relying on a saline solution to minimise electrical impedances. The positioning of the net was aligned to three anatomical points including two preauricular points and the nasion. In all EEG studies, a sampling rate of 1000 Hz was used, with a recording band-pass filter set at 0.001 – 200 Hz. Finally, electrode Cz was used as the reference electrode, whilst the COM electrode was used as the ground.

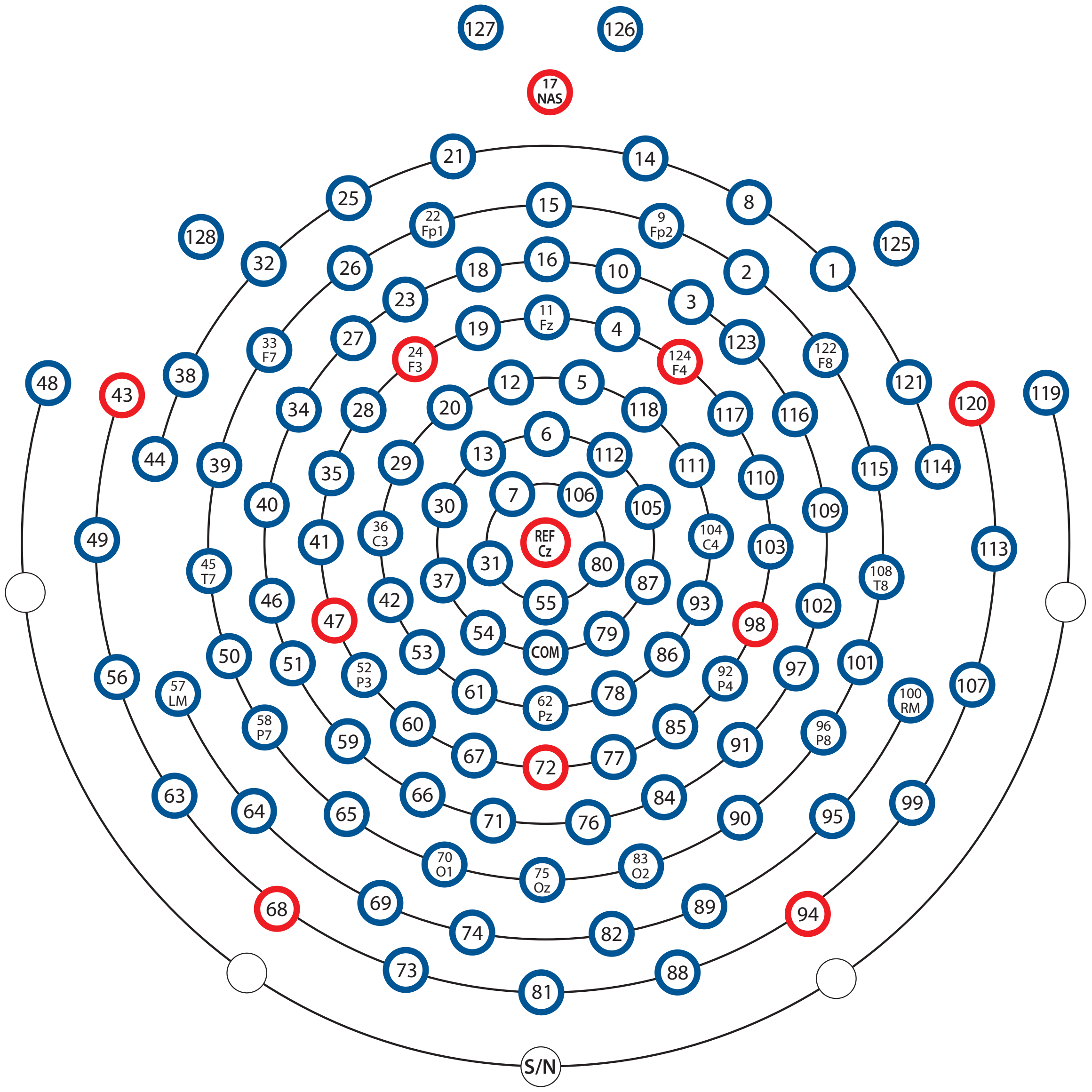


Figure 2.1 Schematic diagram of the distribution of the Geodesic sensor net.

2.1.3 Artefact Correction

The EEG signal is highly susceptible to artefacts (Keil et al., 2022). Artefacts can be segmented into two components: physiological and non-physiological (Gabard-Durnam et al., 2018; Luck, 2014; Teplan, 2002). Physiological artefacts are associated with biological processes and include electrocardiographic (ECG), electrooculographic (EOG), electromyographic (EMG), and electrodermal activity (EDA; Keil et al., 2022; Luck, 2014; Teplan, 2002). Whereas non-physiological artefacts result from hardware issues and electrical interference (Keil et al., 2022). EEG suffers from electrical artefacts arising from alternating mains power supply which introduces cyclic interference at 50 Hz in Europe and 60 Hz in the US, referred to as line noise (Luck, 2014; Teplan, 2002). Finally, EEG is susceptible to other hardware issues, such as fluctuations in impedances or cable and connector issues which cause interference and artefacts (Teplan, 2002).

The simplest artefact correction approach is trial rejection after manual inspection. Here, researchers visually inspect the data for artefacts, identifying contaminated trials, which are marked for rejection. However, manual inspection is laborious, time-consuming, and open to subjectivity, hindering reproducibility (Gabard-Durnam et al., 2018). Additionally, manual rejection is not feasible for large EEG datasets. Consequently, both automated and semi-automated approaches have been developed. Recurring artefacts (e.g., EOG) can be corrected using principal component analysis (PCA). One implementation of PCA-based methods is an adaptive artefact-correcting algorithm implemented in the Brain Electrical Source Analysis Software (BESA; MEGIS GmbH, Germany; Berg & Scherg, 1994; Ille et al., 2002). The method implements a spatial filter technique which decomposes the EEG signal into either brain activity or artefact (Berg & Scherg, 1994; Ille et al., 2002). Given that eye components and genuine neural data are overlapping processes, they can be separated by their topographies (Berg & Scherg, 1994). Example artefact topographies are calculated using PCA and the artefact is removed by projecting out the artefact component’s waveform from the data (Berg & Scherg, 1994).

2.1.4 Automated EEG Pre-processing

Recently, automatic pre-processing pipelines are gaining popularity as they afford the consistent application of the artefact rejection criteria, easy application to large samples, and facilitate comparison and collaboration across research labs (Gabard-Durnam et al., 2018). Several programs have been developed including the Harvard automated processing pipeline for electroencephalography (HAPPE; Gabard-Durnam et al., 2018), the standardised early-stage EEG processing pipeline (PREP; Bigdely-Shamlo et al., 2015), and fully automated statistical thresholding for EEG artefact rejection (FASTER; Nolan et al., 2010) to name but a few. In this thesis, HAPPE was used to pre-process the data of Chapters 5 and 6. Consequently, in this section, we outline the HAPPE pipeline.

HAPPE is an automated EEG pre-processing software written in MATLAB, which utilises EEGLAB functions (Delorme & Makeig, 2004; Gabard-Durnam et al., 2018). The software can pre-process both resting state and task-related EEG for low (≤ 32 Channels) and high-density (> 32 Channels) systems (Gabard-Durnam et al., 2018). Filtering, artefact rejection, and re-referencing to prepare the data for time-frequency analysis can be conducted using HAPPE (Gabard-Durnam et al., 2018). Recently, HAPPE has been extended for ERP studies (Monachino et al., 2022).

Low-pass, high-pass, and line noise filtering can be conducted using user-specified parameters within HAPPE (Gabard-Durnam et al., 2018). Line noise removal is achieved using CleanLine (Mullen, 2012), which utilises a multi-taper regression to remove line noise around (± 2 Hz) the user-specified frequency (Gabard-Durnam et al., 2018). The data can be subsequently downsampled to 250, 500, or 1000 Hz. For ERP studies, HAPPE allows the user to select filter cut-offs using either a Hamming windowed sinc FIR filter or an IIR Butterworth filter (Monachino et al., 2022).

Bad channel detection can also be performed. Here, the software locates bad channels that exceed three standard deviations from the mean of the normed joint probability of the average log power between 1 and 125 Hz (Gabard-Durnam et al., 2018; Monachino et al., 2022). For ERP analysis, further steps are conducted including detecting flatline channels (duration > 5s), rejecting channels > 3 or <-5 standard deviations from the mean power, rejecting channels with remaining line noise contamination (> 6 standard deviations from the line noise mean), and rejecting channels with outliers based on their correlation with other channels. Correlation coefficients less than .8 lead to channel rejection (Monachino et al., 2022).

To minimise trial rejection, HAPPE conducts artefact correction using either wavelet thresholding or independent component analysis (ICA; Gabard-Durnam et al., 2018; Monachino et al., 2022). Wavelet thresholding detects artefacts using both time- and frequency-localisation and removes it without distorting the brain signal (Gabard-Durnam et al., 2018; Monachino et al., 2022). Wavelet thresholding is more computationally efficient than ICA, completing within the order of seconds (Gabard-Durnam et al., 2018; Monachino et al., 2022). In wavelet thresholding for ERPs, soft or hard margins, which determine the stringency of the correction depending on the data type (e.g., soft margin for adult samples; Monachino et al., 2022) are applied. Alternatively, ICA can be applied, which decomposes the data into independent components reflecting artefact and neural data (Monachino et al., 2022). We use wavelet thresholding in Chapters 5 and 6 due to its effectiveness and computational efficiency.

Subsequently, data segmentation is conducted by defining a time window relative to a trigger. If ERP analysis is required, baseline correction can be applied. Furthermore, segments containing bad data are interpolated, with channels containing artefacts being interpolated using the FASTER software (See Nolan et al., 2010). Segment rejection is then performed by defining an amplitude threshold (e.g., -150 to 150 for adults), and/or through joint probability rejection, which identifies artefacts such as muscle movements (Monachino et al., 2022). Two criteria are used for segment rejection. Firstly, single electrode probability is measured by computing the joint probability of a channel's activity at a given segment relative to the activity of the same channel across all other segments. Secondly, electrode group probability is calculated for each section by computing the joint probability between activity at a given channel relative to the activity of all other electrodes in the same segment (Monachino et al., 2022). Segments exceeding three standard deviations from the mean on either criterion are rejected. Following segment rejection, bad channels are interpolated using spherical interpolation (Gabard-Durnam et al., 2018; Monachino et al., 2022). The data can then be re-referenced by performing average re-referencing or using a subset of channels.

2.1.5 Time-Frequency Analysis

Electrophysiological recordings demonstrate neural rhythmicity, characterised as brain oscillations (Keil et al., 2022). Neural oscillations demonstrate a periodic pattern of activity from synchronous neuronal populations (Kirschfeld, 2005). The rhythmicity of EEG was first observed during EEG recordings conducted in the 1920s by Hans Berger, one of the inventors of EEG (J. A. Kim & Davis, 2021; Prerau et al., 2017). Berger initially characterised waves of the first order, or alpha rhythms (İnce et al., 2021), which were oscillations at 10Hz observed near the occipital cortex in awake subjects with eyes closed (Adrian & Matthews, 1934; J. A. Kim & Davis, 2021; Kropotov, 2009). Berger also noted that alpha waves were attenuated once the subject opened their eyes, which were defined as waves of the second order or beta waves, which were faster and had a lower amplitude (J. A. Kim & Davis, 2021; Kropotov, 2009).

Cortical oscillations can be quantified using metrics including frequency, time, amplitude, phase, and morphology, following spectral decomposition. Spectral decomposition, or estimation, aims to separate a waveform into distinct component oscillations, which are linked by the component frequency (Prerau et al., 2017). Much of the research performing EEG time-frequency analysis relies on averaging spectral power across frequencies to summarise the activity in a frequency band. The spectrum can be divided into canonical bands such as: delta (< 3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz) and gamma (> 30 Hz; Keil et al., 2022; Schomer & Lopes, 2010). However, minor variations in the definitions of the frequency bands exist. In this thesis, namely Chapters 4 and 6 we use the following definitions: theta (4 – 7 Hz), alpha (8 – 12 Hz), lower beta (16 – 24 Hz), upper beta (25 – 32 Hz) and gamma (33 – 70 Hz).

2.1.5.1 Discrete, Fast, and Short-Time Fourier Transformations

The Fourier transform is imperative for signal processing and is the foundation of EEG data analysis (Cohen, 2014). The theoretical underpinning of Fourier analysis is that any signal can be represented by a combination of sine waves, provided each wave has a distinct frequency, amplitude, and phase. That is, any signal can be represented by a linear combination of trigonometric functions that have different frequencies and amplitudes (Prerau et al., 2017; Sawa et al., 2022). Consequently, Fourier analysis decomposes a time-domain signal by calculating the dot product, the sum of two vectors after elementwise multiplication, between the EEG time series and sine waves with different frequencies (Cohen, 2014). The Fourier transformation is theoretically lossless, providing a perfect representation of the original signal.

Several implementations of the Fourier Transform have been developed including the Discrete Fourier Transform, Fast Fourier Transform (FFT) and the Short-Time Fast Fourier Transform. The Discrete Fourier Transform involves constructing a sine wave of equal length to the signal for a given frequency and computing the dot product. This process is repeated for *n* sine waves of varying frequency, where *n* is the number of time points in the original signal. Moreover, the Fast Fourier transform refers to a family of more efficient algorithms (e.g., Cooley-Turkey algorithm; Cooley & Tukey, 1965), which are several orders of magnitude faster than the Discrete Fourier transform (Cohen, 2014).

Whilst Fourier analysis provides the average power across the signal, the Short-Time Fast Fourier Transform segments the signal, iteratively calculating the spectral characteristics using a sliding time window (Cohen, 2014; Subha et al., 2010). First the segment is tapered to attenuate the edges (due to edge spectral distortions referred to as edge artefacts), reduce spectral leakage (smearing of power due to non-periodicity in the window), and control frequency smoothing (Cohen, 2014; Kropotov, 2009). The Fourier transform is then applied to the tapered signal. The sliding window is then shifted by *t* time points and the process is repeated (Cohen, 2014). Common types of window functions include Boxcar, Hann/Hanning, Hamming, and Blackman windows (Sawa et al., 2022).

2.1.5.2 The Multi-taper Method

An alternative approach to time-frequency decomposition is the multi-taper method. The multi-taper method calculates an average across several tapers to extract different frequency components (Prerau et al., 2017). Tapers are discrete prolate spheroidal sequences, known as Slepian sequences (Cohen, 2014; Slepian & Pollak, 1961). The sequences are orthogonal, meaning that they extract different frequency spectra properties (Cohen, 2014; Keil et al., 2022; Prerau et al., 2017). The multi-taper method first extracts a segment, which is multiplied with several tapers simultaneously (the number of tapers dictates the smoothness of the resulting spectra), which results in a tapered time series (Cohen, 2014; Prerau et al., 2017). After obtaining the set of tapered time series, the FFT of each series is calculated and the resulting spectra are averaged together (Cohen, 2014; Keil et al., 2022; Prerau et al., 2017). A schematic of the multi-taper method from Prerau et al. (2017) is presented in Figure 2.2**.**

The multi-taper method has several advantages and limitations when compared to other approaches. Firstly, the multi-taper method is advantageous in noisy data and small trial samples, as the method limits the impact of noise (Cohen, 2014). Secondly, the multi-taper method is appropriate for single-trial analyses and is preferred for the analysis of frequencies greater than 30 Hz (Cohen, 2014). For example, at gamma-band frequencies (>60 Hz), the multi-taper method improves the signal-to-noise ratio due to frequency smoothing (Cohen, 2014). Furthermore, the multi-taper method reduces edge artefacts, as information lost through certain tapers can be captured using alternative tapers (Babadi & Brown, 2014; S. E. Kim et al., 2018). However, the multi-taper method results in the smearing of lower frequencies (<30 Hz), meaning it can be challenging to isolate discrete time-frequency characteristics (Cohen, 2014). However, this is less problematic when assessing power at bandwidth resolution, when compared to finer resolutions. In addition, other methods may be more appropriate when precise temporal characteristics are required, as the multi-taper method suffers from relatively low temporal resolution (Cohen, 2014). In this thesis, we apply the multi-taper method due to its suitability for analysing single-trial and higher frequency data (e.g., gamma band; Cohen, 2014).

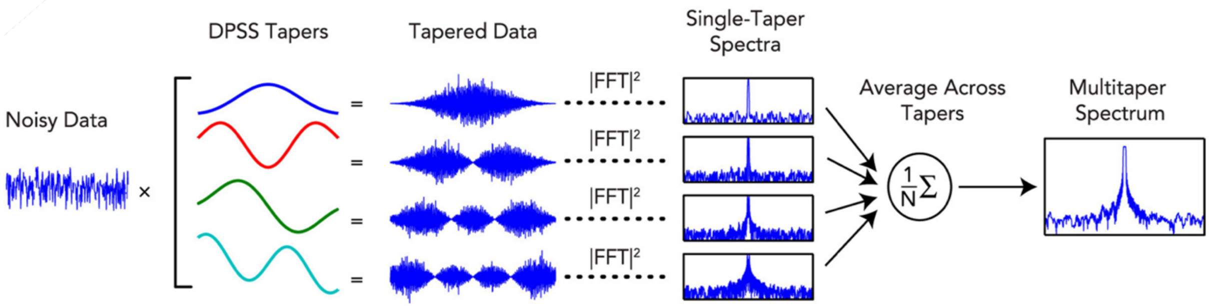


Figure 2.2 Schematic of the multi-taper method for time-frequency decomposition. Adapted from Sleep Neurophysiological Dynamics Through the Lens of Multitaper Spectral Analysis by M. J. Prerau, R. E. Brown, M.T. Bianchi, J.M. Ellenbogen, and P. L. Purdon, 2017, Physiology, 32(1), 60-92, Copyright (2017) by Int. Union Physiol. Sci./Am. Physiol. Soc. Adapted with permission.

2.1.6 Quantification of Event-related Changes in Cortical Oscillations

EEG oscillations can be assessed in terms of amplitude (µV) or power (µV^2; Mathewson et al., 2015). In addition, power spectra can be evaluated using either absolute or relative power. The absolute power of a frequency band is calculated by integrating all power values within a given frequency range (Govindan et al., 2017; Yuvaraj et al., 2014). It is notable that due to anatomical variations, absolute power differs significantly across individuals (Kropotov, 2009). Alternatively, relative power can be calculated by dividing the absolute power of each frequency band by the sum of powers across all frequency bands (Govindan et al., 2017). Relative power has lower inter-individual variation when compared to absolute power and is more suited to comparing across subjects (Harmonya et al., 1993; Nuwer, 1988).

Changes in time-locked oscillations which are associated with experimental stimuli can be evaluated using the event-related desynchronisation method (ERD; Pfurtscheller & Aranibar, 1979). Traditionally, ERD values are calculated across trials, whereby trials are averaged to reduce variability and improve the signal-to-noise ratio (Pfurtscheller & Aranibar, 1977; Pfurtscheller & Lopes da Silva, 1999). In the current thesis, single-trial ERD values are used to train ML models. Whilst single-trial data is inherently noisier, information is lost during the averaging process. ERD can be calculated for a given trial using the equation below.

$$ERD\ (\%) = \left(\frac{A - R}{R}\right) * 100$$

ERD reflects the change in power, represented as a percentage, following the onset of a given event, or active period (A), relative to the corresponding baseline period (Pfurtscheller &

Aranibar, 1977, 1979). Negative ERD values represent a decrease in band power, indicating increased cortical excitation. Positive values, known as event-related synchronisation (ERS), represent band power increases, reflecting cortical inhibition (Pfurtscheller & Aranibar, 1977, 1979).

2.1.7 Event-related Potentials

An alternative approach for analysing EEG data is the event-related potential (ERP) technique. Time-locked voltage changes in the EEG signal due to the onset of an event or stimulus (e.g., visual, auditory, somatosensory) are commonly referred to as Event-related potentials (Lopes da Silva, 2011; Luck, 2014; Sur & Sinha, 2009). ERP analysis assumes that the electrical response evoked by the presentation of stimulus is time-locked, or delayed relative to the onset of the stimulus and that the ongoing EEG activity is stationary (Lopes da Silva, 2011). Consequently, like the ERD method, ERP identification relies on improving the signal-to-noise ratio, which is achieved by averaging across numerous trials (Lopes da Silva, 2011; Luck, 2014). Averaging trials improves the signal-to-noise ratio as time-locked activity is preserved, whilst voltage fluctuations which are not time-locked to the stimulus (e.g., noise) are minimised (e.g., positive, and negative deflections cancel out and approach zero; Luck, 2014). Therefore, the number of trials is an important consideration as the signal-to-noise ratio improves proportionally to the square root of the number of trials (Lopes da Silva, 2011).

To compute an ERP it is important to define a period of fixed length and extract segments or epochs from continuous data (Luck, 2014). Epochs contain a baseline period, spanning a few hundred milliseconds before stimulus onset, and a period after stimulus presentation, which often has a duration of 500-1500ms depending on the component of interest (Luck, 2014). To account for offset and drift in the EEG, which can be produced by factors such as skin hydration, a baseline correction procedure is applied (Luck, 2014). Pre-stimulus EEG is often appropriate for baseline correction as the period provides a good estimate of the EEG offset and does not contain EEG activity relating to the stimulus (although this assumption is sometimes violated as the pre-stimulus period may contain residual activity from the preceding event such as anticipation effects; Luck, 2014). Baseline correction is conducted by calculating the average voltage during the pre-stimulus period and subtracting this value from the entire waveform (Luck, 2014). Without performing baseline correction, the data would include additional cross-trial variability due to different offsets, significantly increasing data variance, and impairing statistical analysis (Luck, 2014). Finally, in traditional ERP analysis, following the baseline correction, the data is averaged by summing all trial waveforms and dividing by the number of waveforms (Luck, 2014).

Positive and negative deflections are referred to as components. Specifically, ERP components exist within the overall complex waveform but are represented by specific characteristics such as positive and negative deflections. ERP components are defined by their polarity (P: positive; N: negative) and their post-stimulus latency (e.g., P300) or the order of the component (e.g., P3; Luck, 2014; Woodman, 2010). Exogenous components (early components approximately occurring between 60 and 100ms) are triggered by the onset of a stimulus and reflect automatic processing and the physical features of the stimulus (Luck, 2014; Sur & Sinha, 2009). Whereas endogenous components (later components > 150ms) are entirely related to the task and reflect cognitive processing, which can be altered due to factors such as attention (Luck, 2014; Sur & Sinha, 2009). In addition to time-frequency analysis, the analysis of ERPs is also explored in this thesis (See Chapter 5).

2.1.8 Strengths and Limitations of EEG

EEG benefits from several strengths. Firstly, EEG has an excellent temporal resolution which enables sub-millisecond sampling (Lau-Zhu et al., 2019; Michel & Murray, 2012; Ploner & May, 2018; Tivadar & Murray, 2019). The temporal resolution facilitates an improved understanding of the neural underpinnings of stimuli processing and responses, which is challenging to measure accurately using behavioural approaches such as reaction times. Secondly, EEG is advantageous over other neuroimaging techniques because the method provides a direct measure of population-level neural activity (Cohen, 2017; Lau-Zhu et al., 2019). Alternative methods rely on proxy measures of neural activity, such as the haemodynamic response in fMRI, which is dependent on the assumption that increased neural activity is correlated with increased blood flow to that region (Logothetis, 2008). Relying on haemodynamic measures of neural activity results in poor temporal resolution, as haemodynamic responses occur over a time scale of several seconds, whilst neural activity occurs within the millisecond range (Glover, 2011). Thirdly, EEG hardware is inexpensive compared to other popular neuroimaging techniques and there are many free software packages for data analysis (Lau-Zhu et al., 2019; Ploner & May, 2018; Tivadar & Murray, 2019). Fourthly, EEG is easy to use and is accessible across the entirety of the human lifespan (e.g., useable for both neonatal and elderly populations) and in clinical settings and populations (Lau-Zhu et al., 2019; Tivadar & Murray, 2019). The utility of EEG is further enabled as some modern EEG systems are portable (e.g., Mindo Triolbite 32-channel dry EEG system, Mindo, National Chiao Tung University Brain Research Centre, Taiwan), which do not require dedicated environments such as Faraday cages (Tivadar & Murray, 2019). These systems often demonstrated comparable results to standard EEG systems (Hinrichs et al., 2020). Finally, EEG can be easily combined with other neuroimaging methods such as fMRI, neuromodulation techniques such as transcranial magnetic stimulation (TMS) and pharmaceutical interventions to name but a few (Ploner & May, 2018; Tivadar & Murray, 2019).

Whilst EEG offers many advantages over alternative neuroimaging methods, it also suffers from several considerable limitations. The most prominent limitation of EEG is poor spatial resolution (Lau-Zhu et al., 2019; Michel & Brunet, 2019; Ploner & May, 2018; Tivadar & Murray, 2019). EEG electrodes record the electrical activity of the brain using electrodes located on the scalp. Consequently, EEG is insensitive to deep and sub-cortical regions (Ploner & May, 2018). In addition, neural activity is attenuated by resistive tissues such as the meninges or the skull before being recorded by the EEG system (Nunez et al., 1997; Nunez & Srinivasan, 2006; Srinivasan et al., 1996). Therefore, EEG is subject to the inverse problem; the determination of the intracranial sources that contribute to the recorded scalp potential (Michel & Brunet, 2019). That is, the inverse problem aims to identify the source of the signal, given that the scalp potentials and volume conduction models are known entities (Caune et al., 2014; Olejniczak, 2006). The solution of the inverse problem is not unique as there are infinite solutions and combinations which could have resulted in the observed scalp potential (Grech et al., 2008; Pascual-Marqui, 1999). Estimates of the source can be obtained using source localisation methods (Grech et al., 2008). However, source localisation can be affected by head and source modelling errors and EEG noise (Whittingstall et al., 2003). Finally, EEG suffers from a low signal-to-noise ratio (Tivadar & Murray, 2019). The magnitude of the true EEG signal is significantly smaller than common artefacts, which hampers the signal-to-noise ratio.

2.2 Principles of Machine Learning

ML is a subfield of Artificial Intelligence (AI) that uses data to develop models for pattern recognition, classification and prediction using principles from numerous disciplines including computer science and statistics (Jordan & Mitchell, 2015; Tarca et al., 2007). In 1959, Arthur Samuel defined ML as a discipline that allows computers to learn without explicit programming (Samuel, 1959). Generally, ML can be considered an area of applied statistics, where the objective is to statistically estimate complex functions (Goodfellow et al., 2016). ML algorithms learn directly from data by altering the model's parameters as a function of experience (e.g., training), aiming to identify parameters which produce the optimal solution (Jordan & Mitchell, 2015). Following training, ML models can be used to make predictions on novel data (Goodfellow et al., 2016; Vu et al., 2018).

Unsupervised, supervised and reinforcement learning algorithms are arguably the core three pillars of ML (Jordan & Mitchell, 2015; J. H. Lee et al., 2018; Tarca et al., 2007; Vu et al., 2018). Unsupervised learning aims to identify representations (e.g., clustering) in unlabelled data, whilst reinforcement learning algorithms aim to learn actions that maximise a reward and are analogous to conditioning (Jordan & Mitchell, 2015; Vu et al., 2018). Supervised learning algorithms are trained on labelled data to make predictions on novel data (e.g., classification or regression; Jordan & Mitchell, 2015; LeCun et al., 2015). Only supervised ML algorithms are implemented in this thesis. Therefore, throughout this section, we will provide an overview of supervised learning. For a broad review of ML that includes both unsupervised and reinforcement learning see Jordan and Mitchell (2015).

Supervised ML involves training a model to predict the outcome or response variable, given a model and input data (LeCun et al., 2015; Pereira & Borysov, 2019; Tarca et al., 2007). The aim of supervised ML is to learn a function that achieves the optimal mapping between input-output pairs using a range of probabilistic, optimisation and statistical techniques (Jordan & Mitchell, 2015; LeCun et al., 2015; Osisanwo et al., 2017; Pereira & Borysov, 2019; Samuel, 1959; Uddin et al., 2019; Vu et al., 2018). Essentially, supervised learning aims to identify a function, *f*, that achieves the optimal mapping of input data, *X*, to an output or label, *Y* (Jordan & Mitchell, 2015; Osisanwo et al., 2017; L. Yang & Shami, 2020). That is, supervised ML algorithms infer a function from input training data, to make predictions on unseen data.

Classification, where models are required to separate the data into discrete categories, is an example of supervised ML (Goodfellow et al., 2016; Jordan & Mitchell, 2015; LeCun et al., 2015). Here, the output or target variable is a discrete class (e.g., cat or dog), which can be represented by integers ($y_i \in \mathbb{Z}$). Classification tasks can be either binary, where the number of classes is equal to two, or multiclass which extends to three or more classes (Goodfellow et al., 2016; Sokolova & Lapalme, 2009; Uddin et al., 2019). Additionally, regression tasks, where the model is required to predict a real value given the input data, can also be achieved using supervised learning ($y_i \in \mathbb{R}$; Uddin et al., 2019). Figure 2.3 presents illustrations of classification and regression for both linear and non-linear data in two-dimensional space.

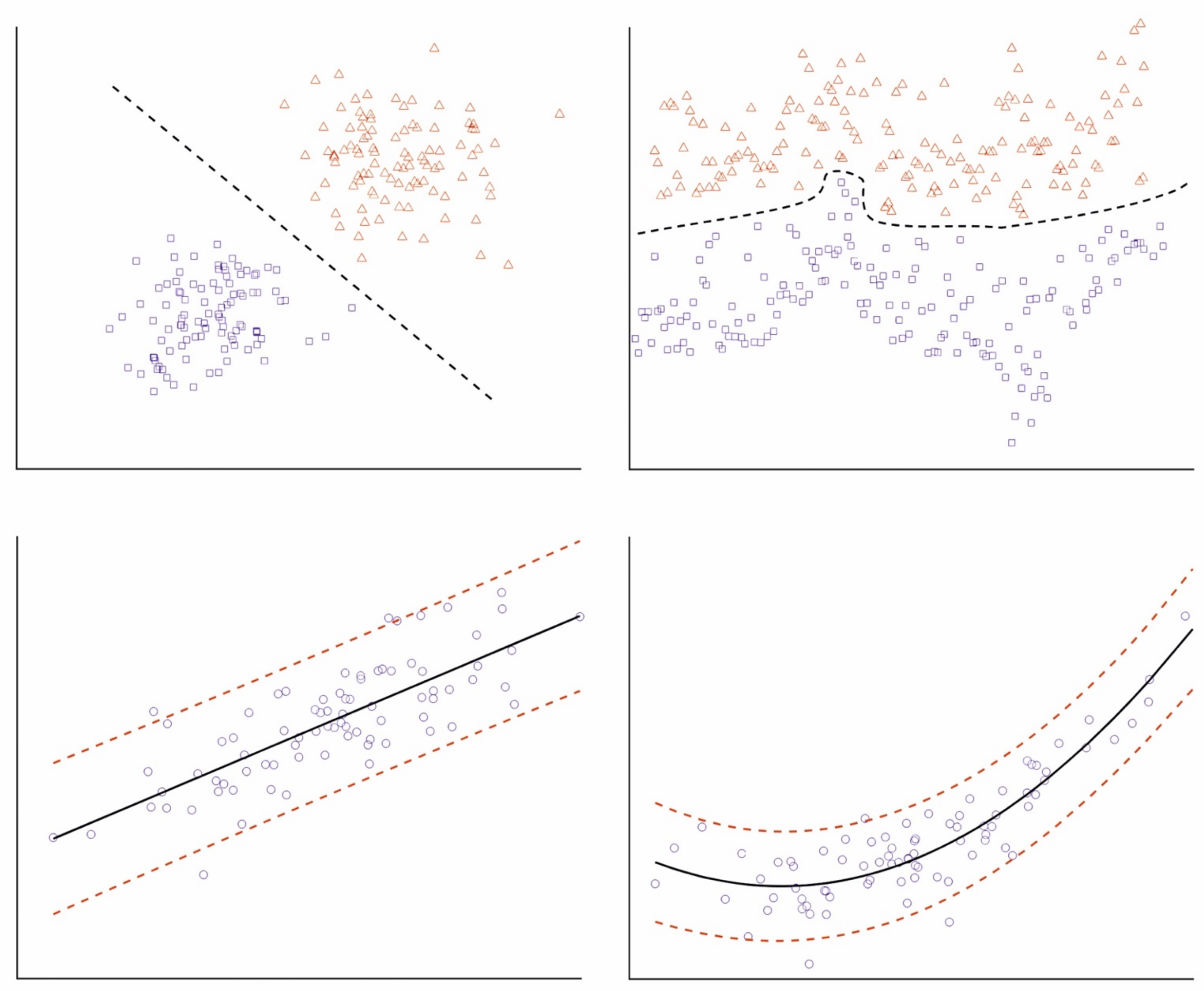


Figure 2.3 Schematic diagrams of classification (top two panels) and regression (bottom two panels) for both linear (left panels) and non-linear (right panels) data. ***Top left panel****. Illustration of a linear classification model. The data can be perfectly separated by a straight line.* ***Top right panel****. Non-linear classification diagram. The data cannot be completely separated by a straight line. Note, in both classification plots, the black dotted line represents the decision boundary of the classifier.* ***Bottom left panel****. Example of linear regression.* ***Bottom right panel****. Example of non-linear (quadratic) regression. Note, in both regression panels, the red dotted line represents 95% confidence intervals.*

The process of training allows the model to learn from experience to predict the output for a given observation (Jordan & Mitchell, 2015; Mahesh, 2020). Model performance iteratively improves until convergence during training. Here, an objective function (also referred to as a loss or cost function) is computed which measures the degree of error between the predicted and true values (LeCun et al., 2015; Q. Wang et al., 2022; Yamashita et al., 2018). During training, ML models alter internal, learnable parameters to minimise the objective (error) function using optimisation algorithms such as gradient descent (LeCun et al., 2015; Yamashita et al., 2018). To adjust internal parameters to minimise the loss, the algorithm calculates a gradient vector for each parameter, which illustrates the change in error given an adjustment of the model parameter (LeCun et al., 2015). The model then adjusts the parameter in the direction that reduces the error (LeCun et al., 2015). This process can be repeated until the optimal parameters have been identified.

2.2.1 Supervised ML Algorithms: Random Forest

In this thesis, we implement 7 traditional supervised ML algorithms including an adaptive boosting algorithm (AdaBoost), linear discriminant analysis (LDA), logistic regression (LR), gaussian naïve Bayes (NB), random forest (RF), support vector machine (SVM), and an extreme gradient boosting algorithm (XGBoost). Additionally, a long short-term memory network (LSTM) is developed in Chapter 6. These algorithms were implemented as they were amongst common reported in the literature (e.g., SVM, LDA, RF) or represented an area of novelty, as they had not been previously assessed (e.g., boosting algorithms; Mari et al., 2022). The RF model demonstrated superior performance to alternative models, which is unsurprising given that it has been shown to perform well on real-world data, requires minimal hyperparameter optimisation, and is robust to problems of overfitting (Bergstra & Bengio, 2012; Dong et al., 2020; Fernández-Delgado et al., 2014; Géron, 2019; T. Jiang et al., 2020; Mienye & Sun, 2022; L. Yang & Shami, 2020). Given that the RF model demonstrated the best performance of all traditional models tested in Chapter 4 of this thesis, only the RF is described in this section for succinctness. Comprehensive overviews of alternative ML methods are presented elsewhere (T. Jiang et al., 2020; Larrañaga et al., 2006; LeCun et al., 2015; Osisanwo et al., 2017; Sarker, 2021; Tarca et al., 2007; Uddin et al., 2019). Additionally, efficient implementations of these models are readily available through software packages such as Scikit-learn (Pedregosa et al., 2011).

Given that RF models are comprised of a series of decision trees (DT; Dong et al., 2020; Sagi & Rokach, 2018), we first provide a brief overview of DTs. DTs have a hierarchical arrangement, resembling a tree-like structure, consisting of a sequence of hierarchical binary partitions of input data (Alloghani et al., 2020; Uddin et al., 2019; Venkatasubramaniam et al., 2017). The DT is comprised of nodes, which have multiple levels and begin with a root node (Alloghani et al., 2020; Uddin et al., 2019; Venkatasubramaniam et al., 2017). Internal nodes aim to separate the data into two disjoint categories based on set criteria, with the categories below the node referred to as the branches (Venkatasubramaniam et al., 2017). The segmentation of the data into classes continues recursively along the structure of the DT until a leaf node is reached, where a stopping rule is implemented (Alloghani et al., 2020; Uddin et al., 2019; Venkatasubramaniam et al., 2017). Predictions can be made on novel observations by passing the data through the DT.

RFs are a type of ensemble learning algorithm, which combines multiple learners to solve a ML task (Dong et al., 2020; Sagi & Rokach, 2018). By combing predictions from multiple learners (e.g., DTs), the prediction error of a single learner will be negated by other learners, resulting in improved performance (Bentéjac et al., 2021; Dong et al., 2020; Sagi & Rokach, 2018). Bagging and boosting approaches are utilised during ensemble learning to improve performance. Independent frameworks (e.g., bagging - RF) involve creating an inducer that is independent of others, meaning that the inducer does not impact the output of other learners (Sagi & Rokach, 2018). Whereas dependent approaches (e.g., boosting) use the output from one inducer to create the next inducer (Sagi & Rokach, 2018).

RFs utilise bagging, which involves building and averaging across numerous DTs with high variance to reduce overfitting (Breiman, 2001; Dong et al., 2020; Mienye & Sun, 2022; Sagi & Rokach, 2018). Here, each model is trained on a bootstrapped copy of the dataset. For a sample of *n* length, each model is trained on a bootstrapped sample with *n* observations, ensuring a sufficient training sample size (Bentéjac et al., 2021; Sagi & Rokach, 2018). Bootstrapped samples may contain repeated observations (approx. 37%), meaning that some samples will be duplicates, whilst others will be omitted for a given learner (Bentéjac et al., 2021; Sagi & Rokach, 2018). The individual models are trained on variations of the bootstrapped data, which is parallelisable, and the resulting models are combined (e.g., majority voting), creating the final classifier (Dong et al., 2020; Sagi & Rokach, 2018).

RFs aim to maximise the difference between trees or decrease dependency, which introduces additional randomness (Sapir-Pichhadze & Kaplan, 2020). Firstly, each DT within the RF is trained on a random subset of data where it randomly selects a set of attributes to identify the optimal split (Bentéjac et al., 2021; Dong et al., 2020; T. Jiang et al., 2020; Sagi & Rokach, 2018). RFs use a majority voting-based approach to produce an output. The instance is classified as the class with the largest number of votes (Bentéjac et al., 2021; Dong et al., 2020; González et al., 2020). In a regression RF model, the output reflects an average of the prediction from each tree. Consequently, the RF architecture minimises overfitting due to majority voting and through several injections of randomness (Dong et al., 2020; González et al., 2020; T. Jiang et al., 2020; Mienye & Sun, 2022).

2.2.2 Feature Selection

ML models are subject to the curse of dimensionality, the notion that increasing the number of features exponentially increases the search space, resulting in an increased likelihood of overfitting (Sagi & Rokach, 2018). Dimensionality reduction using feature extraction and selection reduces overfitting likelihood (Cai et al., 2018). Feature extraction or transformation, reduces the dimensionality of the data, whilst retaining maximum information (e.g., Principal component analysis (PCA); Cai et al., 2018; Khalid et al., 2014). Whereas, feature selection involves identifying a subset of features which are relevant to the prediction task based on evaluation criteria (Cai et al., 2018). Feature selection removes superfluous or irrelevant predictors, improving model performance, computation time, and interpretability (Cai et al., 2018). In this section, we describe two common approaches for feature selection, namely filter and wrapper methods which were implemented throughout this thesis.

Filter methods are applied before modelling to remove irrelevant variables by ranking the features that have the strongest association with the labels (Cai et al., 2018; Chandrashekar & Sahin, 2014; J. Miao & Niu, 2016). Therefore, features that are independent of class labels will be ranked as the least important (Cai et al., 2018; Chandrashekar & Sahin, 2014). The optimal feature set can be determined using a stopping criterion i.e., feature selection is iteratively conducted until a predefined, arbitrary threshold is obtained (Hsu et al., 2011). Numerous evaluation criteria can be used to rank the variables, which are dependent on type of the features and labels (e.g., continuous, categorical, etc.). Examples include the correlation coefficient, mutual information, chi-squared, F-score, and maximum relevance-minimum redundancy algorithm, to name but a few (Cai et al., 2018; Chandrashekar & Sahin, 2014; Guyon & Elisseeff, 2003; Hsu et al., 2011; J. Miao & Niu, 2016). Filter methods are simple to implement, computationally inexpensive, and perform well in applications (Chandrashekar & Sahin, 2014; Hsu et al., 2011).

Alternatively, wrapper methods, which use ML performance as the objective function, can be implemented (Cai et al., 2018; Chandrashekar & Sahin, 2014; J. Miao & Niu, 2016). Wrapper methods include heuristics search and sequential selection algorithms (Chandrashekar & Sahin, 2014). Only variations of sequential selection algorithms are implemented in this thesis, for an overview of heuristics methods see Chandrashekar and Sahin (2014). Sequential feature section initialises with an empty set of features and iteratively adds variables until the optimal combination has been identified (Chandrashekar & Sahin, 2014). Alternatively, the inverse procedure can be implemented, starting with a set of features and recursively removing features until the objective function is maximised (Chandrashekar & Sahin, 2014). Wrapper methods can theoretically achieve higher classification accuracy than filter methods, but often have poor generalisation metrics and are more computationally expensive (Cai et al., 2018; Chandrashekar & Sahin, 2014; Hsu et al., 2011). In practice, filter and wrapper methods are combined which improves performance (Hsu et al., 2011). The dataset is initially filtered, reducing the feature space by eliminating irrelevant features, followed by wrapper methods, which identify the optimal configuration of the remaining features (Cai et al., 2018). The combination of filter and wrapper methods is implemented in this thesis.

2.2.3 Hyperparameter Optimisation

ML models have two types of parameters that can be adjusted. Model parameters are internal values that are initialised and updated through training, whilst hyperparameters are parameters that cannot be identified during training and must be prespecified (L. Yang & Shami, 2020). Many ML algorithms require hyperparameter tuning to achieve optimal performance (Syarif et al., 2016). Hyperparameter optimisation algorithms are effective at identifying optimal values, especially when compared to hand-tuning (Bergstra & Bengio, 2012; L. Yang & Shami, 2020). In this section, we provide descriptions of the hyperparameter optimisation techniques implemented in this thesis, namely grid and random search. Additionally, we present a table adapted from Yang and Shami (2020) which overviews the hyperparameters of many ML algorithms (see Table 2.1).

Grid search exhaustively explores the hyperparameter search space for the optimal combination in a set of predefined values (Bergstra & Bengio, 2012; Syarif et al., 2016; L. Yang & Shami, 2020). That is, grid search evaluates all hyperparameter combinations within a set of user-defined values to identify the optimal configuration. Grid search is easy to implement and can be effectively parallelised (L. Yang & Shami, 2020). However, it is unlikely to identify the true global optimal solution as the input values are user specified. Moreover, grid search is computationally complex, increasing exponentially for every additional hyperparameter value (L. Yang & Shami, 2020). Finally, whilst grid search is optimal for categorical hyperparameters, it is less practical for numerical hyperparameters, which have infinite potential values (L. Yang & Shami, 2020).

Alternatively, random search explores random hyperparameter values between user-defined upper and lower bounds for a set number of iterations (Bergstra & Bengio, 2012; Géron, 2019; L. Yang & Shami, 2020). Theoretically, given a large enough search space, the global optimum hyperparameters can be identified, which is advantageous over grid search (L. Yang & Shami, 2020). Random search is more efficient than grid search and can search over a significantly larger space, which is beneficial for numerical hyperparameters (L. Yang & Shami, 2020). Random search can also be parallelised, whilst the number of iterations can be defined, providing control over the computational complexity (Géron, 2019; L. Yang & Shami, 2020). However, random search may include redundant iterations, as it does not consider previous values (L. Yang & Shami, 2020).

Table 2.1 A comprehensive overview of common ML models, their hyper-parameters, suitable optimization techniques, and available Python libraries. Reprinted from Neurocomputing, 415, by L. Yang & A. Shami. "On hyperparameter optimization of machine learning algorithms: Theory and practice", 295-316, copyright (2020), with permission from Elsevier.

ML Algorithm	Main HPs	Optional HPs	HPO Methods	Libraries
Linear Regression	-	-	-	-
Ridge & Lasso	Alpha	-	BO-GP	Skpot
Logistic Regression	Penalty, c, Solver	-	BO-TPE, SMAC	Hyperopt, SMAC
KNN	n_neighbours	Weights, p, Algorithm	BOs, Hyperband	Skpot, Hyperopt, SMAC, Hyperband
SVM	C, Kernel, Epsilon (for SVR)	Gamma, Coef0, Degree	BO-TPE, SMAC, BHOB	Hyperopt, SMAC, BOHB
NB	Alpha	-	BO-GP	Skpot
DT	Criterion, Max_depth, Min_samples_split, Min_samples_leaf, Max_features	Splitter, Min_weight_fraction_leaf, Max_leaf_nodes	GA, PSO, BO-TPE, SMAC, BOHB	TPOT, Optunity, SMAC, BOHB
RF & ET	n_estimators, Max_depth, Criterion, Min_samples_split, Min_samples_leaf, Max_features	Splitter, Min_weight_fraction_leaf, Max_leaf_nodes	GA, PSO, BO-TPE, SMAC, BOHB	TPOT, Optunity, SMAC, BOHB
XGBoost	n_estimators, Max_depth, Learning_rate, Subsample, Colsample_bytree,	Min_child_weight, Gamma, Alpha, Lambda	GA, PSO, BO-TPE, SMAC, BOHB	TPOT, Optunity, SMAC, BOHB
Voting	Estimators, Voting	Weights	GS	Sklearn
Bagging	Base_estimator, n_estimators	Max_samples, Max_features	GS, Bos	Sklearn, Skpot, Hyperopt, SMAC

AdaBoost	Base_estimator, n_estimators, Learning_rate	-	BO-TPE, SMAC	Hyperopt, SMAC
Deep learning	Number of hidden layers, ‘units’ per layer, Loss, Optimizer, Activation, Learning_rate, Dropout rate, Epochs, Batch_size, Early stop patience	Number of frozen layers (if transfer learning is used)	PSO, BOHB	Optunity, BOHB
K-means	n_clusters	Init, n_init, Max_iter	BOs, Hyperband	Skpot, Hyperopt, SMAC, Hyperband
Hierarchical Clustering	n_clusters, Distance_threshold	Linkage	BOs, Hyperband	Skpot, Hyperopt, SMAC, Hyperband
DBSCAN	eps, min_samples	-	BO-TPE, SMAC, BOHB	Hyperopt, SMAC, BOHB
Gaussian mixture	n_components	Covariance_type, Max_iter, Tol	BO-GP	Skpot
PCA	n_components	Svd_solver	BOs, Hyperband	Skpot, Hyperopt, SMAC, Hyperband
LDA	n_components	Solver, Shrinkage	BOs, Hyperband	Skpot, Hyperopt, SMAC, Hyperband

BO, Bayesian optimization; DBSCAN, Density-based spatial clustering of applications with noise; DT, Decision tree; ET, Extra trees; GA, Genetic algorithm; GP, Gaussian process; GS, Grid search; HB, HyperBand; HP, Hyperparameter; HPO, Hyperparameter optimisation; KNN, k-nearest neighbours; LDA, Linear discriminant analysis; NB, Naïve Bayes; PCA, Principal component analysis; PSO, Particle swarm optimisation; RF; Random forest; SKOPT, Scikit-optimize; SMAC, Sequential model-based algorithm configuration; SVM, Support vector machine; SVR, Support vector regression; TPE, Tree-structured Parzen estimators.

2.2.4 Model Evaluation

2.2.4.1 Internal Validation

ML models require validation, which evaluates model performance using novel data, providing an estimate of the model's generalisability (Maleki et al., 2020; Vabalas et al., 2019). Validation methods include both internal and external validation. Internal validation is an approach to evaluate the predictive capabilities of the model (Moons et al., 2015). Typically, internal validation involves partitioning a single dataset into a training set, which is used to develop the model and a testing set which evaluates the predictive capability of the model, which can be achieved using cross-validation approaches (Browne, 2000; Collins et al., 2015; Koul et al., 2018). Throughout the remainder of this section, we provide a brief overview of cross-validation techniques, before discussing the importance of external validation.

Hold-out validation is the simplest internal validation method and involves randomly splitting a single dataset into training and testing sets (T. Jiang et al., 2020; Koul et al., 2018; Maleki et al., 2020). Here, both the training and testing sets contribute to model development, with the model being trained using the training data, whilst model performance is evaluated on the test set (Maleki et al., 2020). The dataset can be divided in numerous ways, but common splits include 80% - 20% or 70% - 30% for training and testing, respectively. Alternative partitions include 70% for training, 15% for validation, which includes hyperparameter optimisation, and 15% for testing (Maleki et al., 2020). Whilst hold-out validation is simple and computationally inexpensive, model performance fluctuates due to random partitioning (T. Jiang et al., 2020; Koul et al., 2018; Maleki et al., 2020).

Alternatively, k-fold cross-validation, where the data is randomly segmented into k folds, with k-1 folds used for model training and the remaining fold used for validation, can be used (Maleki et al., 2020). Here, the ML model is trained k times, until all folds have been used as the validation set and model performance is calculated as the average of all iterations (Fushiki, 2011; T. Jiang et al., 2020; Luo et al., 2016; Maleki et al., 2020; Wong, 2015). Stratified k-fold cross-validation preserves class distributions in each set, whereas traditional k-fold cross-validation splits the data entirely randomly (Luo et al., 2016; Maleki et al., 2020; Wong, 2015). Stratified k-fold cross-validation is therefore advantageous, as issues arising due to class imbalance are mitigated (Maleki et al., 2020). K-fold cross-validation reduces the randomness associated with a single random split, achieving more reliable estimates of model performance (Koul et al., 2018; Maleki et al., 2020). However, the method is more computationally expensive compared to hold-out validation as each model is trained k times (Koul et al., 2018; Maleki et al., 2020). k-fold cross-validation is parallelisable, which can reduce overall execution time (Maleki et al., 2020).

Finally, leave-one-out cross-validation is a special instance of k-fold cross-validation, where the number of folds is equal to the number of observations (e.g., k=n; Koul et al., 2018; Maleki et al., 2020; Wong, 2015). Consequently, the test set is a single observation, and training is repeated for all observations. Leave-one-out cross-validation is optimal for small datasets, as it maximises the training data and is not subject to the noisy performance estimates associated with both k-fold and hold-out validation (T. Jiang et al., 2020; Koul et al., 2018; Wong, 2015). However, as the model is trained n times, computational complexity is significantly increased (Maleki et al., 2020). In this thesis, we use k-fold validation, as it provides an effective trade-off between computation time and robustness.

2.2.4.2 External Validation

Whilst internal methods are sufficient to provide an estimate of model generalisability, they are not robust enough to provide evidence for practical purposes or clinical translation (Bleeker et al., 2003; Ramspek et al., 2021; Vabalas et al., 2019). Internal validation methods are at risk of overfitting, resulting in overly optimistic estimates of model performance (Cabitza et al., 2021; Vabalas et al., 2019; Varma & Simon, 2006). Simulation research has shown that cross-validation methods (e.g., k-fold cross-validation) are prone to overfitting, especially in small samples, which is common in neuroscientific research (Vabalas et al., 2019). Overfitting is the ability of the model to fit both signal and noise, meaning they can fit random noise with a seemingly high degree of accuracy (e.g., 81%; see Vabalas et al., 2019). The risk of overfitting is further enhanced when hyperparameter optimisation and evaluation are combined, which is a common practice (Arbabshirani et al., 2017; Cawley & Talbot, 2010; Lever et al., 2016; Varma & Simon, 2006). Therefore, ML performance metrics established using internal validation methods may not provide accurate estimates of the model's generalisability. Consequently, it is imperative to assess model performance on independent data, which was not used during model development or internal validation (Lever et al., 2016).

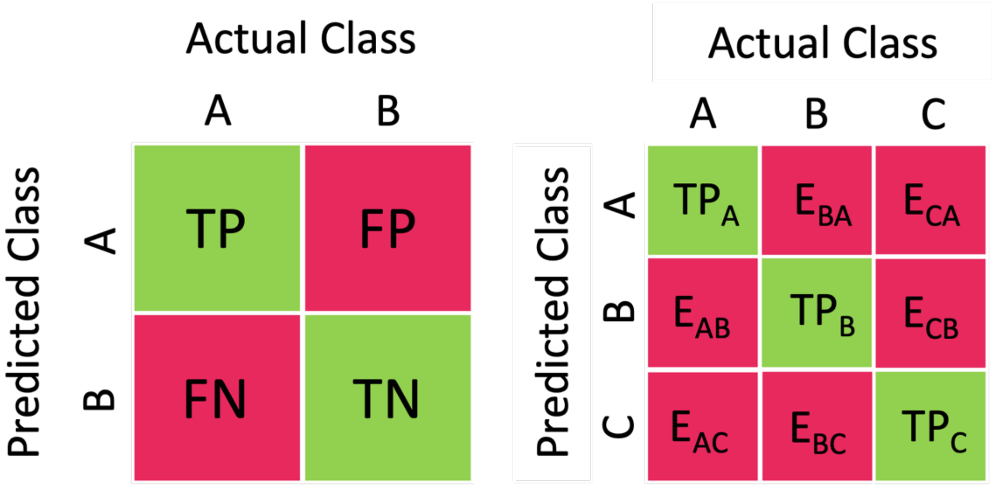
Given the limitations associated with internal validation, external validation is imperative (Bleeker et al., 2003; Steyerberg & Harrell, 2016). External validation involves assessing model performance on data independent of model development (Cabitza et al., 2021). Specifically, external validation can be implemented using data collected from other cohorts (e.g., a different population), facilities or locations (e.g., geographical validation), repositories, or collected at a different time (e.g., temporal validation) or using a different experimental paradigm (Cabitza et al., 2021; Collins et al., 2015). As the external validation dataset is independent from the training and internal validation datasets, any overfitting would fail to generalise (Ho et al., 2020). Consequently, external validation provides robust estimates of model performance and generalisability, which is essential for models with potential clinical applications.

The importance of external validation cannot be understated. A clinical prediction model that is poorly validated would result in suboptimal patient care and negative outcomes (e.g., risk of undertreatment or even mortality; Ramspek et al., 2021). Research has demonstrated that model performance on external data is reduced compared to internal validation (X. Li et al., 2019; Mari et al., 2023; Siontis et al., 2015). Despite the evident importance of external validation, it is rarely assessed, with only 5% of prediction modelling studies published on PubMed reporting external validation in either the title or abstract (Ramspek et al., 2021). In Chapter 3, we conducted a systematic review to explore the effectiveness of ML to predict pain intensity, phenotype, or response to treatment. Our results demonstrated that none of the 44 studies included conducted or reported external validation results (Mari et al., 2022). Therefore, the performance of ML models in the field is likely inflated and not an accurate representation of the current state of the art. Consequently, an important contribution of this thesis is to externally validate ML algorithms to provide robust estimates of model performance for predicting pain-related outcomes.

2.2.4.5 ML Discrimination and Performance Metrics

One approach for quantifying ML performance is by assessing model discrimination, which refers to the ability of the model to accurately differentiate between samples in one condition against other conditions (Alba et al., 2017; Moons et al., 2015). ML discrimination performance can be assessed using numerous metrics depending on the task (e.g., classification or regression; Alba et al., 2017; Powers, 2011; Sokolova & Lapalme, 2009). In this section, we provide an outline of ML discrimination and corresponding performance metrics that are implemented throughout this thesis.

For regression tasks, metrics that describe the error between the predicted outcome and observed value are commonly used such as the mean absolute error (MAE) and root mean square error (RMSE; Chai & Draxler, 2014; Pereira & Borysov, 2019; Willmott & Matsuura, 2005). For classification tasks, the performance can be assessed using metrics derived from a confusion matrix (Pereira & Borysov, 2019; Tharwat, 2021). Confusion matrices provide information regarding the number of correctly and incorrectly classified points per outcome label, which can be applied to both binary and multiclass classification problems (Sokolova & Lapalme, 2009; Tharwat, 2021; Uddin et al., 2019). The confusion matrix describes the number of true positives (TP), which are the positive instances that were correctly classified as positive, true negatives (TN), which are negative samples that were correctly identified as negative, false positives (FP), which reflect negative samples that were incorrectly classified as positive, and false negatives (FN), which represent positive samples that were incorrectly classified as negative (Tharwat, 2021; Uddin et al., 2019). Figure 2.4 provides a schematic of a confusion matrix for both binary and multiclass classification.



Predicted Class		Actual Class A	Actual Class B
	A	TP	FP
	B	FN	TN

Predicted Class		Actual Class A	Actual Class B	Actual Class C
	A	TP_A	E_{BA}	E_{CA}
	B	E_{AB}	TP_B	E_{CB}
	C	E_{AC}	E_{BC}	TP_C

Figure 2.4 Schematic diagram of confusion matrices for binary classification (left) and multiclass classification (right). For multiclass classification, E represents error. Adapted from Classification assessment methods by A. Tharwat, 2021, Applied Computing and Informatics, 17(1), 168-192, Copyright (2018) by Alaa Tharwat. Adapted with permission.

Metrics that describe different aspects of a classification model's performance can be calculated using the confusion matrix. Table 2.2 provides descriptions of performance metrics for both classification and regression tasks which are present throughout this thesis. Comprehensive reviews of these topics have been published elsewhere (Alba et al., 2017; Assel et al., 2017; Chai & Draxler, 2014; Powers, 2011; Sokolova & Lapalme, 2009; Tharwat, 2021; Willmott & Matsuura, 2005). Usually, performance metrics for classification tasks are observed on a scale between 0 and 1, with 1 representing perfect performance, whilst 0.5 is the chance level for a binary classification task. However, for the Brier score, which is affected by both model discrimination and calibration, perfect predictions would output 0.

Table 2.2 ML performance metrics for both classification and regression tasks.

Task	Performance Metric	Notation	Explanation
Classification			
	Accuracy	$\frac{tp + tn}{tp + fp + tn + fn}$	Represents the model's overall effectiveness. Calculated as the ratio of correctly classified observations over all observations.
	Balanced Classification Accuracy	$\frac{1}{2}\left(\frac{tp}{tp + fn} + \frac{tn}{tn + fp}\right)$	Represents the average of sensitivity and specificity, which is useful for imbalanced datasets (e.g., where one class has more observations).
	Brier Score	$\frac{1}{n}\sum_{i=1}^{n}(p_i - o_i)^2$	Represents the mean squared error of the model probability forecast (p_i) and the event outcome (o_i).
	F1 Score	$\frac{2tp}{2tp + fp + fn}$	Represents the harmonic mean between recall and precision.
	False Negative Rate	$\frac{fn}{fn + tp}$	Represents the ratio of positive labels incorrectly classified as negative labels over the total number of positive observations.
	False Positive Rate	$\frac{fp}{fp + tn}$	Represents the ratio of negative labels incorrectly classified as positive labels over the total number of negative observations.
	Negative Predictive Value	$\frac{tn}{tn + fn}$	Represents the ratio of true negatives over the total number of negatively predicted labels.

Positive Predictive Value (Precision)	$\frac{tp}{tp + fp}$	Represents the ratio of true positives over the total number of positively predicted labels.
Sensitivity (Recall)	$\frac{tp}{tp + fn}$	Represents the ratio of true positive samples correctly classified over the total number of positive observations.
Specificity	$\frac{tn}{tn + fp}$	Represents the ratio of true negative samples correctly classified over the total number of negative observations.
Regression		
Mean Absolute Error	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$	Represents the mean absolute error between the true values (y_i) and predicted values ($\hat{y}_i$).
Root Mean Square Error	$\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$	Represents the average error between the true values (y_i) and predicted values ($\hat{y}_i$).

A receiver operating characteristic (ROC) curve is an alternative way to evaluate ML classifiers (Fawcett, 2006; Hajian-Tilaki, 2013; Hoo et al., 2017). ROC curves were developed during World War two to assess the ability of radar operators to detect true aircraft signals from noise (J. Fan et al., 2006). Nowadays, the ROC can be used to determine the effectiveness of diagnostic tests or evaluate ML performance (Fawcett, 2006; Hajian-Tilaki, 2013; Hoo et al., 2017). A ROC is plotted in two dimensions, with the model's true positive rate (sensitivity) on the Y axis and the false positive rate (1 – specificity) on the X axis, providing a visual representation of the true positive and negative trade-off (J. Fan et al., 2006; Fawcett, 2006; Hoo et al., 2017). A classifier with no discriminative ability would be represented by a 45^o line (i.e., y=x) on the plot, as the classifier does not exceed chance classification, producing equal amounts of true and false positives (Fawcett, 2006; Habibzadeh et al., 2016; Hoo et al., 2017). A classifier with reasonable performance should fall above the reference line (Fawcett, 2006; Hoo et al., 2017). The point (0,1) represents perfect discrimination (Fawcett, 2006; Hoo et al., 2017). ROC curves are primarily used with binary classification models, but can be extended to multiclass classification problems (Fawcett, 2006; Mandrekar, 2010; Saito & Rehmsmeier, 2015; Wandishin & Mullen, 2009). Figure 2.5 illustrates a ROC curve for binary classifiers.

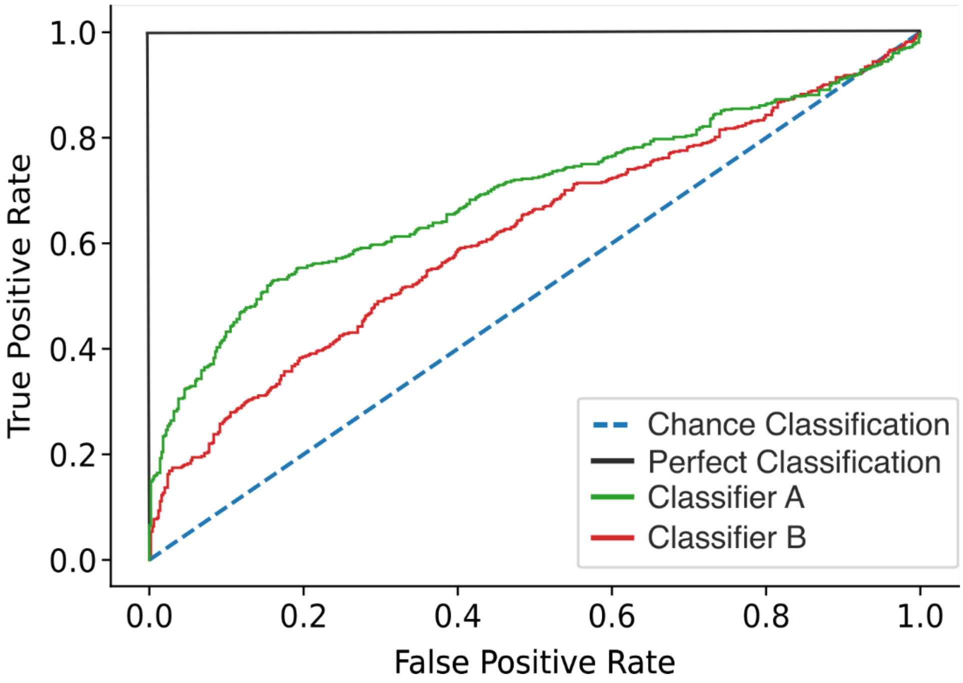


Figure 2.5 The performance of two binary ML classifiers plotted on a ROC curve. The black line represents a perfect classifier, whilst the blue dotted line represents chance classification. In this example, classifier A is the better model.

The area under the ROC curve (AUC) is a popular metric for evaluating model performance (Bradley, 1997). The AUC represents the model's ability to correctly distinguish between two classes (Faraggi & Reiser, 2002; Hoo et al., 2017). The AUC is the two-dimensional area under the ROC curve which provides a summary of model performance at all classification thresholds (Bradley, 1997; Hajian-Tilaki, 2013; Hoo et al., 2017; Kamarudin et al., 2017). The AUC score takes values in the interval between 0 and 1, where 0.5 represents a chance/uninformative model (J. Fan et al., 2006; Habibzadeh et al., 2016). A score of 1 represents perfect discrimination, meaning the model is both 100% sensitive and specific (J. Fan et al., 2006). AUCs can also be used as a threshold for clinical utility. Previously, it was proposed that AUCs ≤ 0.75 are not clinically useful (J. Fan et al., 2006). Therefore, reporting the AUC is imperative and is a requirement of the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis checklist (Collins et al., 2015).

2.2.4.6 Calibration

Prediction models may demonstrate poor risk estimates despite good discrimination performance (Alba et al., 2017; Van Calster et al., 2019). Consequently, evaluating discrimination is insufficient to assess the model's predictive capability (Alba et al., 2017). Therefore, it is imperative to assess calibration, which refers to the agreement between the ML model predictions and observed or reference values (Alba et al., 2017; Luo et al., 2016; Moons et al., 2015; Van Calster et al., 2019). Using a diagnostic example, if a model predicts a 40% risk of a condition being present, then the observed frequency of that condition should be approximately 40 out of 100 events (Alba et al., 2017; Luo et al., 2016; Van Calster et al., 2016, 2019). A poorly calibrated model may over or underestimate the probabilities of the incidence (Alba et al., 2017; Van Calster et al., 2019). For example, a model for predicting mortality rates after cardiac surgery demonstrated good discrimination performance (e.g., AUC of .80 for predicting mortality), but demonstrated poor calibration, resulting in significantly inflated mortality estimates and potentially negative outcomes (Alba et al., 2017). Therefore, calibration assessment is essential to ensure accurate probability estimates (Van Calster et al., 2019). However, it is rarely reported (Moons et al., 2015).

Calibration can be assessed using mean, weak, moderate, and strong calibration, with each level increasing the stringency (Van Calster et al., 2016, 2019). Mean calibration involves comparing the model average incidence with the true event rate (Van Calster et al., 2016, 2019). If the average predicted value is greater than the observed event rate, the algorithm overestimates event probabilities and vice versa (Van Calster et al., 2016, 2019). Weak calibration assesses whether, on average, the model is overfitting or underfitting, providing inaccurate estimations (Y. Huang et al., 2020; Van Calster et al., 2016, 2019). Weak calibration measures the robustness of the estimates, ensuring they are not consistently on the boundaries of the scale (0 or 1) or modest (too similar to the observed prevalence of the condition; Van Calster et al., 2019). Weak calibration can be assessed using a calibration (Cox) slope and intercept, which are equal to 1 and 0, respectively, for perfect performance (Y. Huang et al., 2020; Van Calster et al., 2016, 2019). Moderate calibration assesses whether the model-predicted probabilities are comparable to the observed values (e.g., if the model predicts a 20% risk of a condition being present, then the true condition prevalence should be 20%; Van Calster et al., 2016, 2019). Moderate calibration can be assessed using calibration curves (Moons et al., 2015; Van Calster et al., 2019). Finally, strong calibration assesses whether predicted risks are comparable to the observed event probabilities for all combinations of predictor values (Van Calster et al., 2016, 2019). However, strong calibration assessment is unrealistic requiring an infinitely large dataset to obtain accurate estimates (Van Calster et al., 2016).

Calibration curves (moderate calibration) are the preferred approach to evaluating model calibration (Moons et al., 2015; Van Calster et al., 2016). Calibration curves display the relationship between the estimated or predicted risks, which are plotted on the x-axis, and the observed probabilities, which are plotted on the y-axis (Van Calster et al., 2019). The probabilities are separated into bins (e.g., 10 equal intervals between 0 and 1) and the probabilities in each bin are plotted (Y. Huang et al., 2020). Perfect calibration occurs when the predicted probabilities perfectly match the true probabilities, which can be represented by a 45^o line (e.g., y=x; Huang et al., 2020; Van Calster et al., 2019). Figure 2.6 shows examples of simulated calibration curves (as well as slopes and intercepts) from Van Calster and colleagues (2019). We assess ML models' calibration using calibration curves in this thesis, as it is the recommended assessment method.

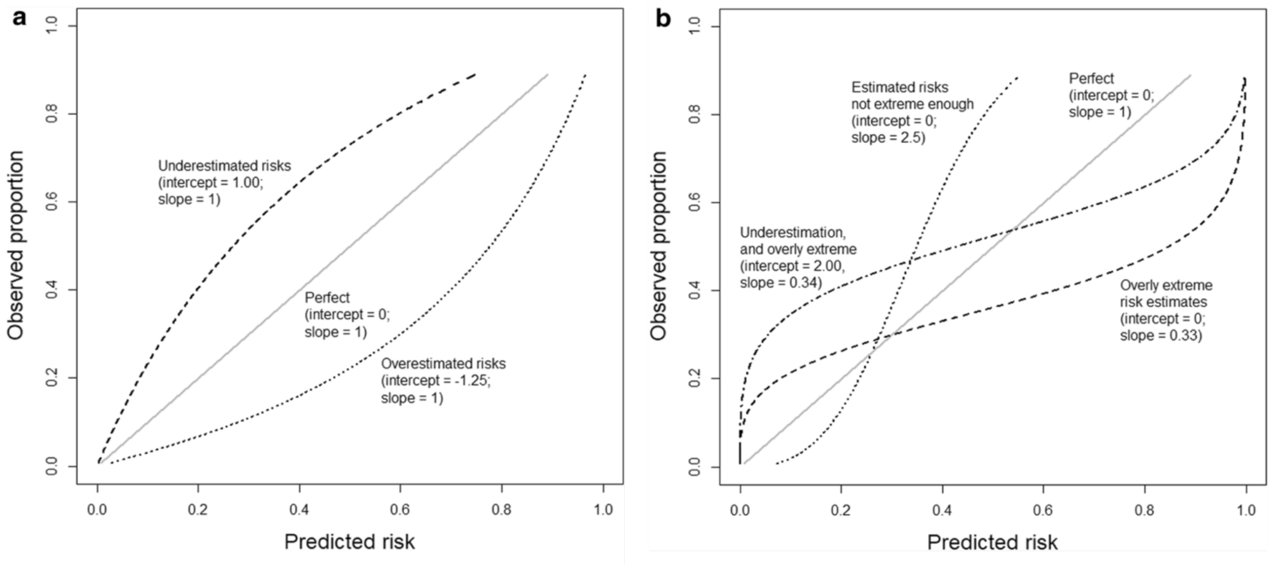


*Figure 2.6 Schematics detailing the different types of miscalibration. **A**) Under- and overestimation of probabilities. **B**) Model predictions that are too extreme or too moderate. Graphics are reflective of a model with an AUC of 0.71 with an event rate of 25%. The slope and intercept of the curves are also provided in the figure. Adapted from Calibration: the Achilles Heel of predictive analytics by B. Van Calster, D.J. McLernon, M. van Smeden et al., 2019, BMC Medicine, 17(230), 1-7, Copyright (2019) by B. Van Calster et al., with permission from Springer Nature.*

2.3 Systematic Review Methodology

A systematic review provides a summary of a research area, utilising reproducible methods to systematically identify, critically assess and synthesise a collection of research studies (Gopalakrishnan & Ganeshkumar, 2013; Uman, 2011). By using a systematic process, systematic reviews often exhibit less bias, such as selection bias, than other review methodologies (e.g., narrative review; Uman, 2011). Systematic reviews are therefore important tools for healthcare, aiding the development of clinical guidelines (Gopalakrishnan & Ganeshkumar, 2013; Shamseer et al., 2015). However, due to the potential implications and impact of systematic review conclusions, they are required to be conducted rigorously following stringent methodology aided by the creation of a protocol (Shamseer et al., 2015), which should be carefully followed and any deviations reported and justified. This process can be done internally, alternatively open-access sources exist (e.g., PROSPERO; https://www.crd.york.ac.uk/prospero/), which allow systematic reviews to be registered online, with the registration identified being cited in the final published manuscript, allowing readers to assess any deviations.

Conducting systematic reviews using reproducible methods remains imperative. However, it is also essential to accurately report the method and results of a systematic review. Fortunately, guidelines exist for conducting systematic reviews, enabling a high degree of credibility and reproducibility, which is critical to their value. The Preferred Reporting Items for Systematic Review and Meta-Analysis (PRISMA) statement provides clear guidance and recommendations to successfully conduct a systematic review (Page, McKenzie, et al., 2021).

Systematic reviews usually follow the process of formulating a review question, defining the inclusion and exclusion criteria, creating the search strategy, selecting the studies, extracting data, performing study quality assessment, analysing and interpreting the results, and disseminating the findings (Uman, 2011). The PRISMA statement provides guidance for all of the stages of reporting the results from the title, where the publication must be identified as a systematic review and/or meta-analysis, to the additional information, where the source and role of the funders must be accurately document (Page, Moher, et al., 2021). Authors of systematic reviews are encouraged to complete the checklist and use the information in the explanation and elaboration process to ensure the review is correctly and rigorously reported (Page, Moher, et al., 2021).

For this thesis, we conducted a systematic review which investigated the effectiveness of ML algorithms and EEG for the prediction of pain intensity, phenotype, and response to treatment. The systematic review is presented in chapter 3. The review is reported in-line with the recommendations and guidelines from PRISMA.

Chapter 3:

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 work.

This systematic review summarises and evaluates the current state-of-the-art machine learning approaches for the prediction of subjective pain intensity, pain phenotypes, and response to treatment using electroencephalography data.

This paper was published in the Journal of Pain (2022). DOI:
https://doi.org/10.1016/j.jpain.2021.07.011
The format of the text has been modified to match the style of this thesis.

The roles of the co-authors are summarised below:

Tyler Mari: Conceptualisation, Methodology, Formal analysis, Investigation, Data curation, Writing – Original Draft, Writing – Review and Editing, Visualisation. **Jessica Henderson:** Investigation, Formal analysis, Writing – Review and Editing. **Michelle Maden:** Methodology Investigation, Data curation, Writing – Review and Editing. **Sarah Nevitt:** Methodology, Formal analysis, Writing – Review and Editing. **Rui Duarte:** Conceptualisation, Methodology, Formal analysis, Investigation, Writing – Original Draft, Writing – Review and Editing, Supervision. **Nicholas Fallon:** Conceptualisation, Methodology, Formal analysis, Investigation, Writing – Original Draft, Writing – Review and Editing, Supervision.

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.

3.1 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 (Dansie & Turk, 2013; Simons et al., 2014; Younger et al., 2009). Traditionally, pain is evaluated using interviews, observations, psychological screening and rating scales (Breivik et al., 2008; Dansie & Turk, 2013; Haefeli & Elfering, 2006; Williamson & Hoggart, 2005). 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 (Breivik et al., 2008; Herr et al., 2011). Moreover, imperfect tools, coupled with the complexity of pain, also inhibit accurate diagnoses and treatment, further limiting the management of clinical pain (Breivik et al., 2008; Fine, 2011; Varrassi et al., 2010). 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 (van der Miesen et al., 2019). ML refers to an algorithm that learns complex data patterns and makes predictions without being explicitly programmed (Samuel, 1959). 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 (Kotsiantis, 2007; Lundervold & Lundervold, 2019; Uddin et al., 2019; Vu et al., 2018). This is achieved by altering internal weights; minimising the error between the input and the predicted label using optimisation algorithms, such as gradient descent (LeCun et al., 2015; Lundervold & Lundervold, 2019;

Whittington & Bogacz, 2019). Therefore, the algorithm learns from experience and can then be used to predict the labels of novel, unseen data (Lötsch & Ultsch, 2018). We focus on the application of ML on Electroencephalogram (EEG), as it is inexpensive and accessible, making it an excellent candidate for clinical applications (Gram et al., 2017; Ta Dinh et al., 2019). 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 (Bargshady et al., 2020; Littlewort et al., 2009; Roy et al., 2016). 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 (Alloghani et al., 2020; Dey, 2016; Jordan & Mitchell, 2015; Lötsch & Ultsch, 2018; Sarker, 2021; Uddin et al., 2019).

By applying supervised ML, researchers have successfully decoded patterns of neuronal activation arising from pain-related outcomes (van der Miesen et al., 2019). 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 (Davenport & Kalakota, 2019; Davis et al., 2017). For example, predicting treatment effectiveness could reduce ineffective trial-and-error treatment and improve patient outcomes (Ginsburg & McCarthy, 2001; Gram et al., 2015, 2017). Despite their potential, pain biomarkers have not significantly impacted public health or clinical practice to-date (Woo et al., 2017). 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.

3.2 Methods

This systematic review is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Moher et al., 2009). The review protocol was registered on PROSPERO on June 5th, 2020 as CRD42020172091.

3.2.1 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.

3.2.2 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 3.1. Reviewer discrepancies at either stage were resolved through discussion or consultation with a third reviewer (NF), who acted as an arbiter, if necessary.

Table 3.1 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.

Abbreviations: EEG, electroencephalogram; ML, machine learning.

3.2.3 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 3.2. A comprehensive discussion of the performance metrics has been reported elsewhere (Chai & Draxler, 2014; Hossin & Sulaiman, 2015; Powers, 2011; Sokolova & Lapalme, 2009; Tharwat, 2020; Willmott & Matsuura, 2005).

Table 3.2 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.

Abbreviations: 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.

3.2.4 Risk of Bias

Assessment of risk of bias (ROB) was performed by using the prediction model risk of bias assessment tool (PROBAST; Wolff et al., 2019), which contains 20 signalling questions to assess ROB across four domains: participants, predictors, outcomes and analysis (Moons et al., 2019; Wolff et al., 2019). 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 (Wolff et al., 2019). 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 (Wolff et al., 2019). 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).

3.2.5 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 (Collins et al., 2015). 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 (Yusuf et al., 2020), 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).

3.2.6 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 (Campbell et al., 2020). 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.

3.3 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 3.1. Subsequently, a total of 44 results were included in this review, with 22 evaluating the prediction of pain intensity (Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; Bai et al., 2016; T. Cao et al., 2020; Elsayed et al., 2020; Furman et al., 2018; Hadjileontiadis, 2015; Kaur et al., 2019; Kimura et al., 2021; L. Li et al., 2018; Misra, Wang, et al., 2017; Nezam et al., 2021; Okolo & Omurtag, 2018; Prichep et al., 2018; Sai et al., 2019; Schulz et al., 2012; Tripanpitak et al., 2020; Tu et al., 2016; Vatankhah et al., 2013; Vijayakumar et al., 2017; M. Yu, Sun, et al., 2020; M. Yu, Yan, et al., 2020), 15 of pain phenotypes (Akben et al., 2012, 2016; Z. Cao et al., 2018; De Tommaso et al., 1999; Frid et al., 2020; Graversen et al., 2011; Levitt et al., 2020; Paul et al., 2019; Saif et al., 2021; Sarnthein et al., 2006; Subasi et al., 2019; Ta Dinh et al., 2019; Vanneste et al., 2018; Vuckovic et al., 2018; Wydenkeller et al., 2009) and seven of response to treatment (Gram et al., 2015, 2017; Graversen et al., 2012, 2015; Grosen et al., 2017; Hunter et al., 2009; Wei et al., 2020).

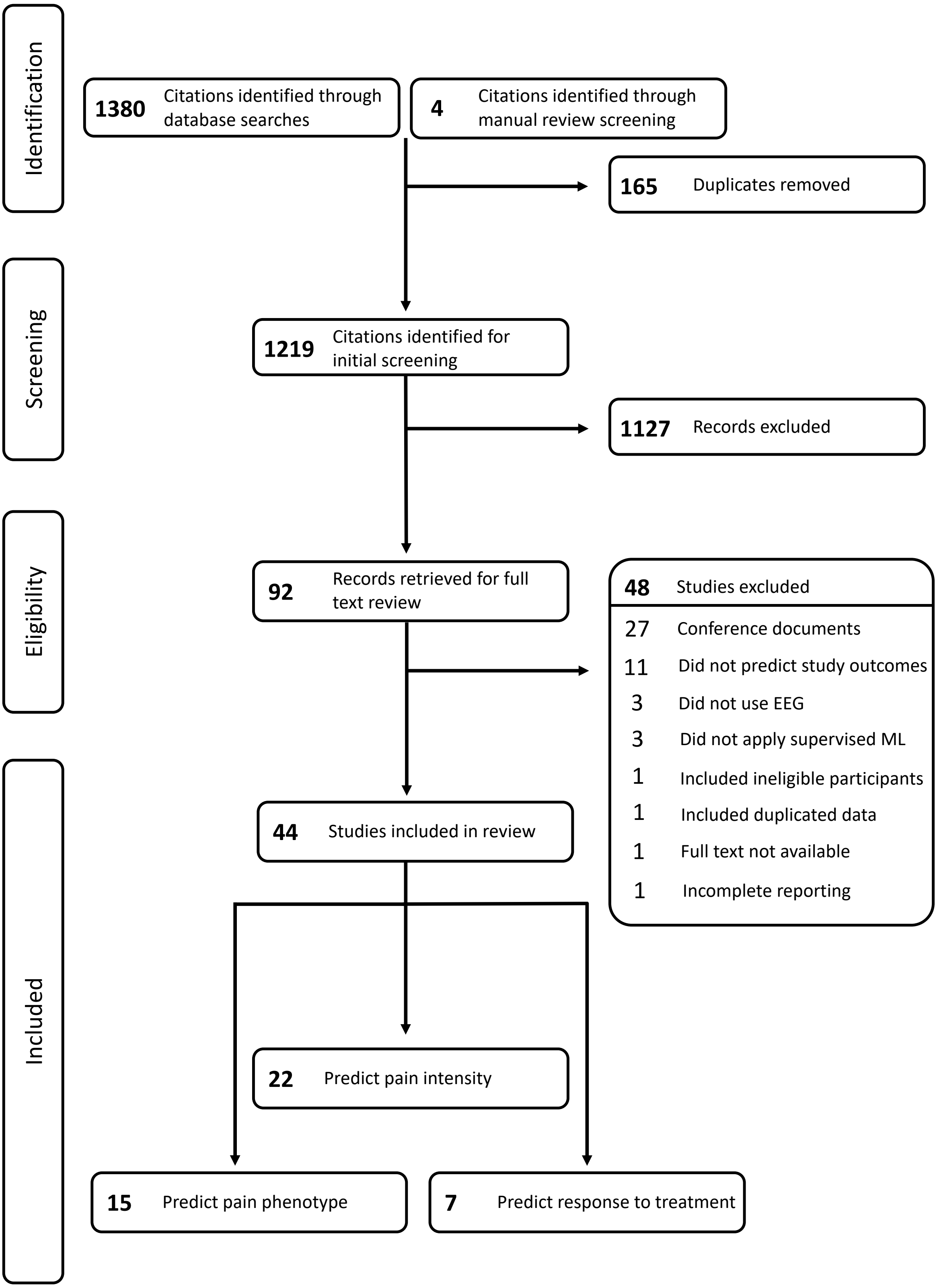

Identification
1380 Citations identified through database searches
4 Citations identified through manual review screening
165 Duplicates removed
Screening
1219 Citations identified for initial screening
1127 Records excluded
Eligibility
92 Records retrieved for full text review
48 Studies excluded
27 Conference documents
11 Did not predict study outcomes
3 Did not use EEG
3 Did not apply supervised ML
1 Included ineligible participants
1 Included duplicated data
1 Full text not available
1 Incomplete reporting
Included
44 Studies included in review
22 Predict pain intensity
15 Predict pain phenotype
7 Predict response to treatment

Figure 3.1 PRISMA flowchart

3.3.1 PROBAST Assessment

The ROB assessment demonstrated that 42 of the 44 studies in this review were categorised as high ROB, as summarised in Figure 3.2. The full assessment is presented in supplementary material 1. 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; Moons et al., 2019), 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.

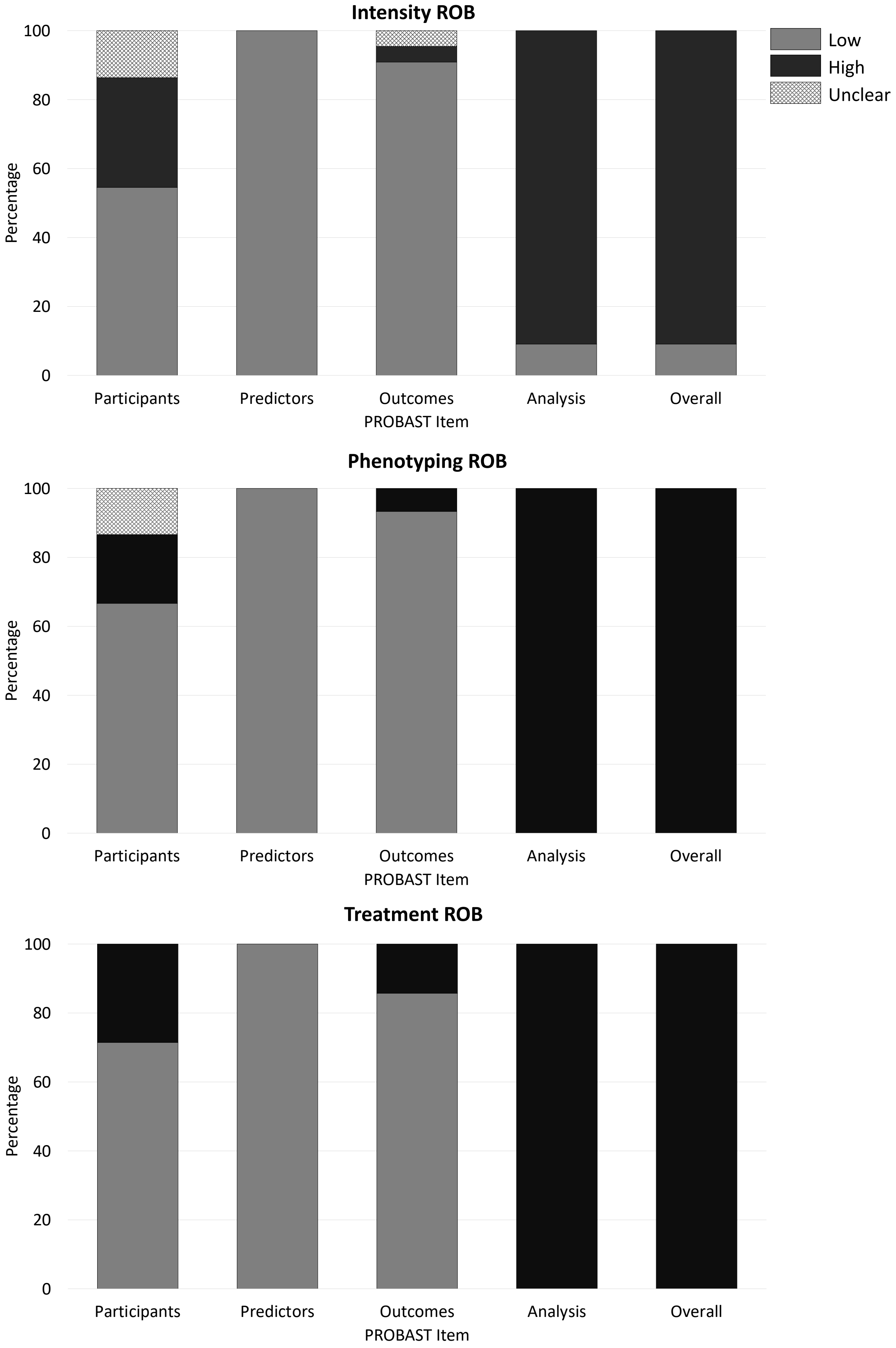


Figure 3.2 PROBAST assessments for pain intensity, pain phenotyping and response to treatment studies.

The applicability assessment demonstrated that only one of the 44 included studies was deemed as having applicability concerns (Okolo & Omurtag, 2018). The study was deemed as having high applicability concerns on the outcome domain (Okolo & Omurtag, 2018). 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 1.

3.3.2 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 3.3. However, the low adherence to reporting guidelines could be partially explained by the compatibility of the tools used (see review limitations).

Table 3.3 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 prediction model, including how and when they were measured.	22 (100%)	14 (93%)	7 (100%)
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%)

3.3.3 Pain Intensity

The characteristics of the 22 included studies which investigated pain intensity are reported in Table 3.4. 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 (Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; Bai et al., 2016; T. Cao et al., 2020; Hadjileontiadis, 2015; Kaur et al., 2019; Misra, Wang, et al., 2017; Okolo & Omurtag, 2018; Sai et al., 2019; Schulz et al., 2012; Tu et al., 2016; Vatankhah et al., 2013; Vijayakumar et al., 2017), eight implement multiclass classification (Alazrai, Momani, et al., 2019; Elsayed et al., 2020; Kimura et al., 2021; Nezam et al., 2021; Tripanpitak et al., 2020; Vijayakumar et al., 2017; M. Yu, Sun, et al., 2020; M. Yu, Yan, et al., 2020) and five conduct continuous prediction (Bai et al., 2016; Furman et al., 2018; L. Li et al., 2018; Prichep et al., 2018; Tu et al., 2016). The algorithms used within these studies are Support Vector Machines (SVM; Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; Kimura et al., 2021; Misra, Wang, et al., 2017; Nezam et al., 2021; Okolo & Omurtag, 2018; Sai et al., 2019; Schulz et al., 2012; Tu et al., 2016; Vatankhah et al., 2013), with one study using a Support Vector Regression (SVR; Tu et al., 2016), regression models, including linear and logistic (Bai et al., 2016; Furman et al., 2018; L. Li et al., 2018; Prichep et al., 2018), 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; T. Cao et al., 2020; Elsayed et al., 2020; Kaur et al., 2019; Tripanpitak et al., 2020; M. Yu, Sun, et al., 2020; M. Yu, Yan, et al., 2020), Linear Discriminant Analysis (LDA; Bai et al., 2016), Random Forest models (RF; Vijayakumar et al., 2017) and one study used a Mahalanobis classifier (Hadjileontiadis, 2015).

Table 3.4 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 Precision Recall F1	94.83% 93.92% 95.14% 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 Test Accuracy CV Accuracy	97.29% 90% 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 Precision[+] Recall[+] F1[+]	79.6%[++] 78.28% ± 6.03%[++] 77.03% ± 9.05%[++] 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 Misclassification	89.58% 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	Accuracy Specificity Sensitivity	83% ± 5% 91% ± 4% 93% ± 5%
						No Pain vs. Low Pain vs. Moderate Pain vs. High Pain vs. Intolerable Pain	Accuracy Specificity Sensitivity	62% ± 6% 78% ± 3% 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 represent the average across participants or condition, 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% (Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; T. Cao et al., 2020; Hadjileontiadis, 2015; Kaur et al., 2019; Okolo & Omurtag, 2018; Sai et al., 2019; Vatankhah et al., 2013; Vijayakumar et al., 2017), with eight of nine studies obtaining an accuracy greater than 85% (Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; T. Cao et al., 2020; Hadjileontiadis, 2015; Kaur et al., 2019; Okolo & Omurtag, 2018; Vatankhah et al., 2013; Vijayakumar et al., 2017). Additionally, five of the studies included in this review attempt to discern low pain and high pain classes (Bai et al., 2016; Misra, Wang, et al., 2017; Okolo & Omurtag, 2018; Schulz et al., 2012; Tu et al., 2016). 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% (Misra, Wang, et al., 2017; Tu et al., 2016). 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% (Misra, Wang, et al., 2017).

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 3.4. Nevertheless, the accuracy range for the classification of three or more pain classes is between 62% and 100% (Alazrai, Momani, et al., 2019; Elsayed et al., 2020; Kimura et al., 2021; Nezam et al., 2021; Tripanpitak et al., 2020; Vijayakumar et al., 2017; M. Yu, Sun, et al., 2020; M. Yu, Yan, et al., 2020). These results suggest that pain classification at a finer resolution is achievable, with half of the eight studies achieving accuracies between 90% and 100% (Alazrai, Momani, et al., 2019; Elsayed et al., 2020; Tripanpitak et al., 2020; M. Yu, Sun, et al., 2020).

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 (Bai et al., 2016; L. Li et al., 2018; Prichep et al., 2018; Tu et al., 2016)**,** whilst one study attempted pain prediction using a 0 to 100 scale (Furman et al., 2018). 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 (Bai et al., 2016; L. Li et al., 2018; Tu et al., 2016). 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 (Furman et al., 2018; Prichep et al., 2018).

3.3.4 Pain Phenotypes

The characteristics of the 15 phenotyping studies are reported in Table 3.5. 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 (Levitt et al., 2020). 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 3.5.

Six of the 15 phenotyping studies attempt to predict migraine phenotypes (Akben et al., 2012, 2016; Z. Cao et al., 2018; De Tommaso et al., 1999; Frid et al., 2020; Subasi et al., 2019). Within these six studies, four classified migraine versus healthy controls (Akben et al., 2012, 2016; De Tommaso et al., 1999; Subasi et al., 2019), one classified migraine with aura versus migraine without aura (Frid et al., 2020) and one classified the interictal phase versus the preictal phase of migraine (Z. Cao et al., 2018). Furthermore, five of the 15 studies predicted neuropathic or neurogenic pain (Saif et al., 2021; Sarnthein et al., 2006; Vanneste et al., 2018; Vuckovic et al., 2018; Wydenkeller et al., 2009). Four of the five studies above predicted the presence of neuropathic pain or neurogenic pain versus healthy controls (Saif et al., 2021; Sarnthein et al., 2006; Vanneste et al., 2018; Vuckovic et al., 2018), and one study classified neuropathic patients into two groups: pain below the lesion versus without pain below the lesion (Wydenkeller et al., 2009). Furthermore, one study classified a broad group of chronic pain patients versus healthy controls (Ta Dinh et al., 2019). 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 (Paul et al., 2019). Moreover, one study classified radiculopathy versus healthy controls (Levitt et al., 2020). 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 (Graversen et al., 2011). SVM was the most common algorithm (Akben et al., 2016; Z. Cao et al., 2018; Frid et al., 2020; Levitt et al., 2020; Paul et al., 2019; Saif et al., 2021; Ta Dinh et al., 2019; Vanneste et al., 2018) including SVR (Graversen et al., 2011), whilst ANN (Akben et al., 2012; De Tommaso et al., 1999), discriminant analysis (Sarnthein et al., 2006; Vuckovic et al., 2018; Wydenkeller et al., 2009) and RF models (Subasi et al., 2019) were also used.

Table 3.5 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 Sensitivity Specificity	93.33% 93.33% 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 Sensitivity Specificity	88.40% 90% 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 Sensitivity (Recall) Precision (PPV) F1	76% ± 4% 75% ± 5% 75% ± 5% 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 FPR	95.83% 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 AUC	84.62% 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	82.50%
							AUC	0.8225
	Multiclass					Healthy Control vs. Radiculopathy vs. Pre-SCS	Accuracy	71.90%
							AUC (Radiculopathy)	0.828
							AUC (Healthy)	0.842
							AUC (Pre-SCS)	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	96.15%
							Sensitivity	96.88%
							Specificity	95.65%
							Precision (PPV)	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	Accuracy	99% ± 0.49%
						Healthy Control vs. PNP	Accuracy	97% ± 0.6%
						PWP vs. PNP	Accuracy	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	87%[††]
							CI	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	85.95%
							Sensitivity	85.20%
							Specificity	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 Sensitivity Specificity	57% ± 4% 60% ± 5% 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 Sensitivity (TPR) FPR AUC	92.53% ± 1.59% 93% ± 2% 21% ± 2% 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] Sensitivity [95 CI] Specificity [95 CI]	88% ± 10% [86%-89%] 89% ± 7% [88%-90%] 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 (Akben et al., 2012, 2016; Z. Cao et al., 2018; De Tommaso et al., 1999; Frid et al., 2020; Subasi et al., 2019). 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 (Akben et al., 2012, 2016; Z. Cao et al., 2018; Frid et al., 2020; Subasi et al., 2019), five report sensitivity (Akben et al., 2012, 2016; Z. Cao et al., 2018; De Tommaso et al., 1999; Subasi et al., 2019) and three report specificity (Akben et al., 2012, 2016; Subasi et al., 2019). 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 (Graversen et al., 2011; Levitt et al., 2020; Paul et al., 2019; Saif et al., 2021; Sarnthein et al., 2006; Ta Dinh et al., 2019; Vanneste et al., 2018; Vuckovic et al., 2018; Wydenkeller et al., 2009). 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 (Paul et al., 2019; Ta Dinh et al., 2019; Vanneste et al., 2018; Vuckovic et al., 2018) and three studies (Paul et al., 2019; Ta Dinh et al., 2019; Vuckovic et al., 2018), 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 (Graversen et al., 2011; Paul et al., 2019; Saif et al., 2021; Sarnthein et al., 2006; Vanneste et al., 2018; Vuckovic et al., 2018).

3.3.5 Response to Treatment

The characteristics of the seven treatment response studies are reported in Table 3.6. Two of the six studies classified active treatment or placebo conditions (Graversen et al., 2012, 2015), whilst a further four predicted whether treatment was successful (Gram et al., 2015, 2017; Grosen et al., 2017; Wei et al., 2020). The final study for the response to treatment conducted a continuous prediction to assess the change in the brief pain inventory score after medication (Hunter et al., 2009). The models used within the response to treatment studies include SVMs (Gram et al., 2015, 2017; Graversen et al., 2012, 2015), regression models, including linear and logistic (Grosen et al., 2017; Hunter et al., 2009) and a k-nearest neighbours algorithm (Wei et al., 2020)

Table 3.6 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 PPV NPV	71.90% 70% 75%
						Responders vs. Non-Responders (Response to Opioid; Placebo day)	Accuracy PPV NPV	71.90% 75% 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 PPV NPV	65% 76% 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 SE CI	1.18[++] 0.09 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% (Gram et al., 2015, 2017; Wei et al., 2020). Here, two studies achieved a positive predictive value (PPV) and negative predictive value (NPV) between 70% and 76% and 53% and 75% (Gram et al., 2015, 2017), 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% (Wei et al., 2020). Regarding the classification of active treatment versus placebo groups, the two studies achieved accuracies between 85.7% and 95.24% (Graversen et al., 2012, 2015). The remaining two studies used regression models to predict treatment response (Grosen et al., 2017; Hunter et al., 2009).

3.4 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 (Davis et al., 2020; Mackey et al., 2019; van der Miesen et al., 2019). 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% (Alazrai, Momani, et al., 2019; Alazrai, AL-Rawi, et al., 2019; T. Cao et al., 2020; Hadjileontiadis, 2015; Kaur et al., 2019; Okolo & Omurtag, 2018; Sai et al., 2019; Vatankhah et al., 2013; Vijayakumar et al., 2017). Regarding multiclass prediction, five out of eight studies demonstrated an accuracy of over 85% (Alazrai, Momani, et al., 2019; Elsayed et al., 2020; Tripanpitak et al., 2020; Vijayakumar et al., 2017; M. Yu, Sun, et al., 2020), with two surpassing 97% (Tripanpitak et al., 2020; M. Yu, Sun, et al., 2020). Furthermore, continuous pain ratings can be predicted with an error of approximately 10% on a 10- or 11-point rating scale (Bai et al., 2016; L. Li et al., 2018; Tu et al., 2016). 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% (Akben et al., 2012, 2016; Graversen et al., 2011; Paul et al., 2019; Saif et al., 2021; Sarnthein et al., 2006; Subasi et al., 2019; Vanneste et al., 2018; Vuckovic et al., 2018). Regarding migraine, all relevant studies achieved over 75% accuracy, sensitivity and specificity, with four and three studies surpassing 85% sensitivity (Akben et al., 2012, 2016; De Tommaso et al., 1999; Subasi et al., 2019) and specificity (Akben et al., 2012, 2016; Subasi et al., 2019), respectively. Moreover, regarding the prediction of pain conditions relative to controls, six of nine studies achieved accuracies of over 85% (Graversen et al., 2011; Paul et al., 2019; Saif et al., 2021; Sarnthein et al., 2006; Vanneste et al., 2018; Vuckovic et al., 2018). Additionally, three studies demonstrated a sensitivity of over 85% (Paul et al., 2019; Vanneste et al., 2018; Vuckovic et al., 2018), with one demonstrating a sensitivity of almost 97% for detecting individuals with fibromyalgia relative to healthy controls (Paul et al., 2019). 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% (Gram et al., 2015, 2017; Wei et al., 2020), whilst treatment and placebo groups can be predicted with accuracies greater than 85% (Graversen et al., 2012, 2015). 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 (Senn, 2003; Snapinn & Jiang, 2007). 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 (Gram et al., 2015, 2017). 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 (Ginsburg & McCarthy, 2001; Gram et al., 2015, 2017). 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 (Akben et al., 2016; Dansie & Turk, 2013; Pryse-Phillips et al., 1997), with many tools failing in those who cannot accurately communicate their pain, such as individuals with dementia (Breivik et al., 2008; Herr et al., 2011). 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 (Kimura et al., 2021), which demonstrates the potential for 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 (Davenport & Kalakota, 2019; Davis et al., 2017). For example, ML may allow clinicians to identify information in brief appointments, which currently cannot be achieved (Davis et al., 2020); 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 (Akben et al., 2016; Pryse-Phillips et al., 1997), 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 (Gram et al., 2017).

Should these algorithms be clinically validated, they could be applied throughout the clinical process, providing this is implemented ethically (Davis et al., 2012, 2017). ML should not replace clinicians but instead, be used as an additional tool; automating routine tasks and increasing time with patients (Ahuja, 2019).

The promise of ML is exciting but not without challenges. Evidence demonstrating that ML significantly improves patient care is scarce (Mateen et al., 2020). Consequently, substantial clinical implementation is unlikely until the end of the decade (Davenport & Kalakota, 2019). 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 (van der Miesen et al., 2019). It is unknown whether models trained on lab-based samples are ecologically valid and generalise to other samples or clinical settings (Davis et al., 2017; van der Miesen et al., 2019). 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 (Vabalas et al., 2019). 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 (Vabalas et al., 2019). 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 (Wolff et al., 2019). Developing and validating a model on the same participants is not appropriate evidence for potential clinical applications (Ramspek et al., 2021). Therefore, to demonstrate sufficient evidence for clinical translation, extensive external validation is imperative (Bleeker et al., 2003). 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 (Wolff et al., 2019). 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 (Collins et al., 2015). 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 (Lundervold & Lundervold, 2019; M. Yu, Sun, et al., 2020), could be a potential solution. Only one study reviewed implemented CNNs, achieving 97% accuracy in a three-class paradigm (M. Yu, Sun, et al., 2020). In other medical fields such as skin cancer detection, CNNs demonstrate comparable accuracy to experts (Esteva et al., 2017). However, CNNs are complex to interpret; hindering clinical applications (Rudin, 2019). 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 (Heus et al., 2018; Nagendran et al., 2020; W. Wang et al., 2020; Yusuf et al., 2020). A recent systematic review exploring the reporting quality of ML for medical diagnosis demonstrated that many studies lack sufficient details; hindering interpretation and replication (Yusuf et al., 2020). 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 (Mateen et al., 2020); 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 (Vollmer et al., 2020). Additionally, the recent extensions to CONSORT and SPIRIT guidelines to include AI studies (Cruz Rivera et al., 2020; Liu et al., 2019, 2020) 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 (Collins & Moons, 2019), 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 (Bauder & Khoshgoftaar, 2018; Holder et al., 2017; J. M. Johnson & Khoshgoftaar, 2019). 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.

3.5 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.

Chapter 4:

External validation of binary machine learning models for pain intensity perception classification from EEG in healthy individuals

Tyler Mari[1], Oda Asgard[1], Jessica Henderson[1], Danielle Hewitt[1], Christopher Brown[1], Andrej Stancak[1], Nicholas Fallon[1]

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

This two-study experiment aims to externally validate machine learning classification algorithms and electroencephalography for pain intensity classification.

This paper was published in Scientific Reports (2022). DOI: https://doi.org/10.1038/s41598-022-27298-1
The format of the text has been modified to match the style of this thesis.

The roles of the co-authors are listed below:

Tyler Mari: Methodology, Formal analysis, Visualisation, Writing – Original Draft, Writing – Review and Editing. **Oda Asgard:** Investigation, Writing – Review and Editing. **Jessica Henderson:** Investigation, Writing – Review and Editing. **Danielle Hewitt:** Investigation, Writing – Review and Editing. **Christopher Brown:** Supervision, Writing – Review and Editing. **Andrej Stancak:** Supervision, Writing – Review and Editing. **Nicholas Fallon:** Conceptualisation, Methodology, Formal analysis, Supervision, Writing – Original Draft, Writing – Review and Editing.

Abstract

Discrimination of pain intensity using machine learning (ML) and electroencephalography (EEG) has significant potential for clinical applications, especially in scenarios where self-report is unsuitable. However, existing research is limited due to a lack of external validation (assessing performance using novel data). We aimed for the first external validation study for pain intensity classification with EEG. Pneumatic pressure stimuli were delivered to the fingernail bed at high and low pain intensities during two independent EEG experiments with healthy participants. Study one (n=25) was utilised for training and cross-validation. Study two (n=15) was used for external validation one (identical stimulation parameters to study one) and external validation two (new stimulation parameters). Time-frequency features of peri-stimulus EEG were computed on a single-trial basis for all electrodes. ML training and analysis were performed on a subset of features, identified through feature selection, which were distributed across scalp electrodes and included frontal, central, and parietal regions. Results demonstrated that ML models outperformed chance. The Random Forest (RF) achieved the greatest accuracies of 73.18, 68.32 and 60.42% for cross-validation, external validation one and two, respectively. Importantly, this research is the first to externally validate ML and EEG for the classification of intensity during experimental pain, demonstrating promising performance which generalises to novel samples and paradigms. These findings offer the most rigorous estimates of ML's clinical potential for pain classification.

4.1 Introduction

Establishing an accurate assessment of subjective pain intensity is imperative for the diagnosis, prognosis and treatment of chronic pain conditions (Bendinger & Plunkett, 2016; Fillingim et al., 2016). Current pain assessment methods are contingent on self-report measures, which are not appropriate for individuals who are unable to communicate their pain precisely or entirely, such as those with dementia (Breivik et al., 2008; Herr et al., 2011), disorders of consciousness (e.g., coma; Herr et al., 2011; Schnakers & Zasler, 2007), cognitive impairments (Arbour & Gélinas, 2014; Herr et al., 2011), non-verbal individuals (e.g., non-communicative palliative care patients; Herr et al., 2011; McGuire et al., 2016), and children (e.g., infants and neo-natal populations; Herr et al., 2011; Witt et al., 2016). Furthermore, pain is an inherently subjective and multifaceted sensory process, which is challenging to measure objectively (Bendinger & Plunkett, 2016; Breivik et al., 2008). Taken together, the complexity of accurate pain assessment, particularly in populations with a reduced capacity for self-report, demonstrates the necessity for improved objective evaluation methods.

Recent endeavours to mitigate the necessity of self-report methods have attempted to elucidate biological markers of pain intensity using neuroimaging (see Mari et al., 2022; van der Miesen et al., 2019). ML analysis of neuroimaging data further enables the identification of pain intensity biomarkers. ML refers to algorithms that identify and learn patterns from data to make predictions on novel inputs without being explicitly programmed, which is achieved using optimisation, statistical and probabilistic techniques (Jordan & Mitchell, 2015; Samuel, 1959; Vu et al., 2018). The primary aim of supervised ML is to identify a function, f, that achieves the best mapping of an input X, to an output Y (see equation 1; Jordan &

Mitchell, 2015; Osisanwo et al., 2017). To identify the optimal function, supervised ML algorithms are trained using labelled data to minimise a loss (error) function by altering internal parameters (LeCun et al., 2015; Uddin et al., 2019). Following training, the model is evaluated on novel data to assess its generalisability.

$$f : X \rightarrow Y \tag{1}$$

Pain-related neural activation forms a distributed network (e.g., neurologic signature; R. Coghill et al., 1994; Wager et al., 2013) and includes SI, SII, insula, thalamus, anterior and midcingulate cortex, prefrontal cortex, amygdala, middle frontal gyrus, cerebellum and brainstem (Duerden & Albanese, 2013; Jensen et al., 2016; A. Xu et al., 2020). In addition, different regions encode specific characteristics of pain; SI and SII encode temporal, spatial and intensity features (Bornhövd et al., 2002; Coghill et al., 1999), whilst the insula contributes to encoding stimulus salience (Wiech et al., 2010).

Regarding EEG, pain modulates cortical oscillations in theta, alpha, beta and gamma frequency bands across various cortical sites including frontal, central, parietal, temporal and occipital regions (J. A. Kim & Davis, 2021; Ploner et al., 2017; Zis et al., 2022). Altered theta oscillations (4-7 Hz) are commonly observed in resting state EEG of individuals with chronic pain (Ploner et al., 2017), e.g., in fibromyalgia syndrome patients (Fallon et al., 2018). Moreover, augmented theta oscillations have been observed during pain and touch stimulation over central and parietal regions, with larger increases during painful stimulation (Michail et al., 2016). Additionally, tonic pain stimulation is associated with decreased alpha and increased beta band power (see J. A. Kim & Davis, 2021; Ploner et al., 2017; Zis et al., 2022 for reviews). Research has demonstrated decreased global alpha and increased beta band power in response to tonic cold pain stimulation (Shao et al., 2012). Source analysis identified pain-related oscillations predominantly in prefrontal cortex, SI, SII, insular cortex and cingulate cortex (Shao et al., 2012). Recently, peak alpha frequency has been shown to reliably predict pain sensitivity (Furman et al., 2018, 2020). Finally, gamma oscillations over SI have been shown to predict subjective pain intensity (Gross et al., 2007; Zhang et al., 2012) and stimulus intensity (Gross et al., 2007). Consequently, EEG features may be used as a neural marker of pain intensity.

Previous research has successfully implemented ML to identify pain intensity using EEG (Mari et al., 2022). Our recent systematic review demonstrated that EEG and ML could discriminate the presence or absence of pain with accuracies between 82.73 and 95.33% and predict pain intensity with accuracies between 62 and 100% (Mari et al., 2022). Moreover, ML classified low and high pain intensity, with the best-performing models achieving cross-validated accuracies of up to 62%, 69.20%, 70.36%, 83.50%, 86.30% and 89.58% (Bai et al., 2016; G. Huang et al., 2013; Misra, Wang, et al., 2017; Okolo & Omurtag, 2018; Schulz et al., 2012; Tu et al., 2016). Overall, these findings demonstrate the potential of ML for identifying pain intensity in healthy individuals, with all studies performing significantly better than chance.

Specifically, Misra and colleagues (2017) used a Gaussian support vector machine (SVM) to successfully classify low and high pain using theta and gamma power over the medial prefrontal region and lower beta power over the contralateral sensorimotor region. Moreover, a naïve Bayes classifier has been used to discriminate pain intensity using single-trial laser-evoked potentials (G. Huang et al., 2013). That study found that low and high pain could be classified with accuracies greater than 80% for both within-subject and cross-subject classifications. In the same study, the continuous pain rating (0-10) was predicted with a mean absolute error of less than 2 for both within-subject and cross-subject levels. Furthermore, similar research used EEG and a random forest (RF) to classify pain intensity into 10-classes (1-10); achieving accuracies close to 90% for both within- and cross-subject classifications (Vijayakumar et al., 2017). Interestingly, the study evaluated the relative contributions of each frequency band to the classification performance and found that all frequency bands were important to the classification (delta, theta, alpha, beta, gamma), with gamma being the most important to the classification performance. Therefore, including a diverse array of frequency bands and electrode locations would likely achieve optimal classification performance.

Despite previous research demonstrating promising performance, it is unclear if these models will successfully generalise to new samples. No studies in the existing literature have reported external validation; the process of evaluating a model using novel data, collected at a different time, or geographical location, or using a different experimental paradigm (Collins et al., 2015). Previous research only assessed cross-validation performance. Cross-validation involves partitioning a single dataset into training and testing sets, such that the test set is used to estimate the model's prediction error (Fushiki, 2011). Although cross-validation is essential in model development, it can lead to overly-optimistic estimates of model performance and overfitting (where the model learns idiosyncrasies in the training, which diminishes performance on novel data; Cabitza et al., 2021; Siontis et al., 2015; Vabalas et al., 2019; Varma & Simon, 2006). Consequently, the previous research findings are potentially inflated and may not be generalisable (Mari et al., 2022), which is insufficient evidence for clinical translation (Bleeker et al., 2003; Ramspek et al., 2021). However, a recent study found that pain-free sensorimotor peak alpha frequency could correctly classify pain-sensitive individuals using an external validation paradigm (Furman et al., 2020), providing evidence that EEG and ML could be effectively combined to identify pain outcomes. Nevertheless, external validation has never been attempted for investigations of pain intensity.

The present study aimed to be the first to externally validate ML for EEG pain intensity classification, through a robust two-step process. Given the paucity of external validation research, we aimed to (1) train ML classifiers on EEG data to predict pain intensity (low, high) and evaluate the cross-validation performance, (2a) to externally validate the classifiers on data collected from a novel sample at a different time, which used identical stimulation and (2b) to externally validate the models on data obtained at a different time, which used different stimulation parameters. We conducted this multistep validation to thoroughly assess model performance and generalisability using seven well-researched supervised ML models. We hypothesised that all ML algorithms would classify pain intensity with performance metrics (accuracy and area under the receiver operating characteristics curve, hereinafter AUC) greater than chance level (≈ 50%) on (1) cross-validation and (2a) external validation one (same stimulation parameters) and (2b) external validation two (different stimulation parameters).

4.2 Methods

Two independent experiments, separated by approximately four months, were conducted. Study one was used for training and cross-validation, whilst study two was used for external validation. Moreover, study two included external validation one, which used the same stimulus parameters as study one, and external validation two, which used different parameters (external validation datasets were collected simultaneously). Both studies were processed using a similar pipeline but were managed independently to prevent data leakage (Luo et al., 2016), which could have biased the external validation. The classification was performed across all trials, pooled from every participant. The EEG data is freely available through the Open Science Framework (https://osf.io/uqt9z/).

4.2.1 Participants

Forty healthy subjects (29 female) aged between 18 and 37 years were recruited across both studies using opportunity sampling. Twenty-five participants (19 female) aged 18 – 37 years (Mean = 23.64 years, SD = 4.04) completed study one, whilst 15 participants (10 female) aged between 19 – 28 years (Mean = 22.13 years, SD = 2.95) completed study two. Both studies were temporally independent, with different participants in each study. Only one participant from study one also completed study two. Participant overlap was not a concern, as we aimed to temporally validate the ML models. The sample size was consistent with previous research (See Mari et al., 2022). All participants had normal or corrected-to-normal vision, and no neurological disorders, chronic pain disorders or acute pain at the time of participation. Participants were reimbursed £10 per hour for their time. Participants provided fully informed written consent at the beginning of both experiments. Both studies achieved ethical approval from the University of Liverpool Health and Life Sciences Research Ethics Committee. All methods in both studies were conducted in compliance with the Declaration of Helsinki.

4.2.2 Pneumatic Pressure Stimulator

For both studies, tonic pain stimulation was delivered to the finger-nail bed of the left-hand index finger using a custom-built pneumatic pressure stimulator (Dancer Design, St. Helens, UK), as utilised in previous pain research (Watkinson et al., 2013). The pneumatic stimulator consisted of a pneumatic force controller, which directed compressed air from an 11.1-litre aluminium cylinder into the stimulator, which lowered a 1cm^2 probe to deliver the desired stimulation force. The stimulator was controlled using a LabJack U3 printed circuit board for interface. The pressure was limited to a maximum of 3.5 bar (9kg/cm^2) to prevent injury.

4.2.3 Experimental Procedure

4.2.3.1 Study One

Following the EEG cap fitting, participants were seated 1-meter from a 19-inch LCD monitor inside a Faraday cage. Participants placed their left-hand index finger into an individualised mould that correctly positioned the finger underneath the stimulator probe. A thresholding procedure was employed to identify participants' pain threshold and high pain intensity stimulus. Participants were verbally instructed to rate the pain intensity of each stimulus on an 11-point visual analogue scale (0 – 10) by using the mouse in their right hand to click the desired rating. On the rating scale, 0 reflected no sensation, 3 represented pain threshold and 10 reflected extreme pain. Participants were informed that any rating below 3 represented non-painful sensations. Following the instructions, a staircase thresholding procedure was implemented. The stimulus intensity was initialised at 0.5 bar pressure and incremented in steps of 0.2 bar (0.1 if preferred at higher levels) up to a maximum of 3.5 bar. The intensity that elicited repeated responses of 6 (±1) on the 11-point scale on three successive trials was used as the high pain intensity stimulus. Moreover, the stimuli intensity that produced a repeated rating of 3 was determined as the pain threshold. Finally, an additional stimulus intensity was defined as two-thirds of the participant's pain threshold stimulus intensity and reflected non-painful touch stimulation.

During the experiment, participants were requested to focus on a fixation cross, displayed on the monitor to minimise eye movements. Each trial consisted of the stimulus delivery and the post-stimulus rating. The stimuli delivery consisted of the rise time (time for the stimulation to increase from 0 bar to the desired intensity) followed by a 3-second hold time (duration the desired stimulus was delivered). For the rise time, the stimuli increased by 1/10th of the desired pressure every 0.1 seconds (to achieve the desired stimuli after 1-second). Subsequently, the stimulus intensity was maintained for three seconds before the probe was released, and a fixation cross was presented for a rest period of 5-seconds. Participants subsequently rated the pain intensity on a 101-point visual analogue scale, using the mouse in their right hand. The scale was anchored at 0, which reflected no sensation, and 100, which represented extreme pain. The rating phase continued until the participant successfully rated the stimuli. The rating phase was followed by a 2-second rest period and instructions for participants to place their finger back into the mould if they had removed it. Participants underwent a further 2-second rest period before progressing to the next trial.

The experiment contained three blocks, lasting approximately 15-minutes each, separated by intervals of 5 – 10 minutes. Forty trials with a minimum interstimulus interval (ISI) of 16-

seconds were delivered per block, consisting of the three stimuli intensities. The stimuli were pseudo-randomised, such that no two consecutive trials consisted of the same intensity and that an equal number of stimuli were presented in each block. There were 13 trials of each of the two conditions and 14 trials of the remaining condition in each block, such that all stimuli conditions were delivered 40 times over the entire study. Consequently, a total of 120 stimuli were delivered in the experiment. Following the completion of all blocks, the EEG cap was removed, and participants were debriefed.

4.2.3.2 Study Two

Study two used similar procedures to study one but consisted of different stimulation parameters. A 2 x 2 factorial design was employed with 4 conditions: low pain fast rise time, low pain slow rise time, high pain fast rise time, and high pain slow rise time. The low and high pain intensities were determined using the same thresholding procedure as study one. The high and low pain fast rise time conditions were identical to the stimulation in study one (1-second rise, 3-second hold). For the slow rise time conditions, the speed at which the probe lowered onto the left-hand index finger was reduced, increasing the rise time to three seconds. The stimuli increased from 0 bar to the desired intensity, in 1/30th increments of the desired stimuli every 0.1 seconds, until the desired intensity was reached and maintained for 3-seconds. After each stimulus, participants rated their pain on the same 101-point rating scale as study one.

Study two was comprised of three experimental blocks, lasting approximately 20-minutes each. Blocks were separated by 5 – 10-minute intervals. The experiment consisted of 144 trials, with 48 trials with a minimum ISI of 16 seconds in each block. Blocks consisted of 12 trials of the four conditions, which were pseudo-randomised using similar randomisation as study one. On completion of the experiment, the EEG cap was removed, and participants were debriefed. Both experiments were delivered using PsychoPy2 (Peirce, 2007).

4.2.4 EEG Acquisition

EEG recordings were continuously obtained using a 129-channel EGI System (Electrical Geodesics, Inc., Eugene, Oregon, USA) and a sponge-based Geodesic sensor net. Net positioning was aligned with respect to three anatomical landmarks: two pre-auricular points and the nasion. Electrode-to-skin impedances were maintained below 50 kΩ for all electrodes throughout the experiment. A recording bandpass filter was set at 0.001 – 200 Hz, with sampling rate set at 1000 Hz. Electrode Cz was set as the reference electrode.

4.2.5 EEG Pre-processing

EEG pre-processing was performed using BESA 6.1 (MEGIS GmbH, Germany). Firstly, low- and high-pass filters were applied at 70 Hz and 0.5 Hz, respectively. Secondly, a notch filter of 50 ± 2 Hz was implemented. Oculographic and electrocardiographic artefacts were removed using principal component analysis (PCA; Berg & Scherg, 1994). Additionally, electrode channels containing large artefacts were interpolated to a maximum of 10% of channels. None of the data in either study surpassed this threshold. Finally, the data were resampled to 256 Hz. Consequently, according to Shannon Sampling Theory, the theoretical maximum frequency that could be assessed was 128 Hz in this study (sampling rate/2; Keil et al., 2022).

Although, more conservative measures recommend a minimum sampling rate of 2.5 times the maximum frequency of interest; resulting in a maximum frequency of approximately 102 Hz (Bendat & Piersol, 2011).

Spectral analyses were conducted using MATLAB 2020a (The MathWorks, Inc., Natick, Massachusetts, USA) and EEGLAB 2021.1 (Delorme & Makeig, 2004). Firstly, power spectra density (PSD) was estimated using Welch's method. The power spectra computation spanned -4 seconds to 6 seconds relative to the trial onset, in 1-second segments, shifted in 0.05-second increments. The data were smoothed using multi-taper Slepian sequences. Estimates of the PSD were computed between 1 and 70 Hz, with a resolution of 1 Hz. The relative band power change was calculated across every time point and frequency, in the entire epoch using the event-related desynchronisation (ERD) method (Pfurtscheller & Aranibar, 1979) (See Equation 2). The estimate of ERD at each datapoint (e.g., A in the equation) is calculated by subtracting the mean PSD of the baseline period (-3.5 to -0.5; R), followed by a numerical transform to give relative change in power as a percentage value.

$$ERD\ (\%) = \left(\frac{A - R}{R}\right) * 100 \tag{2}$$

Negative ERD values represent decreases of band power in the active, relative to the baseline period, indicating cortical activation, while positive values reflect band power increases, known as event-related synchronisation (ERS). For the ML analysis, ERD data were collapsed across established frequency bands theta (4 – 7 Hz), alpha (8 – 12 Hz), lower beta (16 – 24 Hz), upper beta (25 – 32 Hz) and gamma (33 – 70 Hz). Topographical maps, to illustrate power changes from baseline to both low and high pain stimulation conditions of study one are reported in the results section for illustrative purposes. ERD visualisation was conducted and reported following recommendations from previous research (Pfurtscheller & Aranibar, 1977, 1979).

4.2.6 Classification Procedure

Firstly, we identified the trials relating to low and high pain conditions. In the current study, high and low pain samples were determined by the stimulation intensity rather than the subjective rating, as this may ultimately serve as a proxy measure for subjective reporting for populations who cannot accurately report their pain intensity. Secondly, touch intensity trials from study one were removed as study two did not contain touch trials. EEG data from two participants in study one was heavily contaminated with artefacts. Both participants' data were consistently contaminated with severe artefacts (e.g., muscle movement), which could not be resolved without exclusion. No threshold was used to determine exclusions in this instance, as it was evident from visual inspection that the data was not useable. Therefore, both participants were excluded, resulting in a final population size of 23. One participant was removed from study two due to corrupted data, which affected approximately 1/3 of the data. As a result, the final population was 14 in study two. All 14 participants from study two contributed to both external validation one and two, as both datasets were collected during the same session.

Candidate features were created using the single-trial time-frequency transformed data from study one. We computed 15 candidate features for ERD outputs in each specified frequency band which were calculated over the entire trial window [-4-6s] for all 128 electrodes, resulting in 9600 candidate predictors. The features were primarily descriptive statistics of the relative band power changes in each frequency band including the mean, mode, median, minimum, maximum, standard deviation, root mean squared, variance, skewness, kurtosis, absolute mean, Shannon entropy, log energy entropy, range and squared mean values for the time window of each trial. Candidate features used in this study were selected based on previous pain research (Alazrai, Momani, et al., 2019; Sai et al., 2019), which were calculated using MATLAB built-in functions where possible. Moreover, the features used in this study have been extensively explored in other research domains (Anuragi & Sisodia, 2020; Vargas-Lopez et al., 2021; Vimala et al., 2019; Yasoda et al., 2020). We opted to include this selection of different candidate features as, due to the complexity of EEG and ML, it is challenging to predict the effectiveness of the features and algorithms prior to modelling.

Due to neural variability and volatility of single-trial EEG (Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014), missing values and outliers (values beyond three median absolute deviations) were replaced using linear interpolation. Interpolated values were calculated from neighbouring non-outlier data per condition using the *filloutliers* MATLAB function. Outliers were interpolated as they do not follow patterns, which hinders ML performance (Maniruzzaman et al., 2018). Additionally, outlier management is essential for EEG, as artefacts include non-neural activity (Fatourechi et al., 2007). The data were interpolated to maximise the dataset size, as larger datasets are less susceptible to overfitting (Vabalas et al., 2019). Overall, less than 10% (M = 9.84%, SD = 0.55%) of the data were interpolated.

The features were scaled between 0 and 1 and univariate feature selection was employed to rank feature importance. We opted for a data-driven approach, meaning that all candidate features (e.g., all electrode locations and frequency bands) were evaluated during feature selection. Following feature ranking, a form of sequential feature selection was implemented to identify the optimal number of features. Here, the models were trained and evaluated using cross-validation with only one feature initially. Features were added sequentially until performance stabilised. Through this process, the highest-ranking 50 features were selected as this combination achieved near-optimal cross-validation performance without significantly increasing model complexity. The variables identified by the feature selection algorithm were distributed across various electrode locations and included features from frontal, central, and parietal regions. The electrode locations for all frequency bands assessed are displayed in Figure 4.2. Moreover, the number of trials after pre-processing for both studies are presented in Table 4.1.

Table 4.1 Number of events per condition for each validation procedure

Condition	Training and Cross Validation Sets	External Validation One Set (Identical Stimuli)	External Validation Two Set (Different Stimuli)	Total
Low Pain	919	503	504	1926
High Pain	897	504	504	1905
Total	1816	1007	1008	3831

ML was conducted using Scikit-learn, an open-source ML library written in Python, which offers efficient implementations of many ML algorithms (Abraham et al., 2014; Pedregosa et al., 2011). We implemented an adaptive boosting algorithm (AdaBoost), linear discriminant analysis (LDA), logistic regression (LR), gaussian naïve Bayes (NB), random forest (RF), support vector machine (SVM), and an extreme gradient boosting algorithm (XGBoost; see Osisanwo et al., 2017; Sarker, 2021; Uddin et al., 2019 for overviews). Additionally, hyperparameter optimisation was performed on the cross-validation dataset using grid search, a common technique that assesses a fixed set of potential values for each hyperparameter and evaluates all possible combinations to identify the optimal configuration (Syarif et al., 2016). Grid search has been shown to improve ML performance over unoptimised parameters (Syarif et al., 2016), and previous research has implemented grid search (Levitt et al., 2020; Misra, Wang, et al., 2017). The optimal hyperparameters (except for the NB, which does not require optimisation) are presented in Table 4.3 (see Discrimination and Calibration Results).

4.2.7 Model Evaluation

Cross-validation was performed using stratified k-fold validation, whereby the dataset is divided into *k* partitions, with one partition used for validation and the remaining for training. Each model is trained *k* times, with a different validation set at each iteration, meaning all data is used for validation (Fushiki, 2011; Luo et al., 2016; Wong, 2015). Model performance is then averaged over all iterations. Stratified k-fold is advantageous over traditional k-fold as class distributions are preserved in each partition, rather than being random (Luo et al., 2016; Wong, 2015). We set the value of *k* = 10 (Fushiki, 2011). The models were also assessed using a two-stage external validation procedure. For each validation, we computed accuracy, precision, recall, F1, AUC and brier scores to assess performance (Alba et al., 2017; Assel et al., 2017; Powers, 2011; Sokolova & Lapalme, 2009) (See Supplementary Material 2 for overviews). A flow chart of the classification procedure is presented in Figure 4.1.

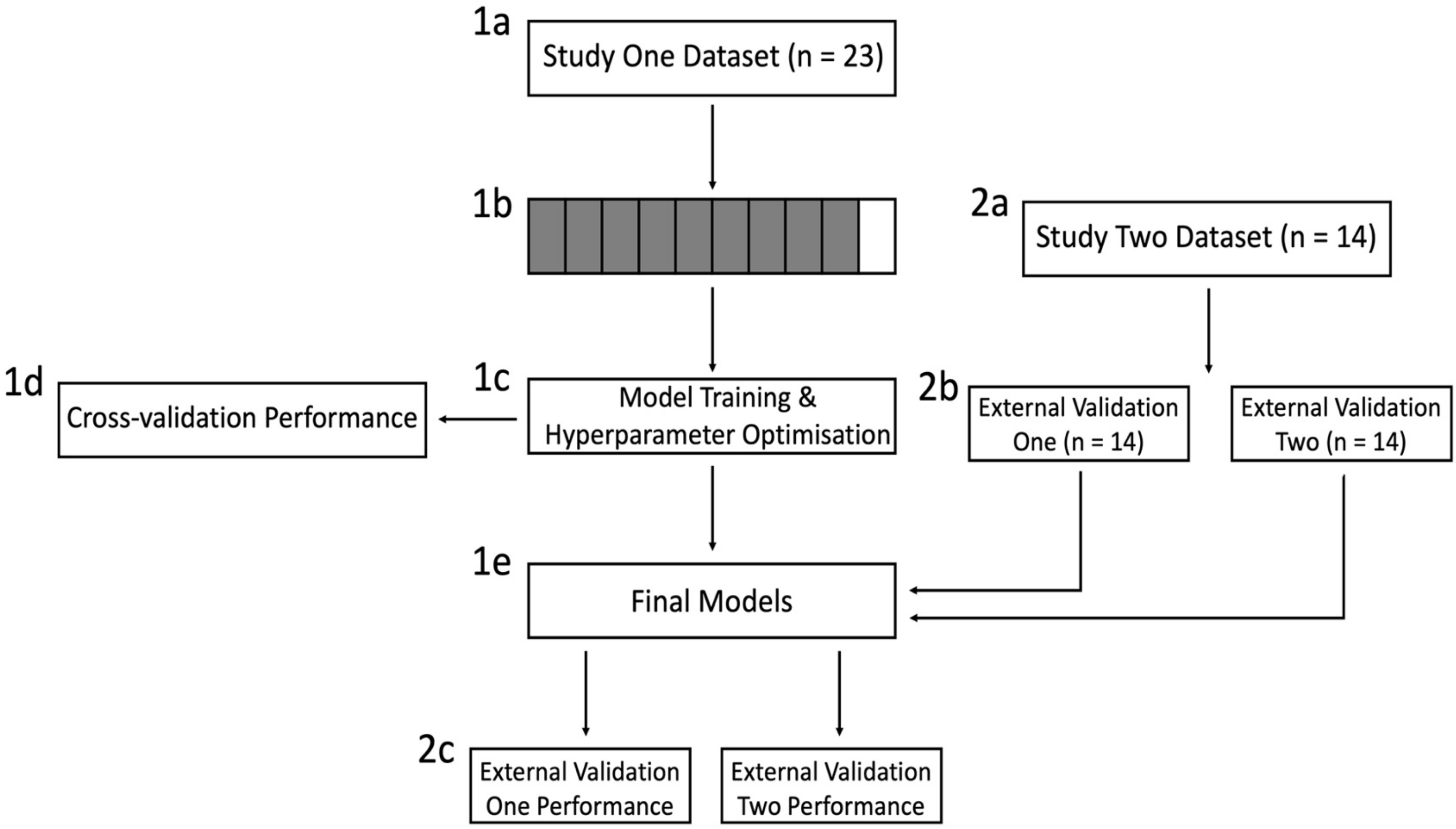

1a
Study One Dataset (n = 23)
1b
2a
Study Two Dataset (n = 14)
1d
Cross-validation Performance
1c
Model Training & Hyperparameter Optimisation
2b
External Validation One (n = 14)
External Validation Two (n = 14)
1e
Final Models
2c
External Validation One Performance
External Validation Two Performance

Figure 4.1 Flow chart of the classification pipeline. The final dataset from study one was cleaned, and features of interest were extracted (1a). The dataset, which was comprised of all 23 participants' data, was split into 10 approximately equal folds (1b), with 9 folds used for training and 1 fold used for testing. Candidate models were then trained 10 times until all folds had been used for testing. During the training process, the hyperparameters of each model were optimised using grid search (1c). After training, the models' cross-validation performance was examined (1d) and the final models and hyperparameters were selected based on the best cross-validation performance (1e). The dataset for study two was prepared using a similar pipeline (i.e., data cleaning) to study one, but was managed independently to prevent data leakage (2a). The dataset for study two was then split into external validation one and two, based on the trial types of the study (fast and slow rise) (2b). All 14 participants in study two contributed to both external validation datasets. Finally, the final models were tested separately on external validation one and two datasets, and model performance (discrimination and calibration) was assessed.

4.2.8 Calibration Assessment

We also assessed model calibration. Calibration assessment evaluates the agreement between the model's prediction and the observed or reference value (Alba et al., 2017; Luo et al., 2016; Van Calster et al., 2019). If a model predicts a 30% risk of an outcome being present, then the observed outcome frequency should be approximately 30 of 100 events (Luo et al., 2016; Steyerberg et al., 2010; Van Calster et al., 2019). For example, in a diagnostical context, in individuals with a predicted risk of *x%* for having a medical condition, *x* out of 100 individuals should have the condition (Van Calster et al., 2016). Calibration is important for model evaluation but is rarely evaluated (Christodoulou et al., 2019; Mari et al., 2022). We assess calibration using calibration curves, whereby the predicted probability is plotted on the x-axis, and the true probability is plotted on the y-axis. Perfect calibration occurs when the predicted probabilities perfectly match the observed probabilities, which is represented by a 45^{o} line in calibration curves. Comprehensive overviews of prediction model calibration assessment have been reported elsewhere (Y. Huang et al., 2020; Van Calster et al., 2019).

4.2.9 Statistical Analysis

Statistical analyses were conducted to investigate self-reported pain ratings for both studies. Firstly, a paired sample t-test assessed whether pain ratings differed between the low and high pain stimuli in study one. For study two, we assessed whether pain ratings differed between low and high stimuli and the fast and slow rise time conditions, using a 2x2 repeated measures ANOVA with the levels being stimuli intensity (low, high) and rise time (fast, slow). Statistical analysis was completed using IBM SPSS 27 (IBM Corp., Armonk, New York, USA).

4.3 Results

4.3.1 Behavioural Pain Ratings

Descriptive statistics for the behavioural pain ratings for both studies are presented in Table 4.2. A paired samples t-test demonstrated that subjective pain ratings in the high pain condition were significantly greater than those in the low pain condition in study one (*t* (22) = 12.71, *p* < .001, *d* = 2.65).

Table 4.2 Descriptive statistics (Mean ± standard deviation) for pain ratings across condition and study paradigm.

Condition	Low Pain	High Pain
Study One		
Cross-validation dataset (fast rise)	36.87 ± 13.44	62.65 ± 15.28
Study Two		
External validation one dataset (fast rise)	50.51 ± 12.96	73.53 ± 10.61
External validation two dataset (slow rise)	47.22 ± 12.55	68.77 ± 9.83

Regarding study two, a 2x2 repeated measures ANOVA demonstrated a significant main effect of stimuli intensity on subjective pain ratings (*F (1,13)* = 53.91, *p* < .001, η_p^2 = .81), with pain ratings being significantly higher in the high pain conditions compared to the low pain conditions. Additionally, the analysis demonstrated a significant main effect of rise type on subjective pain ratings (*F (1,13)* = 14.94, *p* = .002, η_p^2 = .53), with subjective pain intensity being higher in the fast rise time conditions compared to the slow rise time conditions. Finally, the ANOVA demonstrated that there was no significant interaction between stimuli intensity and rise type on subjective pain intensity (*F (1,13)* = 1.25, *p* = .284, η_p^2 = .09).

4.3.2 ERD/S

To provide an overview of the neural characteristics of pain, we provide topographical maps demonstrating the difference between high and low pain conditions. It is important to note that the ML classification process is separate from the visualisation, which is explained hereafter. Figure 4.2 shows the time-frequency changes during rest (-3.5 – -2.5 s relative to the onset of stimulation) and the active period (1 – 2 s relative to the onset of stimulation; representing a period of maximum pressure level following the completion of stimulation rise time) for study one. Topographic plots demonstrating relative band power changes in frequency bands Theta (4 – 7Hz), Alpha (8 – 12Hz), Lower Beta (16 – 24Hz), Upper Beta (25 – 32Hz), and Gamma (33 – 70Hz) are shown. The left pair of columns represent rest and active periods for the low pain condition, whilst the right pair of columns represent the high pain condition. During both low and high pain stimulation conditions, theta-band ERS was evident over anterior frontal regions. In the high pain condition, lateralised ERS over right and left temporal electrodes was also observed in theta frequency range (Figure 4.2**A**). Importantly, strong bilateral ERD in the alpha band was observed over sensorimotor regions in both low and high pain conditions (Figure 4.2**B**) with visibly stronger alpha ERD present in high-pain condition. Bilateral ERD was also evident in both lower and upper beta bands over sensorimotor regions for both pain intensity conditions. ERD is comparatively weaker for upper beta compared to lower beta (Figure 4.2**C/D**). The bilateral ERD observed with painful stimulation in alpha and beta bands is consistent with previous research (Ploner et al., 2006). Finally, for Gamma band changes, we identified bilateral ERD across temporal-parietal regions and ERS over anterior frontal electrodes for both low and high pain conditions (Figure 4.2**E**).

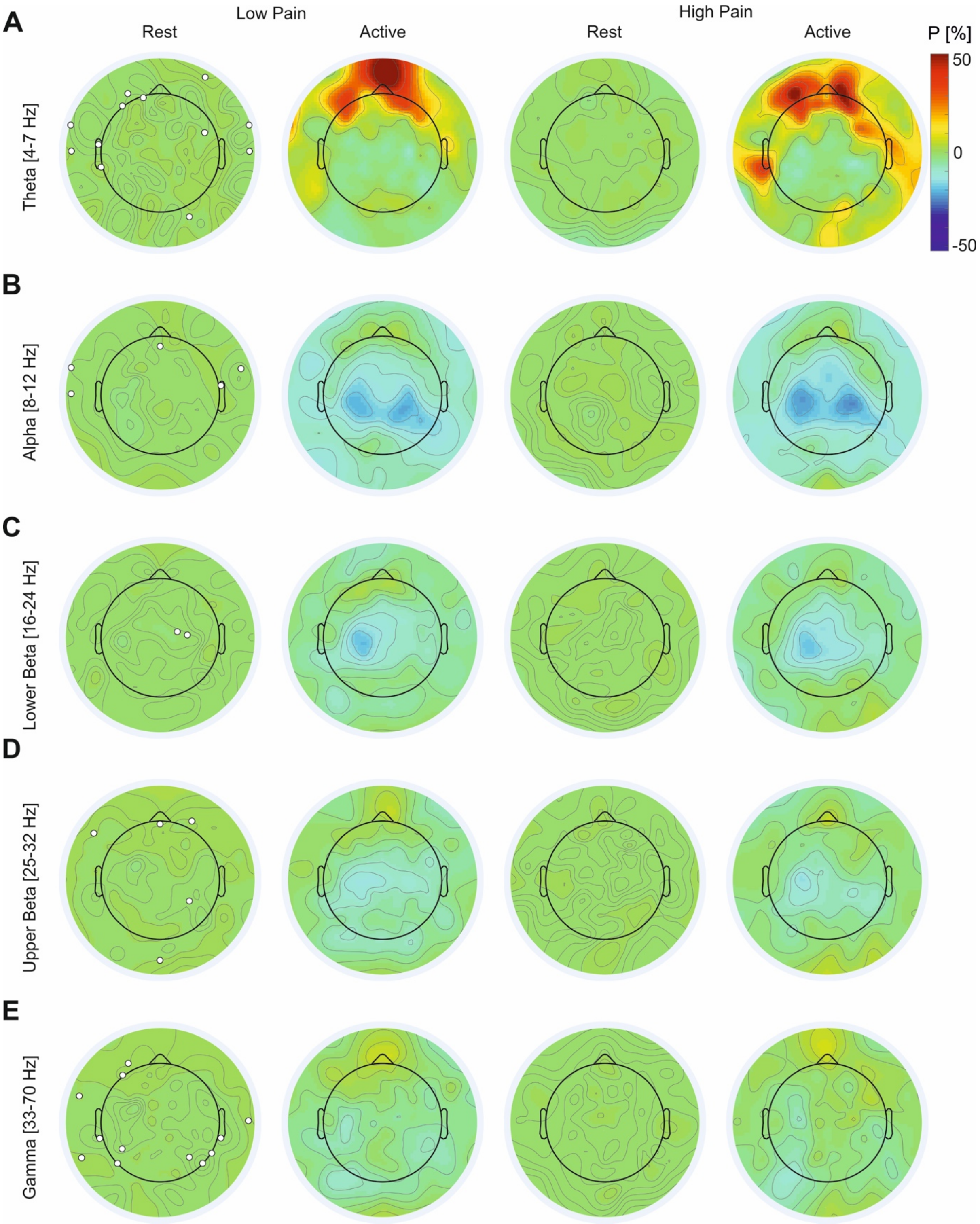


Figure 4.2 Grand average band power changes during rest (-3.5 s – -2.5 s) and during active pressure stimulation (1 s – 2 s) from study one. The trial period spanned from -4 s to 6 s relative to trial onset, with a baseline from -3.5 s to -0.5 s. The active period for visualisation was selected in line with previous recommendations (Pfurtscheller & Aranibar, 1977, 1979) and reflected 1 s of continued pressure after the stimulator reached the desired intensity level. Topographic maps show the band power changes in low and high pain intensity

conditions and from rest to active periods in Theta (**A**), Alpha (**B**), Lower Beta (**C**), Upper Beta (**D**), and Gamma (**E**) for study one. P = percentage power change from baseline. The white circles on the low pain rest plots represent the electrode locations of the features used in the ML models. Note: In the original publication, topographies were displayed in supplementary material.

4.3.3 Discrimination and Calibration Results

The classification performance metrics and optimal hyperparameters are reported in Table 4.3. The ROC curves for both external validation stages are presented in Figure 4.3. In addition, the confusion matrices are reported in supplementary material 2, allowing for the calculation of additional metrics, which may be of interest to readers and to those conducting meta-analyses.

Table 4.3 Classification performance metrics for cross validation and both external validation procedures.

Model	Optimal Parameters	Cross Validation (Mean ± SD)		External Validation One		External Validation Two	
AdaBoost	Learning rate = 0.1, Number of estimators = 2500	Accuracy	0.7732 ± 0.0374	Accuracy	0.6385	Accuracy	0.5595
		AUC	0.8644 ± 0.0199	AUC	0.6995	AUC	0.5823
		Brier	0.2450 ± 0.0011	Brier	0.2473	Brier	0.2488
		F1	0.7596 ± 0.0469	F1	0.6459	F1	0.5681
		Precision	0.7983 ± 0.0538	Precision	0.6336	Precision	0.5573
		Recall	0.7302 ± 0.0717	Recall	0.6587	Recall	0.5794
Linear Discriminant Analysis	Shrinkage = 0.4, Solver = Least squares	Accuracy	0.6965 ± 0.0249	Accuracy	0.6008	Accuracy	0.5625
		AUC	0.7707 ± 0.0307	AUC	0.6248	AUC	0.5724
		Brier	0.2007 ± 0.0135	Brier	0.2609	Brier	0.2888
		F1	0.6809 ± 0.0450	F1	0.5630	F1	0.5127
		Precision	0.7114 ± 0.0473	Precision	0.6226	Precision	0.5786
		Recall	0.6665 ± 0.1042	Recall	0.5139	Recall	0.4603
Logistic Regression	C = 1.0, Penalty = Lasso (L1), Solver = LibLinear	Accuracy	0.6910 ± 0.0301	Accuracy	0.5899	Accuracy	0.5476
		AUC	0.7676 ± 0.0283	AUC	0.6170	AUC	0.5615
		Brier	0.1990 ± 0.0108	Brier	0.2544	Brier	0.2793
		F1	0.6793 ± 0.0391	F1	0.5663	F1	0.5043
		Precision	0.7024 ± 0.0548	Precision	0.6013	Precision	0.5577
		Recall	0.6687 ± 0.0856	Recall	0.5357	Recall	0.4603

Naïve Bayes	-	Accuracy	0.7137 ± 0.0432	Accuracy	0.6395	Accuracy	0.6012
		AUC	0.8011 ± 0.0362	AUC	0.6746	AUC	0.6288
		Brier	0.2382 ± 0.0378	Brier	0.2978	Brier	0.3437
		F1	0.6806 ± 0.0807	F1	0.6142	F1	0.5830
		Precision	0.7532 ± 0.0513	Precision	0.6613	Precision	0.6109
		Recall	0.6377 ± 0.1339	Recall	0.5734	Recall	0.5575
Random Forest	Criterion = Entropy, Maximum depth = 10, Maximum features = Log$_2$, Number of estimators = 350	Accuracy	0.7318 ± 0.0556	Accuracy	0.6832	Accuracy	0.6042
		AUC	0.8129 ± 0.0392	AUC	0.6910	AUC	0.6088
		Brier	0.2008 ± 0.0100	Brier	0.2217	Brier	0.2409
		F1	0.6748 ± 0.0961	F1	0.6216	F1	0.5481
		Precision	0.8315 ± 0.0757	Precision	0.7729	Precision	0.6385
		Recall	0.5830 ± 0.1253	Recall	0.5198	Recall	0.4802
Support Vector Machine	C = 1.0, Gamma = 0.1, Kernel = RBF	Accuracy	0.6773 ± 0.0189	Accuracy	0.6187	Accuracy	0.5645
		AUC	0.7844 ± 0.0226	AUC	0.6647	AUC	0.5956
		Brier	0.1927 ± 0.0084	Brier	0.2369	Brier	0.2653
		F1	0.6669 ± 0.0454	F1	0.6265	F1	0.5675
		Precision	0.7279 ± 0.0515	Precision	0.6145	Precision	0.5636
		Recall	0.6298 ± 0.1013	Recall	0.6389	Recall	0.5714

XGBoost	Column sample by tree = 1.0, Gamma = 1.5, Maximum depth = 2, Minimum child weight = 1, Subsample = 1.0	Accuracy	0.7527 ± 0.0337	Accuracy	0.6246	Accuracy	0.5645
		AUC	0.8362 ± 0.0270	AUC	0.6770	AUC	0.5956
		Brier	0.1657 ± 0.0134	Brier	0.2336	Brier	0.2653
		F1	0.7282 ± 0.0591	F1	0.6205	F1	0.5675
		Precision	0.7922 ± 0.0405	Precision	0.6280	Precision	0.5636
		Recall	0.6845 ± 0.1019	Recall	0.6131	Recall	0.5714

The results can be segmented based on the type of validation performed. Regarding cross-validation discrimination, the results demonstrate that all the models perform better than chance on all metrics. The models achieved accuracies between 67.73 and 77.32% and AUCs between 0.7676 and 0.8644. Out of the seven models tested, four achieved accuracies greater than 70%. Moreover, the AdaBoost model achieved the best performance overall, recording the highest accuracy (77.32%) and AUC (0.8644) during cross-validation.

Regarding external validation one, the results demonstrate that the models performed better than chance on most of the performance metrics. The accuracy of the models ranged from 58.99 to 68.32%, whilst the AUC ranged from 0.6170 to 0.6995. Here, six out of the seven models achieved accuracies greater than 60%. Moreover, the RF model achieved the highest accuracy (68.32%), whilst the AdaBoost model recorded the best AUC (0.6995) on the first external validation dataset. However, it must be noted that the AdaBoost model only marginally exceeded the RF at this validation stage, with the RF achieving an AUC of 0.6910.

Lastly, for the discrimination results, the models achieved accuracies between 54.76 and 60.42% and AUCs ranging from 0.5615 to 0.6288 on external validation two. Two models (RF and NB) achieved accuracies greater than 60%. In line with the first external validation, the RF achieved the best accuracy (60.42%) on the second validation dataset, whilst the NB algorithm achieved the greatest AUC (0.6288).

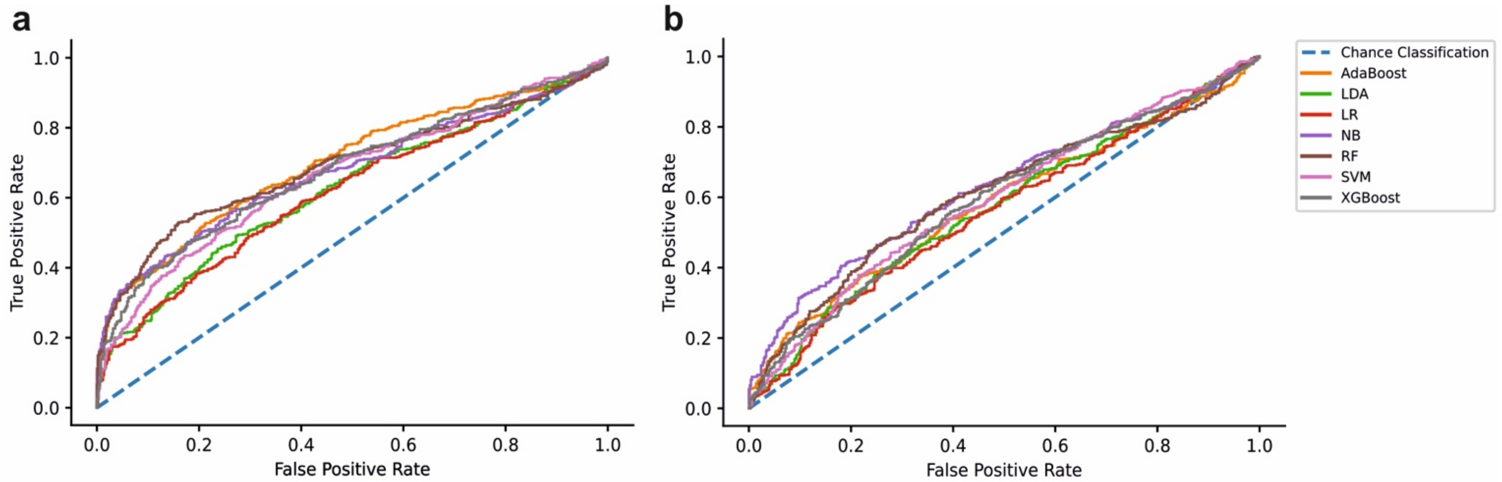


*Figure 4.3 Discrimination results for both external validation stages. (**a**) ROC curve for all models assessed on the first external validation dataset. (**b**) ROC curve assessment on the second external validation dataset. The dotted blue line represents chance classification (a classifier with no skill) as a reference.*

Finally, we also assessed the calibration of the models. The calibration plots for all models across both external validation stages are presented in Figure 4.4. Regarding the interpretation of the calibration curves, if the model line is above the reference line, it suggests that the model is underestimating the probability of the incidence, whilst the inverse insinuates that the model is overestimating the incidence prevalence. Finally, the Brier score provides a metric of the disparity between predicted and true outcome probabilities is reported in Table 4.3.

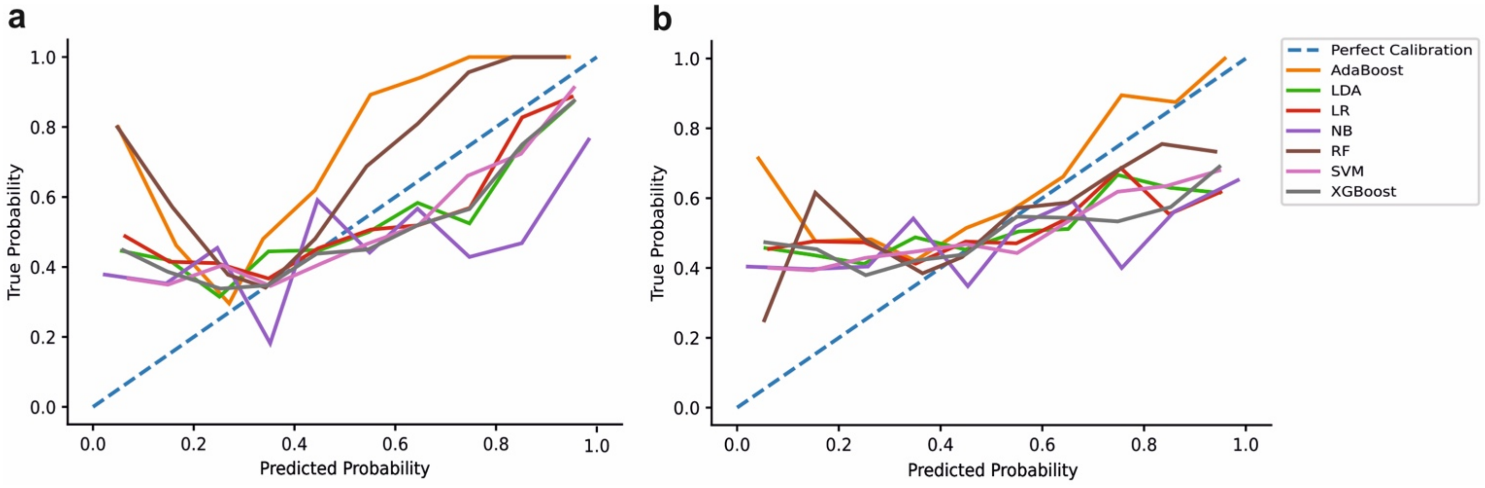


*Figure 4.4 Calibration results for both external validation stages. (**a**) Calibration curve for all models assessed on the first external validation dataset. (**b**) Calibration curve for the second external validation dataset. The blue dotted line (45^o) represents perfect calibration (complete agreement between predicted and observed probabilities). When the colour line is above the reference, the model underestimates the true probability, whilst the model overestimates probabilities when the line is below the reference line.*

4.4 Discussion

This study represents the first successful attempt to externally validate ML to discriminate between pain intensity using EEG. We hypothesised that all ML algorithms would achieve greater than chance performance (≈50%) on (1) cross-validation, (2a) external validation one (same stimulation parameters as training data), and (2b) external validation two (different stimulation parameters to training data). Our results demonstrated that all models surpassed chance performance, achieving accuracies of up to 78%, 69% and 61% on cross-validation and external validation one and two, respectively. The RF model demonstrated the highest accuracy on both external validation stages. Overall, the findings support our hypothesis. This study is the first to demonstrate that ML and EEG can be effectively combined for binary classification of pain intensity with accuracies approaching 70% using external validation. Moreover, the second external validation confirms the robustness of the results, demonstrating that ML can accurately classify experimentally induced pain intensity using different stimulation parameters, which is imperative for translation when minor variations in the nature of pain should not invalidate the algorithm. Therefore, this study advances the field, correcting widespread limitations and providing the first rigorous and generalisable estimates of the effectiveness of ML and EEG for pain intensity classification.

Our findings support previous literature demonstrating that subjective pain intensity can be accurately classified using EEG and ML (Mari et al., 2022; van der Miesen et al., 2019). The cross-validation performance in this study is comparable to previous research (Mari et al., 2022). Previous attempts to classify low and high pain intensity from EEG have produced comparable results, with accuracies ranging between 62 and 89.58% (Bai et al., 2016; G. Huang et al., 2013; Misra, Wang, et al., 2017; Okolo & Omurtag, 2018; Schulz et al., 2012; Tu et al., 2016). Similar research successfully classified 10-classes of pain intensity using a RF model and multichannel EEG (Vijayakumar et al., 2017). Our findings support the existing literature, as both studies demonstrate the importance of using a diverse array of frequency bands to achieve optimal classification performance. In addition, Huang and colleagues (2013) developed models using single-trial laser-evoked potentials, capable of accurately classifying low and high pain for both within-subject and cross-subject predictions. Alternative neuroimaging (e.g., fMRI) approaches also demonstrate promise for pain outcome prediction (van der Miesen et al., 2019). For example, the neurologic signature of pain demonstrated 93% sensitivity and specificity in discriminating between no pain and pain conditions in a novel sample (Wager et al., 2013). Overall, the previous research demonstrates the potential of neuroimaging and ML for pain intensity classification. However, EEG may prove to be the optimal method after further validation, due to the accessibility, ease of use, and low cost

(Mackey et al., 2019; Tivadar & Murray, 2019), which offers potential for the method to be used in a more diverse array of use cases.

Whilst our results are comparable to the best-performing models of the existing literature (e.g., classifying better than chance), it must be noted that several models reported across all studies had reduced performance, demonstrating the importance of careful evaluation. Moreover, the literature is comprised of positive results, which may be a result of publication bias and therefore should be carefully interpreted. In addition, previous research assessed model performance using only internal validation methods (e.g., cross-validation), meaning that overfitting and generalisability had not been sufficiently evaluated (Mari et al., 2022). Therefore, the novelty and impact of the present research stem from the extensive external validation. Presently, the clinical potential of ML and EEG for pain prediction has likely been overestimated (Bleeker et al., 2003; Ramspek et al., 2021; Vabalas et al., 2019) and significant developments are required before the clinical potential can be accurately assessed. However, although our results are modest, the current study extends upon previous research, demonstrating that ML and EEG can accurately classify novel samples which provides more robust evidence for the clinical utility of ML.

Beyond EEG, alternative proxy pain measures have been proposed (e.g., behavioural assessments). Many behavioural approaches rely on facial expressions (e.g., PACSLAC; Fuchs-Lacelle & Hadjistavropoulos, 2004) or ML techniques (Prkachin, 2009), which is time-consuming (Prkachin, 2009) and can be erroneous in individuals with dementia (e.g., Lewy Body; Oosterman et al., 2016), Parkinson's disease (Priebe et al., 2015), or facial paralysis (e.g., locked-in syndrome; Pistoia et al., 2010), as well as children who can suppress pain expressions (Larochette et al., 2006). EEG and ML may provide effective pain assessment in these challenging conditions. Pain-related neural activity is observable across populations (e.g., infants; Slater et al., 2010) and should not be affected by intentional suppression. Therefore, EEG-ML methods could become useful adjunctive pain assessment tools, specifically in situations that have previously proved challenging.

EEG-ML approaches may also prove advantageous over other pain biomarker techniques. Physiological measurements including heart rate variability (HRV), electrodermal activity (EDA), and pupillometry demonstrate potential (Cowen et al., 2015). However, such approaches also exhibit significant limitations, which often result in reduced effectiveness in certain populations (e.g., paediatric postoperative patients (Choo et al., 2010). Moreover, alternative neuroimaging techniques remain promising (e.g., fMRI; van der Miesen et al., 2019; Wager et al., 2013). However, many neuroimaging techniques are impractical for widespread clinical implementation, due to financial and infrastructure restrictions (Mechelli & Vieira, 2020). EEG is inexpensive compared to fMRI and can be easily implemented in a multitude of settings (e.g., doctor’s office) using dry or mobile EEG (Hinrichs et al., 2020; Mackey et al., 2019; Ploner & May, 2018; Tivadar & Murray, 2019). Furthermore, EEG can be used during surgery (X. Xu & Huang, 2020) and can also be further simplified using a single electrode (Kimura et al., 2021). Taken together, EEG may be advantageous over other methods, demonstrating diverse utility in clinical settings.

The findings from this study also highlight the importance of external validation, as cross-validation metrics did not consistently reflect external validation metrics, which challenges previous EEG and ML research. It is established that ML performs better on data from the same cohort (internal validation) when compared to novel samples (external validation; Cabitza et al., 2021; Siontis et al., 2015). Consequently, cross-validated metrics are potentially biased and not representative of prediction errors (Cabitza et al., 2021; Vabalas et al., 2019; Varma & Simon, 2006). In this study, the AdaBoost model achieved the best cross-validation metrics but performed worse than the RF on both external validations. As the RF performance only reduced minimally during external validation, we have increased confidence that the model has learned pain-related information, rather than fitting random noise. Furthermore, small reductions in performance when progressing from cross-to-external validation procedures are common and should not invalidate the model's clinical utility (Cabitza et al., 2021; Salehinejad et al., 2021; Siontis et al., 2015). Given the subjective nature of pain (Bendinger & Plunkett, 2016; Breivik et al., 2008) and variability of neural activity (e.g., single-trial EEG; Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014), a reduction of only 5% demonstrates the RF's robustness, providing evidence for the clinical potential of this approach. Overall, our research emphasises that failing to include external validation in experimental paradigms reduces clinical interpretation (Bleeker et al., 2003; Ramspek et al., 2021) and should be avoided in future research. We also recommend caution when interpreting research that only reports cross-validation, to avoid presenting over-optimistic results, which could hinder future efforts towards clinical translation.

Models that are not sufficiently evaluated are potentially damaging to the clinical utility of ML and EEG. A biased algorithm risks that patients could receive sub-optimal care (e.g., under-treatment), which has significant dangers (Ramspek et al., 2021; Wilson & Pendleton, 1989). Indeed, ML models failing due to biases are common and may be overlooked without sufficient validation (e.g., skin markings in dermoscopic images inflating the probability of an input being classified as a melanoma using a convolutional neural network; Winkler et al., 2019). Such biases may render the algorithm useless. Therefore, our research provides a foundational development toward clinical translation and paves the way for improved standards in ML-EEG studies for pain classification.

ML and artificial intelligence (AI) are rapidly advancing society (e.g., route planning and self-driving vehicles), but successful medical applications are rare (Seneviratne et al., 2020; Shah et al., 2019). Clinical translation requires significant developments spanning external validation to dissemination (Mechelli & Vieira, 2020). Whilst our best model is an important initial development, the performance is not currently clinically applicable. Further external validation is imperative, particularly through international multi-centre collaborations (Mari et al., 2022; Mechelli & Vieira, 2020; van der Miesen et al., 2019) to demonstrate clinically relevant performance. This would evaluate algorithms using larger, more diverse samples, allowing for greater confidence that the algorithm is not biased by dataset idiosyncrasies, which are specific to a single lab's apparatus or procedures (Mackey et al., 2019). Moreover, progression to research in clinical populations which attempts to classify clinical rather than experimental pain is critical to establish the clinical utility of the method. Subsequently, the clinical translation pipeline should be carefully navigated. Real-world and utility assessments (e.g., randomised controlled trials) should ensure the algorithm is useful to clinicians (Mechelli & Vieira, 2020; Seneviratne et al., 2020). Moreover, feasibility, safety, ethical and acceptability considerations will be essential to establish appropriate deployment standards to limit risk before dissemination (Mackey et al., 2019; Mechelli & Vieira, 2020; Seneviratne et al., 2020). However, before attempting these stages significant further research is required. Establishing a substantial body of external validation research, including multi-centre collaborations must be the primary objective. The long-term future of clinical ML applications for pain is contingent on the collective research community successfully addressing the clinical translation stages.

The current study has several limitations. Firstly, the calibration assessment demonstrated that the predicted probabilities were not consistently representative of the true probabilities. Consequently, the clinical potential of the findings at this early stage should be interpreted with caution. Imperfect calibration is suggestive of potential overfitting, reducing validation performance due to the idiosyncrasies in the training data (Van Calster et al., 2019). However, given the volatility of neural activity (Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014), it is to be expected that the models capture some random noise. As calibration is rarely assessed (Christodoulou et al., 2019; Mari et al., 2022), future research should aim to assess and improve model calibration (e.g., Platt scaling; Y. Huang et al., 2020). Moreover, whilst this study consists of two temporally independent datasets, our overall sample size is relatively small, which reduces the confidence in the results. For ML to exhibit clinical relevance, a larger, more diverse sample is required. Future research should increase sample sizes to provide more robust conclusions, which would offer substantial further evidence for clinical translation. In addition, there was some overlap between the samples, with one participant contributing to both the development and validation samples. Future research could avoid participant overlap, or specifically explore the differences between within- and cross-subject prediction. However, in the current study, both samples were temporally independent and consisted of different experimental paradigms. Therefore, participant overlap is unlikely to significantly affect the results. Moreover, although the sampling rate in this study was sufficient (sampling rate > 2.5 times the maximum frequency analysed) to retrieve gamma band frequencies and avoid aliasing issues (Bendat & Piersol, 2011), future research should maximise the sampling rate to ensure that the highest frequencies are precisely sampled.

The current study predicted stimulation intensity rather than subjective intensity, as this may ultimately serve as a better proxy method for individuals who cannot self-report their pain. However, on a trial level, there were a few instances where a low-intensity stimulus produced a high subjective response and vice-versa. Consequently, such trials may have hindered the learning algorithms' performance. Future research should investigate both subjective pain intensity and stimulus intensity. Additionally, it is possible that EEG signals used in the classification were not pain-specific, which should be explored in further research. Research has suggested that EEG responses to pain may be more directly related to stimulus saliency rather than pain perception (Iannetti et al., 2008). Moreover, whilst classifying discrete pain classes has clinical potential, predicting parametric outcomes would improve the impact of the research. The ability to accurately predict subjective pain intensity to a finer resolution would increase clinical utility. Therefore, future research should externally validate regression models to demonstrate greater clinical relevance. Concurrent attempts to improve binary classification performance are also warranted before clinical translation. Finally, although the models in this study outperformed chance, we cannot definitively state that the models are exclusively reflective of neural processing. EEG signals can often contain non-brain responses e.g., muscle movements (Goncharova et al., 2003), which could affect the results. Many of the features were from electrodes located over feasible brain regions and not exclusively from those electrodes most commonly impacted by movement artefacts such as peripheral sites (Goncharova et al., 2003), which provides confidence in the results. Moreover, model performance generalised to two external validation datasets, which included different experimental pain stimulation. Therefore, we can reasonably suggest that pain-related brain information was the predominant contributor to accurate classification. However, despite thorough artefact correction, residual non-brain activity may be present in the EEG signal. Whilst our artefact correction procedure is extensively validated, it is possible residual non-brain activity may still contribute to the features and classification. For example, whilst similar research has used prefrontal theta as a feature for pain classification (Misra, Wang, et al., 2017), we cannot rule out the possibility that residual oculographic (e.g., saccades) or facial muscle movements may also contribute to the EEG data in the present study. Therefore, we propose that the importance of the frontal theta features should be interpreted with caution. Future research should aim to explore the role of non-brain responses on EEG pain classification using additional techniques such as the characterisation of electromyographic (EMG) signals or concurrent evaluation of facial expressions. In addition, future research should investigate the impact of different pre-processing procedures on pain classification performance, with a goal to develop standardised, reproducible pre-processing.

4.5 Conclusion

This research study is the first to demonstrate that ML and EEG can be used in tandem to discriminate between low and high pain intensity using a comprehensive two-stage external validation paradigm. Our best-performing model (RF) classified low and high pain with around 70% accuracy on external validation with matched stimulation and around 60% with different experimental pain stimuli. The results presented here are a significant development for the research field, as we begin to address limitations that have hindered clinical interpretation in the past. Consequently, this study provides the current best estimates of the effectiveness of

ML and EEG for pain intensity classification. Future research should strive to build on the work presented here by consistently externally validating models, before progressing to multi-centre validation studies. Overall, the current study demonstrates the potential of ML and EEG for successful pain intensity prediction and provides the first robust estimates of ML generalisability which have eluded all previous research in this field.

Chapter 5:

Machine learning and EEG can classify passive viewing of discrete categories of visual stimuli but not the observation of pain.

Tyler Mari[1], Jessica Henderson[1], S. Hasan Ali[1], Danielle Hewitt[1], Christopher Brown[1], Andrej Stancak[1], Nicholas Fallon[1]

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

This study aimed to develop ML models using EEG data to classify the observation of pain.

This paper was published in BMC Neuroscience (2023) Doi: https://doi.org/10.1186/s12868-023-00819-y
The format of the text has been modified to match the style of this thesis.

The roles of the co-authors are listed below:

Tyler Mari: Conceptualisation, Methodology, Investigation, Formal analysis, Writing – Original Draft, Writing – Review and Editing. **Jessica Henderson:** Investigation, Writing – Review and Editing. **S. Hasan Ali:** Investigation, Writing – Review and Editing. **Danielle Hewitt:** Investigation, Writing – Review and Editing. **Christopher Brown:** Conceptualisation, Methodology, Supervision, Writing – Review and Editing. **Andrej Stancak:** Conceptualisation, Methodology, Supervision, Writing – Review and Editing. **Nicholas Fallon:** Conceptualisation, Methodology, Formal analysis, Supervision, analysis, Writing – Original Draft, Writing – Review and Editing.

Abstract

Previous studies have demonstrated the potential of machine learning (ML) in classifying physical pain from non-pain states using electroencephalographic (EEG) data. However, the application of ML to EEG data to categorise the observation of pain versus non-pain images of human facial expressions or scenes depicting pain being inflicted has not been explored. The present study aimed to address this by training Random Forest (RF) models on cortical event-related potentials (ERPs) recorded while participants passively viewed faces displaying either pain or neutral expressions, as well as action scenes depicting pain or matched non-pain (neutral) scenarios. Ninety-one participants were recruited across three samples, which included a model development group (n=40) and a cross-subject validation group (n=51). Additionally, 25 participants from the model development group completed a second experimental session, providing a within-subject temporal validation sample. The analysis of ERPs revealed an enhanced N170 component in response to faces compared to action scenes. Moreover, an increased late positive potential (LPP) was observed during the viewing of pain scenes compared to neutral scenes. Additionally, an enhanced P3 response was found when participants viewed faces displaying pain expressions compared to neutral expressions. Subsequently, three RF models were developed to classify images into faces and scenes, neutral and pain scenes, and neutral and pain expressions. The RF model achieved classification accuracies of 75%, 64%, and 69% for cross-validation, cross-subject, and within-subject classifications, respectively, along with reasonably calibrated predictions for the classification of face versus scene images. However, the RF model was unable to classify pain versus neutral stimuli above chance levels when presented with subsequent tasks involving images from either category. These results expand upon previous findings by externally

validating the use of ML in classifying ERPs related to different categories of visual images, namely faces and scenes. The results also indicate the limitations of ML in distinguishing pain and non-pain connotations using ERP responses to the passive viewing of visually similar images.

5.1 Introduction

Machine learning (ML) and EEG have demonstrated promise for predicting discrete categories of visual stimuli (e.g., objects, scenes, faces etc.; Bagchi & Bathula, 2022; Cudlenco et al., 2020; Ghosh et al., 2021; Kaneshiro et al., 2015; Stewart et al., 2014; Yavandhasani & Ghaderi, 2022; Zheng et al., 2020), subjective pain intensity in response to physical pain (Mari et al., 2022, 2023; van der Miesen et al., 2019), and response to pharmaceutical intervention (Gram et al., 2017; Graversen et al., 2012; Jaworska et al., 2019), to name but a few. Research from our group previously demonstrated that high and low pain stimuli can be predicted with approximately 70% accuracy using time-frequency analysis of EEG features distributed across the scalp (Mari et al., 2023). However, the effectiveness of ML and EEG for the classification of human facial expressions and scenes depicting pain and non-pain conditions has yet to be explored. This is despite a wealth of research demonstrating the importance of neurobiological empathic responses to observed pain, which has particular relevance to clinical, physiological, and societal domains (Decety & Jackson, 2004; Lamm et al., 2011; Singer et al., 2004; Singer & Lamm, 2009). For example, elucidating the neurobiology of empathy is important for understanding the development of empathy and for clinical conditions where empathy is reduced or absent (e.g., autism; Decety & Holvoet, 2021; Y.-T. Fan et al., 2014; Oberman et al., 2005). Moreover, from a societal perspective, understanding the neurobiology of empathy may support areas such as medical education (Preusche & Lamm, 2016). Therefore, this study aimed to address this gap by developing ML models using single-trial EEG responses during the passive observation of both facial expressions and action scenes depicting neutral and painful conditions.

Traditional ERP research studies exploring empathic responses to the observation of pain demonstrate differences in ERP amplitudes, which may enable accurate ML classification at the single-trial level. A meta-analysis of up to 36 studies demonstrated an enhanced P3 and late positive potential (LPP) during pain observation, with the maximal effect observed at central-parietal sites (Coll, 2018). Previous research by our lab demonstrated that images depicting pain scenes elicited an enhanced LPP over central-parietal regions compared to situation-matched neutral images in both healthy people and a chronic pain population (Fallon, Li, Chiu, et al., 2015). Therefore, single-trial EEG responses over central-parietal electrode sites may be an important candidate feature for the ML algorithm.

In addition to classifying EEG responses to images depicting neutral and pain conditions, we also aimed to externally validate ML for the classification of single-trial neural responses to broad categories of visual stimuli (faces versus scenes) regardless of the pain component, which to the best of our knowledge has yet to be attempted. Here, the N170 component may be the most informative feature for classification. The N170 component is an early negative waveform deflection which is maximally observed over occipitotemporal regions between 140 and 200ms after stimulus onset, peaking at approximately 170ms, which is enhanced during the observation of faces (Bentin et al., 1996; Bötzel et al., 1995). The N170 is maximal when viewing faces and is attenuated or missing in response to other stimulus categories (Bentin et al., 1996; Itier, 2004). The N170 has been reliably reproduced in stationary and mobile EEG experiments (Bentin et al., 1996; Bötzel et al., 1995; Eimer, 2000; Itier, 2004; Itier & Taylor, 2004; Johnston et al., 2015; Soto et al., 2018). Additionally, the vertex positive potential (VPP), which is a large positive potential across frontal-central regions peaking between 140ms and 180ms, is observed after the presentation of a face stimulus (Bötzel et al., 1995; Jeffreys, 1989, 1996). Given the similarity in the characteristics of the N170 and VPP, the evidence suggests that both components originate from the same neural dipole (Itier & Taylor, 2002; Joyce & Rossion, 2005). Therefore, neural responses located over occipitotemporal and frontal-central regions may enable accurate classification of face versus scene images.

Indeed, previous research has successfully combined EEG and ML to classify neural responses to visual stimuli including faces, objects, and scenes. A support vector machine (SVM) trained on EEG components over occipital electrodes has successfully classified the presence of visual objects in 7 subjects; achieving a cross-validated accuracy and AUC of 87% and 0.7, respectively (Stewart et al., 2014). Additionally, research has demonstrated that neural networks could successfully classify 40 image classes from the ImageNet database (e.g., animals, objects, food) with an average accuracy of 90.16% using EEG recorded from 6 subjects (Zheng et al., 2020). Further research exhibits comparable results in decoding neural responses to objects, scenes, human and animal bodies and faces (Bagchi & Bathula, 2022; Cudlenco et al., 2020; Kaneshiro et al., 2015; Yavandhasani & Ghaderi, 2022). Finally, an attention-based convolutional bidirectional long short-term memory network has been developed to classify EEG responses to familiar and unfamiliar faces (Ghosh et al., 2021). Using time-frequency features from pre-frontal, frontal, and temporal regions, the authors classified familiar and unfamiliar faces with an accuracy of 91.34%. Therefore, the literature suggests that EEG and ML can potentially be used to successfully decode brain responses to categories of visual stimuli.

Despite promising results, the field is not without significant limitations. ML research is often insufficiently validated, with only internal validation methods used to evaluate models. This potentially leads to inflated performance estimates, overfitting and un-generalisable models (Cabitza et al., 2021; Vabalas et al., 2019; Varma & Simon, 2006). Therefore, ML models should be evaluated using data independent of model development (Lever et al., 2016). One such approach is external validation, whereby ML performance is assessed using novel data obtained from other cohorts, facilities, and repositories or collected from a different location (geographical), time (temporal) or experimental paradigm (Cabitza et al., 2021; Collins et al., 2015). Research has demonstrated reduced performance on external validation datasets (X. Li et al., 2019; Mari et al., 2023; Siontis et al., 2015). Due to the omission of external validation, it is challenging to reasonably interpret the generalisability of existing research, as the results are potentially inflated.

The present study aimed to externally validate ML and EEG for visual stimuli decoding both across and within subjects for the first time. Firstly, we trained a Random Forest (RF) model on EEG features to classify data into either faces or scenes. Moreover, we developed two further RF models to classify EEG data into either neutral or pain classes for both scenes and faces respectively. All models were externally validated using two separate samples: cross-subject which consisted of a new cohort, and within-subject which consisted of participants from the model development sample who were recruited for a second experimental session at a later time (temporal validation). We hypothesised that the RF model would classify visual stimuli with an accuracy significantly greater than the chance level (≈ 50%) for each classification task: (1) faces – scenes, (2) scenes: neutral – pain, and (3) faces: neutral – pain for both external validation samples.

5.2 Methods

5.2.1 Participants

A total of three samples, consisting of 116 EEG sessions, were collected for this study. Forty participants (22 female; 7 left-handed) aged between 18 and 52 (Mean = 27.70 years, standard deviation {SD} = 7.43) years were recruited for sample one (model development sample/cross-validation). Sample two (cross-subject validation) consisted of 51 participants (34 female; 6 left-handed) aged between 19 and 60 (Mean = 27.63 years, SD = 9.65), whilst sample three consisted of 25 participants aged between 21 and 53 (14 female; 4 left-handed; Mean = 28.96 years, SD = 8.01). Twenty-five participants from sample one completed a second experimental session a minimum of 12 weeks after their first session (Mean = 108.68 days, SD = 10.92). This cohort represented a temporal within-subject validation sample (sample three) for the ML analysis. We aimed to recruit a large sample, particularly for external validation, to provide robust estimates of model generalisability, as small external validation datasets can also provide imprecise estimates of model discrimination and calibration (K. I. E. Snell et al., 2021). Participants provided written informed consent before participation and all methods were conducted in compliance with the Declaration of Helsinki. The study received ethical approval from the University of Liverpool Health and Life Sciences Research Ethics Committee. Eligibility criteria included: at least 18 years old, normal, or corrected-to-normal vision, no acute pain at the time of participating, no history of chronic pain, and no neurological conditions. Participants were compensated with a total of £40 for time and travel expenses. The raw data is available on reasonable request.

5.2.2 Materials

5.2.2.1 Pain Faces

In the present study, we employed a passive viewing paradigm where participants were required to observe a series of visual stimuli but were not required to respond. This differs from a free viewing task, as participants were requested to pay attention to the image, which imposes a task and is arguably not truly free viewing (A. Li et al., 2020). Here, a 2x2 factorial design was used in this study: faces (expressions) and scenes, each with two levels, namely neutral and pain. The neutral and pain faces were selected from the Delaware Pain Database (Mende-Siedlecki et al., 2020). The Delaware Pain Database is an image database that contains photographs of the faces of individuals who are displaying a painful expression (e.g., grimacing) and matched neutral controls. We selected a total of 56 faces (28 painful and 28 matched neutral images). The faces were selected using several criteria. Firstly, we aimed to broadly recreate the ethnicity and gender distribution of the UK to provide representative stimuli. A total of 22 white subjects (80%) consisting of 11 males and females, 3 Asian subjects (10%) including 2 males and 1 female and 3 black subjects (10%) consisting of 1 male and 2 females were selected, which broadly matched the racial distribution of the UK (Office for National Statistics, 2021). Within the individual categories (e.g., white males) the images with the highest pain rating were selected, providing pain was listed as the dominant emotion. The 28 neutral images were selected as the matched version (e.g., same subject) of the pain expressions. Face images were approximately 1382x925 in size. Figure 5.1A demonstrates an example of neutral and pain expressions.

5.2.2.2 Pain Scenes

Additionally, still, photograph images of action scenes depicting pain or matched non-pain scenarios (hereinafter referred to as neutral or pain scenes) were employed in the present study. The pain scene images consisted of 28 images depicting either hands or feet in scenarios that elicit pain. For example, images of a knife cutting through bread in a way that would endanger the finger (e.g., placed under the knife). Twenty-eight matched neutral scenes, which replicate the scene but did not demonstrate pain, were also used. For example, the image depicted a knife cutting through bread without endangering the finger (e.g., the finger not placed under the knife). The same distribution of ethnicities implemented in the facial expression images was applied to the pain scene images. The images were selected from a larger internal pool of photographs depending on their pain rating. A small pilot study was conducted (n = 5) to rate each of the images in terms of pain intensity. The images that elicited the highest average pain rating in the pilot study were selected for the final experiment. The images used in this study are similar to previous research (Akitsuki & Decety, 2009; Fallon, Li, & Stancak, 2015; Fallon, Li, Chiu, et al., 2015; Y. Fan & Han, 2008; Han et al., 2008). Pain scene images were 774x518 in size. Figure 5.1B demonstrates examples of neutral and pain scene images used in this study.

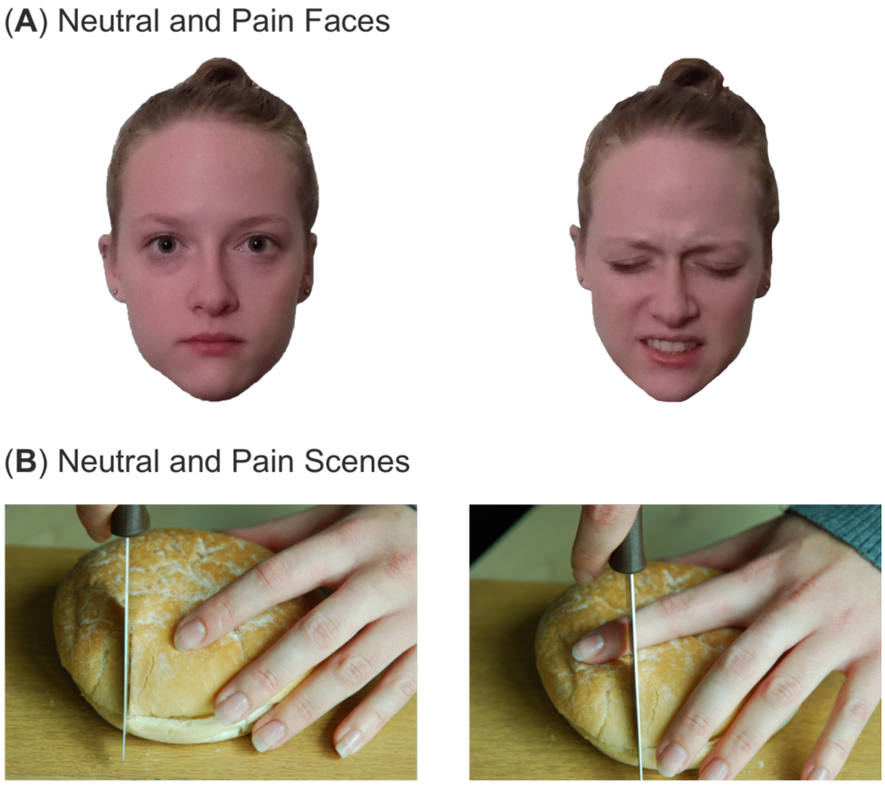


Figure 5.1 (**A**) Example of neutral and pain face stimuli from the Delaware Pain Database (Mende-Siedlecki et al., 2020). (**B**) Example neutral and pain scene stimuli.

5.2.3 Procedure

Participants attended the EEG laboratory at the University of Liverpool between June and October 2022. Following the fitting of the EEG cap, participants were seated inside a Faraday cage 1 metre away from a 23-inch 1080p LCD monitor. The experimenter verbally explained the passive viewing task and the participants' questions were answered. During this time, participants were requested to pay attention to the images and minimise movement during trials. The experiment consisted of a total of 336 trials, split into three blocks of 112 stimuli. Within each block, 28 stimuli for each of the four conditions were presented. Each block lasted 6 minutes and was separated by approximately 15-minute periods. During the block intervals, electrode impedances were checked, and additional saline solution was applied as required.

Each trial was initiated with a 2-second rest interval, where participants were shown a blank grey screen. Following the rest period, a colour photograph, that was randomly selected, was displayed for 1 second. Subsequently, the image disappeared, and the 2-second rest interval occurred before the presentation of the next image. This was repeated until all 112 images had been presented.

Following the completion of all blocks, the EEG cap was removed, and a subjective rating block was completed. Here, participants were informed that they were required to rate their perceived pain intensity of the images on a 0 – 100 scale with 0 reflecting no pain and 100 reflecting extreme pain. The rating scale included vertical bars denoting increments of 10. During the rating period, participants were presented with an image positioned above the rating scale and were required to rate the image by clicking the scale with the mouse in their right hand. The presentation of the images was randomised, and for each image, an infinite response time was employed. Once the participant had successfully rated the image, the screen was cleared, and the next image and scale were presented 100ms later. Following this, participants completed the pain catastrophizing scale (PCS; M. J. L. Sullivan et al., 1995) and were subsequently debriefed and compensated for their time and expenses.

5.2.4 EEG Acquisition

Continuous EEG recordings were acquired using a 129-channel EGI System (Electrical Geodesic Inc., EGI, now Magstim EGI, Eugene, Oregon, USA) and a sponge-based Geodesic sensor net. The net was positioned with respect to three anatomical landmarks: two pre-auricular points and the nasion. Throughout the experiment, electrode-to-skin impedances were maintained below 50 kΩ. A recording bandpass filter was applied between 0.001 – 200 Hz and the sampling rate was set at 1000 Hz. Cz was used as the reference electrode.

5.2.5 EEG Data Analysis

The data were pre-processed using the Harvard Automated Processing Pipeline for Electroencephalography (HAPPE version 3; Gabard-Durnam et al., 2018). Firstly, low-pass and high-pass filters were applied to the data at 45 and 0.1 Hz, respectively. Secondly, the data were downsampled to 500 Hz and re-referenced using the common average approach (Lehmann, 1987). Moreover, bad channel detection and interpolation were performed, and data contaminated by artefacts (e.g., oculographic) underwent wavelet thresholding (soft margin) to separate artefact and neural data. The data were then segmented into epochs of -200ms to 800ms relative to stimulus onset (500 total time points) and baseline corrected (-200ms to 0ms). Automated epoch rejection was then performed based on segment amplitude and similarity criteria. The thresholds were set at minimum and maximum segment amplitude of -150 and 150, respectively in line with HAPPE recommendations (Gabard-Durnam et al., 2018). The number of trials (mean ± SD) retained after automated trial rejection was 60.18 ± 8.44 (72% of total trials) for neutral scenes, 61.23 ± 6.19 (73%) for pain scenes, 62.93 ± 7.87 (75%) for neutral faces, and 62.15 ± 6.90 (74%) for pain faces, in sample one. In sample two, the mean number of trials remaining was 61.88 ± 5.14 (74%) for neutral scenes, 61.78 ± 6.22 (74%) for pain scenes, 62.63 ± 4.81 (75%) for neutral faces, and 62.27 ± 5.19 (74%) for pain faces. Finally, for sample three, the remaining number of trials was 62.76 ± 6.36 (75%) for neutral scenes, 60.20 ± 5.89 (72%) for pain scenes, 63.80 ± 5.97 (76%) for neutral faces, and 64.08 ± 6.49 (76%) for pain faces. Following pre-processing, the ERPs were analysed in MATLAB 2020b (The MathWorks, Inc., Natick, Massachusetts, USA) and EEGLAB 2021.1 (Delorme & Makeig, 2004). Multiple comparisons were accounted for using the false discovery rate (FDR) method. A minimum window width of 10ms was implemented to assess significant differences between the ERP waveforms.

5.2.6 Machine Learning Procedure

Following EEG pre-processing, the data were prepared for ML analysis. Each of the datasets (model development, cross-subject, and within-subject validation sample) were processed independently to prevent data leakage which could bias the external validation procedure (Luo et al., 2016). Candidate features were calculated from single-trial ERP waveforms. A total of 18 candidate features, which primarily represented descriptive statistics of the ERP waveform, were calculated for each trial between 0-800ms relative to stimulus onset. The features consisted of the mean, mode, median, minimum, maximum, standard deviation, root mean squared, variance, skewness, kurtosis, absolute mean, Shannon entropy, log energy entropy, range, mean squared, number of peaks, number of troughs, and the ratio between peaks and troughs. The features calculated in this study are comparable to previous research, both by our lab and external groups (Anuragi & Sisodia, 2020; Mari et al., 2023; Sai et al., 2019; Vargas-Lopez et al., 2021; Vimala et al., 2019). The 18 features were calculated using MATLAB functions, where possible, and were computed for each of the 129 electrodes, resulting in 2322 candidate features.

Single-trial EEG is significantly impacted by noise and variability (Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014). In line with our previous research, outlier feature values, defined as values beyond three median absolute deviations, were linearly interpolated. The interpolated values were calculated from neighbouring non-outlier data points for each condition using the MATLAB function *filloutliers* and were implemented as outliers impair the ML performance (Maniruzzaman et al., 2018). Interpolation was selected over data removal to maximise the dataset, as smaller datasets are more prone to overfitting (Vabalas et al., 2019). A total of 4.77 ± 0.49%, 5.16 ± 0.31%, and 4.74 ± 0.15% of the data were interpolated for the model development sample, cross-subject validation sample, and within-subject validation sample, respectively.

After outlier interpolation in MATLAB, all ML processing and analysis were conducted using Python and Scikit-learn (Pedregosa et al., 2011). Here, the random seed was set to 123 for all ML analyses. The features for each dataset were scaled to between 0 and 1 and univariate feature selection was conducted. All candidate features were ranked in terms of importance using F-tests and a custom sequential feature selection was implemented. Here, a baseline RF model, with no hyperparameter tuning, was developed with one feature initially. Features were sequentially added, up to a maximum of 100 features (to limit computational complexity), to identify the optimal feature configuration. The optimal number of features for each classification task (scenes - faces; scenes: neutral – pain; and faces: neutral – pain) was defined as the baseline model that achieved the best cross-validation accuracy. Stratified k-fold validation (k = 10) was used as the cross-validation procedure.

Following the identification of the optimal features, the final ML model was developed for each task. Here, a RF model was trained on the model development dataset. Hyperparameter optimisation was achieved using random search, which searches within a range of upper and lower bounds for the optimal hyperparameter values for a user-specified number of iterations (Bergstra & Bengio, 2012; Géron, 2019; L. Yang & Shami, 2020). The external validation datasets did not inform model development as this can lead to overfitting. Therefore, hyperparameter optimisation was only performed in relation to cross-validation performance. For training and cross-validation, we evaluated model performance using stratified k-fold validation (k = 10) with accuracy as the scoring function. A maximum of 5000 iterations was specified for hyperparameter tuning. Once the optimal hyperparameters were identified, the model was refitted to the entire training dataset. This resulted in the final model that was evaluated using the external validation datasets.

5.2.7 Model Evaluation: Discrimination and Calibration

The predictive capability of each model was assessed using several performance metrics for each of the validation sets (cross-validation and two external validation datasets). The primary discrimination metrics in this study were the model accuracy and area under the receiver operating characteristics curve (AUC). In addition, we also assessed model performance using alternative metrics including the Brier score, F1 score, precision, and recall. Overviews of these metrics have been reported elsewhere (Alba et al., 2017; Assel et al., 2017; Mari et al., 2022, 2023; Sokolova & Lapalme, 2009). For the external validation datasets, we calculated model performance for each subject and averaged across the entire sample to achieve both individual subject and whole sample accuracies.

In addition to model discrimination performance, we also assessed calibration for models that exceed chance discrimination performance. Prediction algorithms can be subject to bias even when the models demonstrate excellent discrimination performance (Van Calster et al., 2019). Consequently, model calibration, which evaluates the agreement between the model's predicted probability of an event compared to the reference or observed value, should be assessed (Alba et al., 2017; Luo et al., 2016; Van Calster et al., 2019). We assessed model calibration using calibration curves for both the cross-subject and within-subject validation sets, segmenting each dataset into 20 bins (see Van Calster et al., 2019). Calibration curves display the predicted probability on the x-axis and the true probability on the y-axis. Perfect calibration is represented by a 45^{o} line, whereby the predicted and observed probabilities are identical (Mari et al., 2023). Calibration has been extensively reviewed elsewhere (Y. Huang et al., 2020; Van Calster et al., 2019). Calibration assessment is only necessary when the ML models demonstrate good discrimination ability, as models with poor performance do not require additional calibration assessment (Alba et al., 2017).

5.2.8 Statistical Thresholding

Theoretically, the chance level for a binary classification task with infinite sample size is 50%. However, sample sizes are not infinite and are often small in neuroscience, resulting in variable chance levels. To quantitatively evaluate whether the ML model significantly outperformed the chance level for each subject, we implemented a statistical thresholding approach based on a binomial cumulative distribution method proposed by Combrisson and Jerbi (2015). The statistical threshold to exceed the chance level can be calculated using the following approach that applies the *binoinv* MATLAB function:

$$Statistical\ Threshold = binoinv\left(1 - \ \alpha, n, \frac{1}{c}\right) * \ \frac{100}{n}$$

Where α is the significance level, *n* is the number of trials per participant, and *c* is the number of classes.

For a given participant with *n* = 200 and *c* = 2, the model accuracy must be above 56%, 58%, and 61% to be significant at the .05, .01, and .001 levels, respectively (Combrisson & Jerbi, 2015). If the model accuracy exceeds the given threshold, the performance is significantly greater than the chance level. A minimum of 100 data samples is required to achieve comparable results to permutation testing (Combrisson & Jerbi, 2015). For all classification attempts, all subjects had more than 100 trials meaning that the use of binomial testing is acceptable. In all classifications, we use a threshold of p = 0.05. The average chance level for cross-subject and within-subject predictions was 55.20 ± 0.20% and 55.26 ± 0.24%, 57.34 ± 0.37% and 57.41 ± 0.39%, and 57.39 ± 0.36% and 57.24 ± 0.38%, for faces – scenes, scenes: neutral – pain, and faces: neutral – pain classifications, respectively. Finally, to test whether the average sample performance exceeded the average chance threshold for each sample and classification attempt, the individual subject accuracies and chance levels were compared using paired samples t-tests.

5.3 Results

5.3.1 Self-report Ratings

Descriptive statistics of the average self-report pain ratings for each of the four image types across the three samples are presented in Table 5.1. A 2 x 2 repeated measures ANOVA was conducted using IBM SPSS 27 (IBM Corp., Armonk, New York, USA) to assess the differences between participant pain ratings for the different conditions. The data from samples one (model development) and two (cross-subject validation) were combined for the analysis. There was a significant main effect of image type on the participant's perceived pain intensity ratings (F (1,90) = 19.89, $p < .001$, $\eta_p^2 = .18$), with the action scene images being rated as more painful than faces. Moreover, there was a significant main effect of pain condition (F (1,90) = 1568.26, $p < .001$, $\eta_p^2 = .95$). Here, the pain condition images received significantly higher pain ratings than the neutral condition images. Additionally, there was a significant interaction between image type and pain condition (F (1,90) = 22.10, $p < .001$, $\eta_p^2 = .20$). Post hoc paired samples t-tests demonstrated that pain ratings were significantly higher in the pain scenes condition when compared to the pain faces condition (t (90) = 4.89, $p < .001$, d = .51). There was no significant difference between pain ratings for the neutral faces or scenes conditions (t (90) = 0.68, p = .497, d = .07). Furthermore, the pain scene images had significantly higher pain ratings when compared to the neutral scene images (t (90) = 38.72, $p < .001$, d = 4.06). Finally, the pain face images received significantly higher pain ratings when compared to the neutral face images (t (90) = 31.09, $p < .001$, d = 3.26).

Table 5.1 Mean ± SD of perceived pain intensity for each condition and sample.

Sample	Neutral Scenes	Neutral Faces	Pain Scenes	Pain Faces
Development Sample	5.96 ± 8.32	4.87 ± 8.35	61.74 ± 14.04	52.63 ± 18.19
Cross-subject Validation Sample	3.80 ± 3.98	3.93 ± 5.10	63.55 ± 14.49	57.28 ± 14.80
Within-subject Validation Sample	4.87 ± 8.31	4.56 ± 8.91	61.59 ± 10.69	58.38 ± 14.84

5.3.2 ERP Analyses

Figures 5.2A, B, and C show the averaged ERP waveform from select electrodes and the scalp isopotential maps for each condition and comparison (scenes – faces, scenes: neutral – pain, faces: neutral – pain). A significantly stronger negative deflection in response to face images compared to scene images was observed over bilateral occipital-temporal electrodes during the N170 time window (142 – 214ms; peak 170ms; p < .00001). Regarding neutral and pain scene images, a significantly stronger positive deflection was observed in a cluster of central-parietal electrodes during the LPP (524 – 796ms; p < .05), peaking at 578ms. Similarly, for neutral and pain faces, a significantly enhanced P3 potential (270 – 348ms; peak 318ms; p < .05) was observed over central-parietal electrodes in the pain condition relative to the neutral condition.

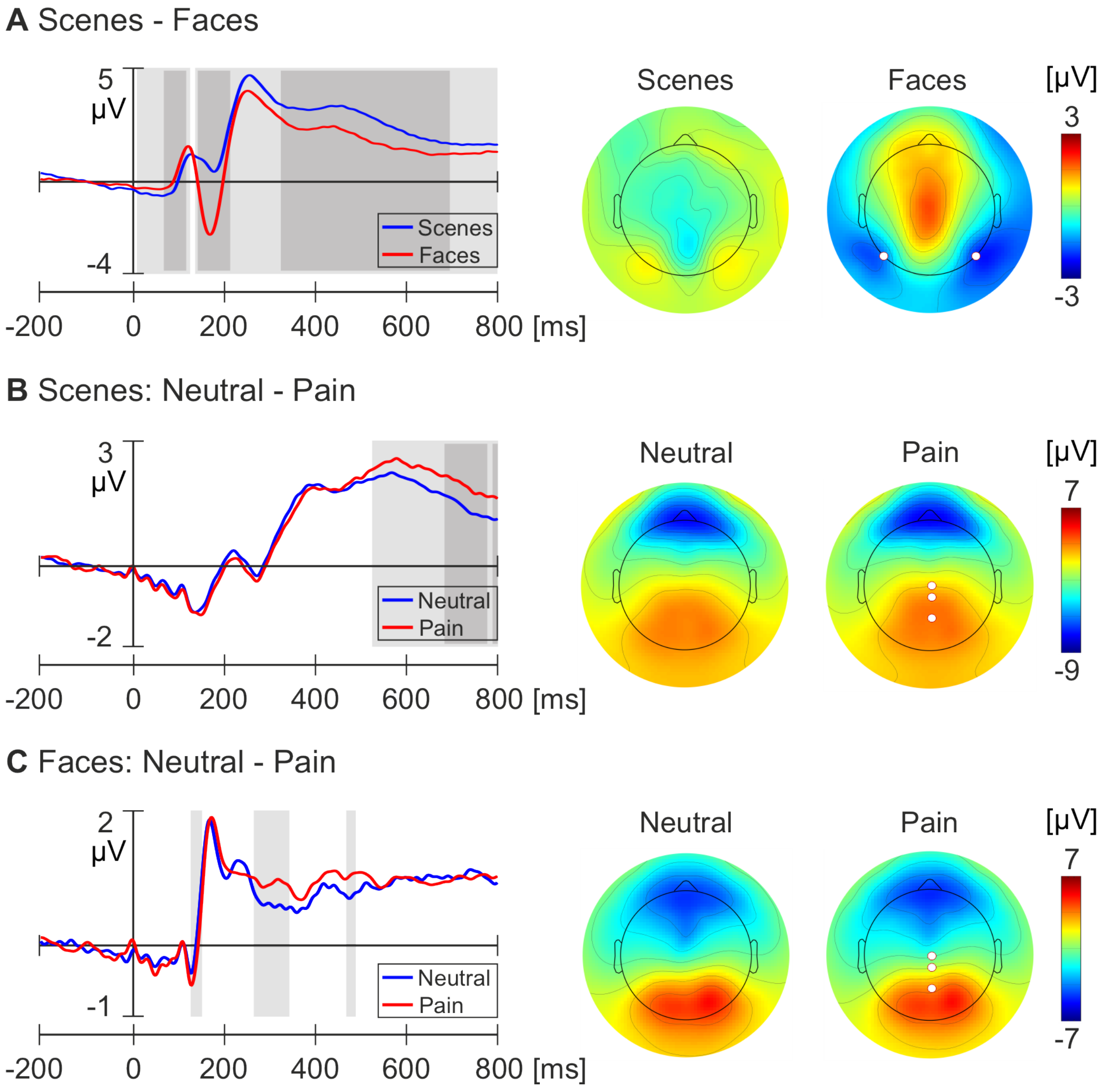


*Figure 5.2 Average ERP waveforms and scalp isopotential maps for each comparison from the unique 91 subjects within samples one and two. (**A**) Brain responses to scene and face images. Left: Average ERP waveforms from electrodes 58 (P7) and 96 (P8) for each condition. Right: Average scalp potential for each condition between 150 and 190ms. (**B**) Brain responses to neutral and pain scenes. Left: Average ERP waveforms from electrodes Cz, 55, and 62 (Pz). Right: Average scalp potential between 524 and 674ms for each condition. (**C**) Brain responses to neutral and pain face images. Left: Average ERP waveforms at electrodes Cz, 55, and 62 (Pz). Right: Average scalp potential between 270 and 348ms for each condition. White circles indicate electrode locations of the average ERP waveforms. Light grey bars denote significant differences at p < .05. Dark grey bars represent significant differences at p < .00001.*

5.3.3 Machine Learning Analyses

Following ERP analyses, the ML analysis was conducted for each of the three classification attempts. From the feature selection procedure, a total of 89, 94, and 90 features were deemed optimal for each classification task, respectively. The scalp locations of the optimal features for each of the different classification paradigms are presented in Figure 5.3. Additionally, the number of trials/observations used in the ML analysis for each condition and each sample is presented in Table 5.2.

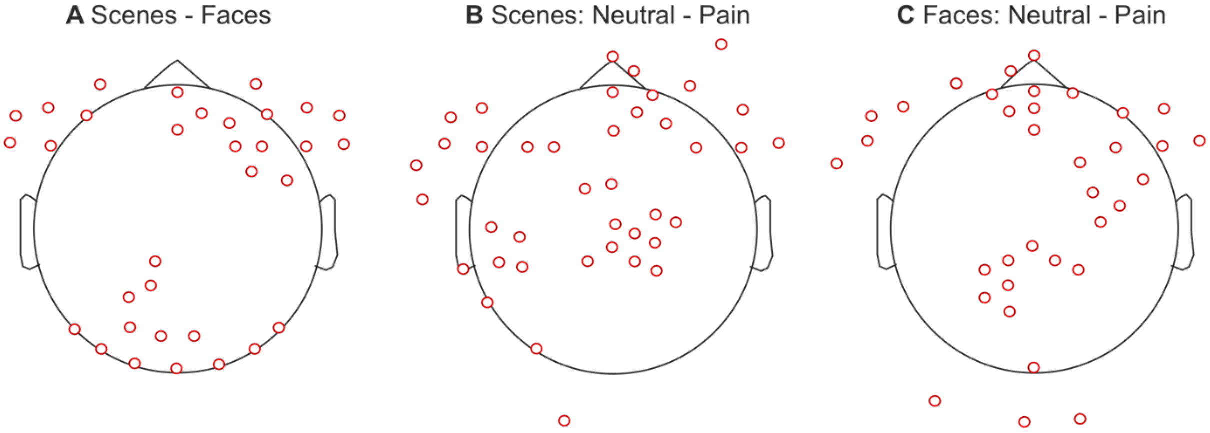


*Figure 5.3 Scalp locations of the important features determined during feature selection and model development for each classification task: scenes – faces (**A**), scenes: neutral – pain (**B**), and faces: neutral – pain (**C**).*

Table 5.2 The number of observations/trials per condition and sample used in the ML analysis.

	Scenes		Faces		
Sample	Neutral	Pain	Neutral	Pain	Total
Development/Cross-validation (n = 40)	2407	2449	2517	2486	9859
Cross-subject (n = 51)	3156	3151	3194	3176	12677
Within-subject (n = 25)	1569	1505	1595	1602	6271
Total	7132	7105	7306	7264	28807

5.3.3.1 Faces – Scenes Classification

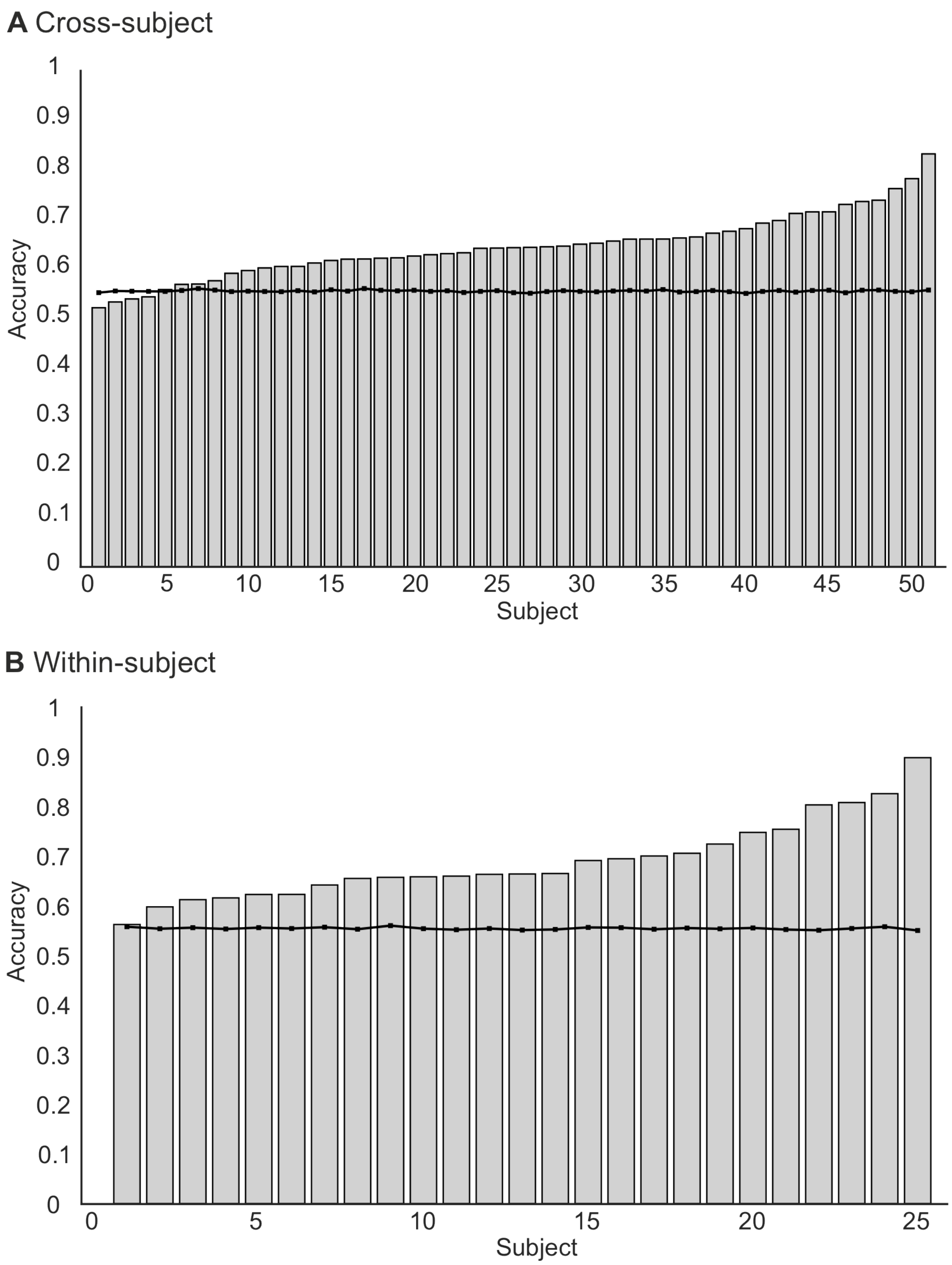
The average of each sample's classification performance metrics and optimal hyperparameters for the classification of face versus scene photographs are reported in Table 5.3. Additionally, Figure 5.4 shows the accuracies and chance thresholds for individual subjects in the cross-subject and within-subject validation samples. The average sample results demonstrate that the RF model achieved an accuracy (±SD) of 0.7456 (0.0459), 0.6415 (0.0634), and 0.6880 (0.0792) on the cross-validation and two external validation sets, respectively. Moreover, the model achieved an average AUC of 0.8189 (0.0406) on cross-validation, 0.7088 (0.0753) on cross-subject validation, and 0.7558 (0.0922) on within-subject validation. Paired samples t-tests demonstrated that the average sample accuracy was significantly greater than chance levels for the cross-subject sample (t (50) = 10.08, p < .001, d = 1.41) and the within-subject sample (t (24) = 8.46, p < .001, d = 1.69).

Table 5.3 Mean sample performance metrics for scenes - faces classification.

	Cross Validation		Cross-subject Validation		Within-Subject Validation	
Metric	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.7456	0.0459	0.6415	0.0634	0.6880	0.0792
AUC	0.8189	0.0406	0.7088	0.0753	0.7558	0.0922
Brier Score	0.1707	0.0164	0.2152	0.0253	0.1970	0.0358
F1 Score	0.7854	0.0299	0.6972	0.0460	0.7388	0.0557
Precision	0.6924	0.0495	0.6129	0.0583	0.6597	0.0959
Recall	0.9111	0.0240	0.8207	0.0890	0.8560	0.0802

Optimal hyperparameters: Number of estimators = 766, Maximum depth = 53, Minimum samples to split = 9, Minimum samples at leaf = 2, Maximum features = sqrt, Bootstrap = False.

Regarding the individual subject classification performance, the results demonstrate that the model accuracy for 47 of 51 subjects was significantly greater than the chance level ($p < .05$) for the cross-subject validation sample. Moreover, for all participants (25/25) in the within-subject sample, the model achieved accuracies significantly greater than the chance levels.



A Cross-subject

B Within-subject

*Figure 5.4 Accuracies for each individual participant for the scenes – faces classification. (**A**) Cross-subject validation dataset. (**B**) Within-subject validation dataset. The black lines denote the significance threshold for chance classification performance at p = .05.*

Finally, we also assessed model calibration for the two external validation datasets. The calibration curves for both validation stages are presented in Figure 5.5. To interpret the plots, if the model line falls above the reference line it is indicative of underestimating the probability of the outcome, whilst a line below the reference suggests the model is overestimating the probability of the event (Mari et al., 2023; Van Calster et al., 2019). The RF model for the faces versus scenes classification task generally demonstrates reasonable calibration for both cross-subject and within-subject datasets. The calibration curves follow the expected trend. Overall, the model is reasonably well-calibrated for both cross-subject and within-subject predictions.

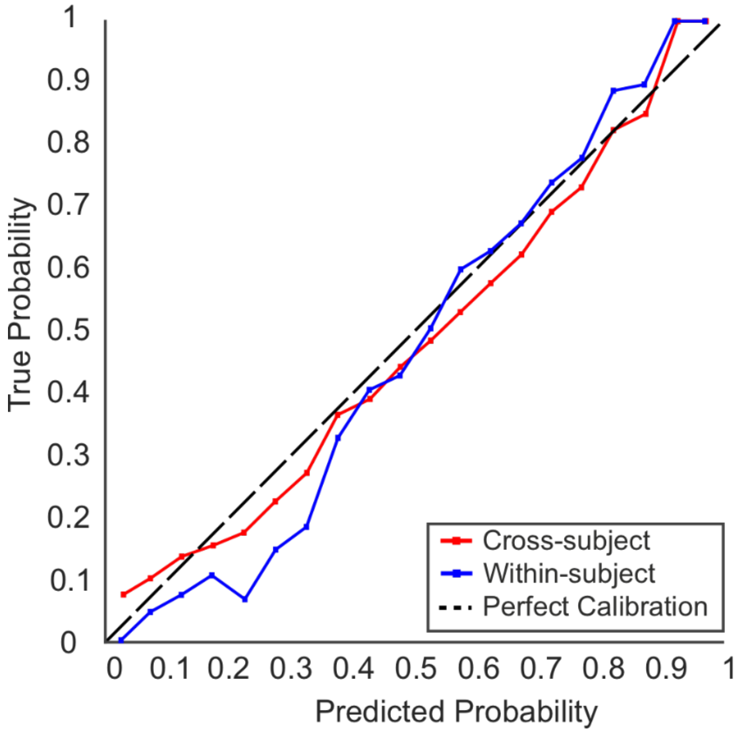


Figure 5.5 Calibration curves for both cross-subject and within-subject validation datasets for the scenes – faces classification task. The black dotted line (45°) represents perfect calibration.

5.3.3.2 Scenes: Neutral – Pain Classification

The average classification performance and optimal hyperparameters for the neutral versus pain scenes classification are reported in Table 5.4. The average accuracy (SD) was 0.8038 (0.0208), 0.2837 (0.0358), and 0.5065 (0.0504) for cross-validation, cross-subject validation, and within-subject validation, respectively. The AUCs produced a similar trend, with the evaluation procedure demonstrating an AUC of 0.8348 (0.0234), 0.2747 (0.0361), and 0.5123 (0.0518) for the three validation stages. Paired samples t-tests demonstrate that both the cross-subject (t (50) = 57.15, p < .001, d = 8.00) and within-subject (t (24) = 6.67, p < .001, d = 1.33) performance is significantly lower than the chance threshold. Regarding individual subject performance, the classification accuracy was less than the chance level for all 51 participants of the cross-subject sample. For the within-subject sample, only 2 of the 25 subjects recorded an accuracy significantly greater than the chance level. The results for individual subjects are reported in Figure 5.6. Finally, as the models do not outperform chance levels for discrimination, we do not assess calibration.

Table 5.4 Mean sample performance metrics for neutral - pain scenes classification.

	Cross Validation		Cross-subject Validation		Within-Subject Validation	
Metric	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.8038	0.0208	0.2837	0.0358	0.5065	0.0504
AUC	0.8348	0.0234	0.2747	0.0361	0.5123	0.0518
Brier Score	0.1480	0.0093	0.3966	0.0232	0.3044	0.0257
F1 Score	0.8344	0.0151	0.3866	0.0423	0.4798	0.0554
Precision	0.7277	0.0231	0.3379	0.0340	0.4960	0.0473
Recall	0.9788	0.0204	0.4553	0.0635	0.4682	0.0758

Optimal hyperparameters: Number of estimators = 735, Maximum depth = 46, Minimum samples to split = 28, Minimum samples at leaf = 17, Maximum features = sqrt, Bootstrap = False.

5.3.3.3 Faces: Neutral – Pain Classification

Finally, the average classification metrics and hyperparameters for the neural and pain faces classification are reported in Table 5.5. The results demonstrated that the RF model achieved an average accuracy (SD) of 0.6132 (0.0300), 0.5473 (0.0501), and 0.5076 (0.0383) for the cross-validation, cross-subject, and within-subject validation samples, respectively. In terms of AUC, the cross-validation AUC was 0.6717 (0.0396), the cross-subject AUC was 0.5629 (0.0667), and the within-subject AUC was 0.5241 (0.0557). Paired samples t-test indicated that the average sample accuracy was significantly lower for the cross-subject validation sample (t (50) = 3.82, p < .001, d = 0.53) and the within-subject sample (t (24) = 8.57, p < .001, d = 1.71). The individual subject accuracies for both the cross and within-subject samples are reported in Figure 5.6. Sixteen participants from the cross-subject sample and 2 participants from the within-subject sample achieved classification accuracies significantly greater than chance. As the model performance did not significantly exceed the chance threshold, we do not assess model calibration.

Table 5.5 Mean sample performance metrics for neutral - pain faces classification.

	Cross Validation		Cross-subject Validation		Within-Subject Validation	
Metric	Mean	SD	Mean	SD	Mean	SD
Accuracy	0.6132	0.0300	0.5473	0.0501	0.5076	0.0383
AUC	0.6717	0.0396	0.5629	0.0667	0.5241	0.0557
Brier Score	0.2268	0.0073	0.2523	0.0155	0.2594	0.0108
F1 Score	0.5944	0.0505	0.5046	0.1053	0.3942	0.1003
Precision	0.6216	0.0353	0.5585	0.0720	0.5182	0.0834
Recall	0.5788	0.0930	0.4932	0.1804	0.3355	0.1200

Optimal hyperparameters: Number of estimators = 161, Maximum depth = 27, Minimum samples to split = 2, Minimum samples at leaf = 4, Maximum features = log2, Bootstrap = False.

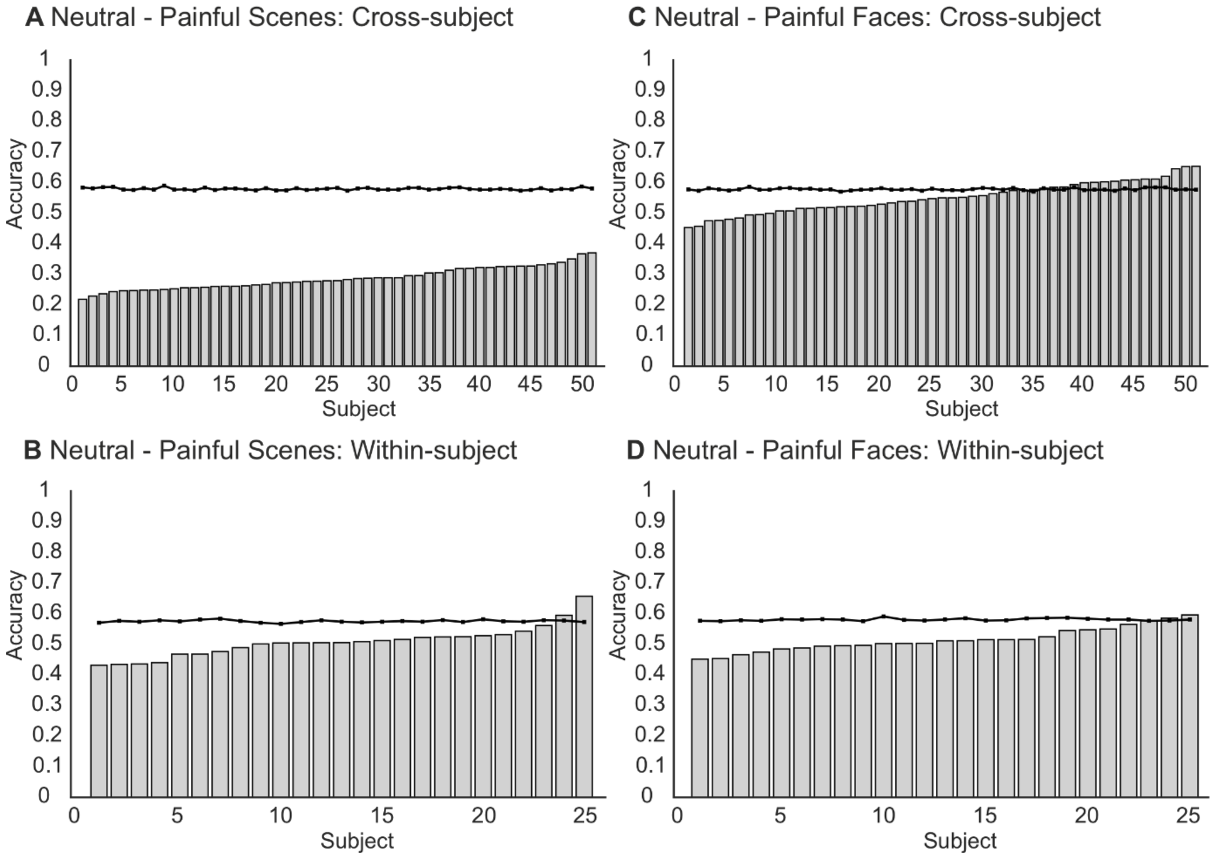


Figure 5.6 Individual subject accuracies for both cross-subject (top panels) and within-subject (bottom panels) for both scenes: neutral – pain (left panels) and faces: neutral – pain (right panels). The black lines denote the significance threshold for above chance classification performance at p = .05.

5.3.3.4 Exploratory Analysis

As the RF model was unable to significantly exceed the chance thresholds for both neutral and pain scenes and faces classification, we performed exploratory analyses to assess whether a different number of features could improve the classification performance on the external validation datasets. To assess this, we developed and evaluated 100 RF models for each classification attempt, sequentially adding features on each iteration. We initially trained the model with 1 feature and progressed to a maximum of 100 features. The model was then assessed on both validation datasets. The RF was trained using the same procedure as the other models developed in this study, but the number of iterations of hyperparameter optimisation was capped at 500 to reduce computation complexity. The mean, standard deviation, minimum, and maximum values for each of the classification tasks that did not exceed chance performance (scenes: neutral – pain and faces: neutral – pain) are reported in Table 5.6. The results of the exploratory analysis demonstrated comparable results to the original models developed. Minor performance improvements were observed, however, the model accuracy for both external validation sets remain around the chance classification level.

Table 5.6 Exploratory analysis results (accuracy) for feature combinations (1-100).

Classification	Sample	Mean	SD	Minimum	Maximum
Scenes	Cross-validation	0.7048	0.1155	0.5282	0.8186
	Cross-subject	0.3968	0.1226	0.2689	0.5362
	Within-subject	0.5063	0.0058	0.4889	0.5218
Faces	Cross-validation	0.5978	0.0014	0.5359	0.6128
	Cross-subject	0.5435	0.0063	0.5148	0.5540
	Within-subject	0.5166	0.0068	0.4952	0.5364

5.4 Discussion

We aimed to externally validate and classify single-trial EEG data elicited in response to visual stimuli using ML. Our results demonstrated that the RF model could classify images of scenes and faces with above-chance classification performance for all samples. However, the ML model could not discriminate between neutral and pain depictions of faces or scenes, achieving accuracies comparable to the chance classification rate, or lower. The results support our first hypothesis that the RF model would outperform the chance level for the scenes versus faces classification task. However, the remaining two hypotheses that the RF model would outperform chance for both cross-subject and within-subject samples on both the neutral and pain conditions for face and scene images were not supported as the model performance was significantly lower than chance on all classification attempts. Consequently, the results suggest that large broad category differences (e.g., faces – scenes) are sufficient to achieve above-chance classification performance using external single-trial EEG data. However, more nuanced differences, such as those observed in the neutral–pain classifications, cannot be used to accurately discriminate classes with novel data using the current paradigm.

Our ERP analysis demonstrated an enhanced N170 over bilateral occipital-temporal electrodes in response to face images when compared to scenes, which has been reliably demonstrated previously (Bentin et al., 1996; Bötzel et al., 1995; Eimer, 2000; Itier, 2004; Itier & Taylor, 2004; Johnston et al., 2015; Soto et al., 2018). Moreover, an increased LPP over a cluster of central-parietal electrodes was identified in the pain scene images compared to the neutral condition. Finally, an increased P3 over central-parietal electrodes was observed in response to pain faces compared to neutral expressions. The ERPs elicited in response to the empathic pain processing are also consistent with previous research (Coll, 2018; Fallon, Li, Chiu, et al., 2015). Meta-analyses of the ERP components observed during the empathic processing of painful stimuli demonstrated a positive shift in both the P3 and LPP components during the observation of painful stimuli, with the effect maximally observed over the central-parietal region (Coll, 2018). Therefore, our ERP analysis validates the data quality and experimental paradigm and replicates the effects previously reported in a comparatively large sample of healthy participants.

The findings from this study are comparable and build upon the findings of previous research which demonstrated that discrete categories of visual stimuli could be accurately classified by ML and EEG. We successfully classified images into either faces or scenes, using features predominately located across frontal-central and occipitotemporal regions, which are active during the observation of faces (e.g., N170 and VPP; Bentin et al., 1996; Bötzel et al., 1995; Jeffreys, 1989, 1996). Previous research has successfully classified neural responses to visual stimuli including faces, objects, and scenes (Bagchi & Bathula, 2022; Cudlenco et al., 2020; Ghosh et al., 2021; Kaneshiro et al., 2015; Stewart et al., 2014; Yavandhasani & Ghaderi, 2022). The present study extends the previous research by externally validating ML and EEG for image classification for both cross and within-subject prediction tasks using a large sample size. Much of the existing literature consisted of small samples (e.g., ≤ 10 subjects; Bagchi & Bathula, 2022; Cudlenco et al., 2020; Kaneshiro et al., 2015; Stewart et al., 2014; Yavandhasani & Ghaderi, 2022; Zheng et al., 2020), which are at higher risk of overfitting, resulting in potentially biased results (Arbabshirani et al., 2017; Vabalas et al., 2019). Furthermore, previous research did not rigorously assess model performance using external validation, which further increases the risk of poor generalisability (Collins et al., 2014). Therefore, the performance and utility of previous models should be interpreted with caution. In addition to generalising to external data, our classification of scenes and faces demonstrated well-calibrated estimates, which provides further evidence of an effective prediction model (Y. Huang et al., 2020; Van Calster et al., 2019). Calibration is often omitted in prediction modelling research, but it is essential to evaluating model performance (Christodoulou et al., 2019; Mari et al., 2022). Consequently, our research provides methodologically superior estimates of the effectiveness of ML and EEG for classifying visual stimuli during passive viewing. To our knowledge, we are the first to externally validate ML models for EEG visual task decoding, providing robust estimates of model discrimination and calibration, and allowing for the interpretation of model generalisability.

The current study demonstrated that ML and EEG were unable to accurately classify neutral or pain faces or scenes. We believe that the low signal-to-noise ratio of EEG and the use of a passive task may have contributed to poor classification performance. Firstly, EEG has a low signal-to-noise ratio which may have resulted in poor discriminative ability for the neutral and pain stimuli classifications (Tivadar & Murray, 2019). The N170 component offers a distinguishing characteristic between images of face and non-face classes. However, the ERP waveforms for neutral and pain images in either face or scene conditions are similar in their spatio-temporal profile, with differences mainly implicated as enhanced or augmented component fluctuations (Coll, 2018; Fallon, Li, Chiu, et al., 2015). Therefore, we can speculate that the differences at the single-trial level may be attenuated by noise and not detectable. Indeed, ML-EEG research often implements spatial filters to improve the signal-to-noise ratio and classification performance (Blankertz et al., 2008; Rivet et al., 2009). However, we opted against spatial filtering as it has a high risk of overfitting (Blankertz et al., 2008; Grosse-Wentrup et al., 2009). Alternatively, the improved signal-to-noise ratio of magnetencephalography may allow for improved classification performance (S. Singh, 2014). Moreover, the use of a passive viewing paradigm may have contributed to the classification performance. Research has demonstrated that passive viewing tasks result in reduced P300 amplitudes when compared to active viewing (Bennington & Polich, 1999), whilst other component amplitudes (e.g., LPP) are associated with, and altered by, attention and engagement (Dunning & Hajcak, 2009; Hajcak et al., 2013; Kam et al., 2014). Therefore, any further attenuation of ERPs arising from passive viewing may have hindered the ML algorithm's ability to detect patterns. Consequently, nuanced differences (such as those elicited due to empathic responses to pain) may not enable accurate classification on the single-trial level during passive viewing. It is possible that active viewing tasks (e.g., requiring image classification performed by the viewer) may improve EEG signal and consequently ML performance. However, requiring input from the subject raises questions about the usefulness of such brain decoding tools, which should preferably allow inferences on behaviour without specific behavioural requirements. Additionally, active viewing may introduce additional confounds, leading to spurious results. Research has demonstrated that stimulus properties could be decoded solely using eye movements in an active viewing task, which was not possible during passive viewing within the same sample (Thielen et al., 2019). Whilst the impact of active viewing on EEG-ML classification systems should be investigated, it is important to note that, for the method to be genuinely useful and offer novel insight, it should preferably be able to accurately classify responses during passive viewing. Overall, the inability of the ML algorithm to classify neutral and pain images likely stems from poor signal-to-noise ratio and attenuated ERP responses.

Our results highlight the importance of external validation in ML research. Without performing robust, external validation, the generalisability of the ML model cannot be effectively assessed as the results may stem from overfitting (Cabitza et al., 2021; Vabalas et al., 2019; Varma & Simon, 2006). Our cross-validation analysis of the pain scenes classification appears promising, with the model achieving an accuracy of approximately 80%. However, by implementing external validation, it was evident that the model was overfitting, achieving an accuracy below the chance level (28%) for the cross-subject dataset and comparable to chance (51%) for the within-subject validation. Therefore, through the external validation protocol, we were able to identify a model with poor generalisability, which may have otherwise been reported as an important finding. Indeed, we are not the first to demonstrate reduced performance when using an external validation (X. Li et al., 2019; Mari et al., 2023; Siontis et al., 2015), which is a significant, but often overlooked consideration when designing applied ML projects. Much of the prediction modelling research (regardless of research domain) does not assess model performance using external validation (e.g., only 5% of prediction modelling articles on PubMed report external validation in the title or abstract; Ramspek et al., 2021). Caution is advised when reporting or interpreting past ML-EEG results which have only been assessed using internal methods such as cross-validation, as the models are prone to overfitting, resulting in inflated, un-generalisable performance metrics (Cabitza et al., 2021; Siontis et al., 2015; Varma & Simon, 2006). Overall, our study highlights the importance of robust evaluation procedures when using ML, to minimise the risk of a new replication crisis (Hutson, 2018).

The present study has several limitations. Firstly, we used a passive viewing experimental paradigm, which may have resulted in attenuated ERP responses (Bennington & Polich, 1999).

Whilst we observed significant differences in both the P3 and LPP components in response to neutral and pain images, the differences between the conditions on a single trial level may have not been preserved due to the reduced neural responses associated with passive viewing, the low signal-to-noise ratio, and single-trial variability which may have contributed to poor ML performance (Blankertz et al., 2011). Additionally, informal feedback from participants indicated that the passive viewing task was perceived as 'boring', which may have reduced attention, further impacting the neural responses (Dunning & Hajcak, 2009; Hajcak et al., 2013; Kam et al., 2014). Therefore, passive viewing may not be appropriate to elicit adequate responses that are detectable using ML at the single trial level using the approach outlined in the present study. Future research should implement active viewing paradigms and assess ML performance to build on our findings. For example, a two-alternative forced choice paradigm whereby participants are required to determine the presence or absence of pain may be more suitable for ML classification than passive viewing tasks. Similar forced choice tasks within pain empathy research have been widely reported (Coll, 2018). Secondly, whilst the images in the study were similar to previous research (Fallon, Li, & Stancak, 2015; Fallon, Li, Chiu, et al., 2015; Y. Fan & Han, 2008; Han et al., 2008; Mende-Siedlecki et al., 2021), they may not be extreme enough to be detectable at the single trial level. Future research may wish to explore more intense pain imagery, such as those depicting injury (Osborn & Derbyshire, 2010), which may elicit larger ERP and behavioural responses. Additionally, the two stimuli categories used in this study (faces and scenes) were not matched for all physical properties (e.g., luminance), which may have confounded the EEG and impacted the classification. Research has demonstrated that properties such as brightness can alter EEG responses (Eroğlu et al., 2020). Therefore, we cannot entirely rule out the notion that confounds such as the physical properties of the image contributed to the classification performance. Moreover, we did not record the racial background of the participants in this study. Research has shown that neural responses during pain observation are attenuated when viewing individuals of a different race (Y. Cao et al., 2015). Therefore, collecting and reporting the racial background of the subjects in this study could have provided important additional insight. Finally, the current study only recorded neural responses. Future research should aim to record composite measures (e.g., galvanic skin response) to supplement the EEG, which may improve classification performance.

The current study has important significance in the research field. Specifically, we provide the most robust estimates of EEG-ML visual stimuli decoding due to the extensive external validation procedure. We identified a potential limit of ML-EEG techniques, as ML models were unable to accurately classify pain observation above chance levels. However, assuming model performance can be improved, developing an empathy classification tool has important applications in healthcare, such as a supplementary tool for empathy training for healthcare workers (Bas-Sarmiento et al., 2020). However, performance improvements are imperative before such applications are considered. Currently, we can reasonably predict whether an individual was observing a face or a scene on external data, which represents an important knowledge contribution. However, the criteria typically applied to clinical contexts suggest that models that demonstrate an AUC less than or equal to 0.75 are not deemed practically useful (J. Fan et al., 2006). Given that most of the AUCs in this study do not exceed this threshold, we recommend that improved model performance is pursued to increase the practical significance of the results, with a particular focus on empathic response prediction.

5.5 Conclusion

To the best of our knowledge, this is the first study to externally validate ML and EEG for the classification of various classes of visual stimuli including pain or neutral facial expressions and scenes with pain being inflicted on another person, or without pain. Our results demonstrate that ML and EEG can be used to decode neural responses and successfully classify face versus scene images with better-than-chance accuracy. However, the ML models were unable to discriminate between neutral and painful depictions of either face or scene images. Additionally, the ML result questions the suitability of passive viewing tasks for brain-based decoding algorithms. Overall, the study demonstrates promising results for decoding discrete categories of visual stimuli but is unable to identify the observation of pain using single-trial ERP responses. Finally, our results reiterate the importance of robust, external validation procedures to sufficiently evaluate ML-EEG performance; without which may lead to a new wave of impressive, but not replicable, findings.

Chapter 6:

External Validation of Machine Learning and EEG for Continuous Pain Intensity Prediction in Healthy Individuals

Tyler Mari[1], Jessica Henderson[1], S. Hasan Ali[1], Danielle Hewitt[1], Christopher Brown[1], Andrej Stancak[1], Nicholas Fallon[1]

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

This study aimed to externally validate ML and EEG for the prediction of continuous pain intensity.

The roles of the co-authors are listed below:

Tyler Mari: Conceptualisation, Methodology, Investigation, Formal analysis, Writing – Original Draft, Writing – Review and Editing. **Jessica Henderson:** Investigation, Writing – Review and Editing. **S. Hasan Ali:** Investigation, Writing – Review and Editing. **Danielle Hewitt:** Investigation, Writing – Review and Editing. **Christopher Brown:** Conceptualisation, Methodology, Supervision, Writing – Review and Editing. **Andrej Stancak:** Conceptualisation, Methodology, Supervision, Writing – Review and Editing. **Nicholas Fallon:** Conceptualisation, Methodology, Formal analysis, Supervision, analysis, Writing – Original Draft, Writing – Review and Editing.

Abstract

Previous research has predicted subjective pain intensity from electroencephalographic (EEG) data using machine learning (ML) models. However, there is a paucity of externally validated ML models for pain assessment, particularly for continuous pain prediction (e.g., decoding pain ratings on a 101-point scale). We aimed to conduct the first external validation paradigm for ML regression models for continuous pain intensity prediction from EEG data. Ninety-one subjects were recruited across three samples. Sample one (n = 40) was used for model development, sample two (n = 51) was used as a cross-subject external validation set, whilst sample three (n = 25) was used as a within-subjects temporal external validation set. Pneumatic pressure stimuli were delivered to the left-hand index fingernail bed at 10 graded intensity levels. Single-trial time-frequency features of peri-stimulus EEG were used to train a Random Forest (RF) model and long short-term memory (LSTM) network to predict continuous (0 – 100) pain intensity responses. Results demonstrated that both the RF model and LSTM network predicted pain intensity significantly more accurately than a random prediction model, with the mean absolute error (MAE) of the RF (best performing model) at 19.59, 21.29, and 18.90 for internal validation, cross-subject external validation, and within-subject external validation, respectively. However, neither model was able to predict pain intensity better than a baseline dummy model, which predicted the mean behavioural rating of the training set and did not have access to neural data. Moreover, in a replication of our recent work, we developed a RF model for the classification of low and high-pain trials, which demonstrated internal and external validation accuracies up to 64% and 58%, respectively. Taken together, our results suggest that using ML and EEG to predict continuous pain ratings is not currently feasible. However, classification models demonstrate some potential,

consistently outperforming chance across validation samples. Further improvements such as composite measures are required to elevate ML performance to a clinically meaningful level.

6.1 Introduction

Pain is subjective, complex, and challenging to measure due to an intricate interplay between biological, psychological, and social factors (Bendinger & Plunkett, 2016; Breivik et al., 2008; Gatchel et al., 2007; Younger et al., 2009). The current gold standard of pain assessment is self-report measures, requiring high-level linguistic and social skills, which are unsuitable for individuals who cannot accurately communicate their pain (Herr et al., 2011; Schiavenato & Craig, 2010). Vulnerable populations including non-verbal individuals (Herr et al., 2011; D. Li et al., 2008; McGuire et al., 2016) or individuals with cognitive impairments (Herr et al., 2011; Voepel-Lewis et al., 2002), traumatic brain injury (Arbour & Gélinas, 2014), dementia (Breivik et al., 2008; Herr et al., 2011; Kunz et al., 2009), or disorders of consciousness (Herr et al., 2011; Schnakers & Zasler, 2007), and children (Herr et al., 2011; Witt et al., 2016) are often unable to self-report their pain, which can prevent effective pain management. Therefore, pain assessment techniques that are independent of self-report may facilitate improved pain management in these populations.

Numerous brain regions contribute to pain processing, including the primary (SI) and secondary (SII) somatosensory cortex, insular cortex, anterior and midcingulate cortex, prefrontal cortex, thalamus, amygdala, periaqueductal grey, cerebellum, and brainstem (Duerden & Albanese, 2013; Jensen et al., 2016; Petre et al., 2022; Peyron et al., 2000; A. Xu et al., 2020). A recent coordinate-based activation-likelihood estimation (ALE) meta-analysis demonstrated consistent pain-related activations independent of stimulus modality, location, and gender in bilateral SII, amygdala, thalamus, brainstem, right middle frontal gyrus, left insula and midcingulate cortex (A. Xu et al., 2020). The brain regions implicated in pain processing are often considered a distinct pattern, such as the pain matrix or neurologic signature, which exhibits activity changes that encode pain intensity (Garcia-Larrea & Peyron, 2013; Wager et al., 2013). Consequently, neural markers of pain may enable proxy pain assessment.

Electroencephalography (EEG) may demonstrate clinical utility as a proxy pain assessment technique as it is low-cost and easy to use (Mackey et al., 2019; Tivadar & Murray, 2019). Importantly, pain-related changes in cortical oscillations are observable across scalp regions in established frequency bands, which may enable pain assessment (J. A. Kim & Davis, 2021; Ploner et al., 2017; Zis et al., 2022). Augmented theta oscillations have been observed during the resting state EEG of individuals with fibromyalgia (Fallon et al., 2018). Additionally, research has demonstrated increased theta amplitudes over central and parietal regions during tactile and painful stimulation, with larger amplitudes observed during painful stimulation (Michail et al., 2016). The contribution of alpha and beta bands in pain processing is well-established, with research consistently demonstrating alpha suppression and beta enhancement during tonic pain stimulation (A. C. N. Chen & Rappelsberger, 1994; Dowman et al., 2008; Huber et al., 2006; Shao et al., 2012). Finally, gamma-band oscillations over SI predict both stimulus and subjective pain intensity (Gross et al., 2007; Zhang et al., 2012). Overall, EEG activity could reliably decode pain intensity.

Supervised machine learning (ML) has been successfully implemented to decode pain-related outcomes using several neuroimaging modalities (Mari et al., 2022; van der Miesen et al., 2019). Specifically, we previously externally validated ML and EEG for low and high pain intensity classification through a multistage validation procedure (Mari et al., 2023). Using 50 time-frequency features consisting of theta, alpha, lower beta, upper beta, and gamma bands from frontal, central, and parietal regions, we classified low and high pain with a cross-validation accuracy of 73.18% using a random forest (RF). Importantly, the model generalised to a novel sample with an accuracy of 68.32%. Further, the model achieved an accuracy of 60.42% on additional external data that used different experimental pain stimulation. Consequently, our results provided robust estimates of ML performance for pain classification on novel samples. However, regression models, which demonstrate finer prediction resolution, should be assessed to improve clinical utility. Obtaining more precise pain estimates may enable improved pain management. For example, predicting continuous ratings enables finer monitoring of pain over time and allows for changes after treatment to be more accurately assessed (e.g., small changes in pain intensity can be identified, which is not possible with broad binary classification; Shirvalkar et al., 2023)**.**

Previous research has used linear regression to predict subjective pain intensity (0-10) from single-trial laser-evoked potentials (LEPs), achieving a mean absolute error (MAE) of 1.03 and 1.82 (lower scores represent better performance) for within- and cross-subject predictions, respectively (G. Huang et al., 2013). Furthermore, Bai and colleagues (Bai et al., 2016) developed a normalisation technique to reduce EEG inter-individual variability and improve model performance, achieving a MAE of 1.17 for cross-subject prediction. Using the same dataset, Li et al. (2018) predicted subjective pain intensity with a MAE of 1.19. Finally, research has demonstrated that subjective pain intensity could be predicted with a MAE of 1.15, using pre- and post-stimulus time-frequency features (Tu et al., 2016). The evidence suggests that EEG and ML can be combined to predict continuous pain ratings.

The previous findings are promising but lack external validation. To comprehensively assess ML performance, models should be evaluated on data that is independent of the training set, as internal validation methods often result in inflated performance metrics (e.g., accuracy) due to overfitting (Cabitza et al., 2021; Lever et al., 2016; Siontis et al., 2015; Vabalas et al., 2019; Varma & Simon, 2006). External validation, which evaluates model performance on novel data obtained from different cohorts, facilities, repositories or collected at a different time, location or using a different experimental paradigm, is essential to obtain robust estimates of model generalisability during prediction model development (Cabitza et al., 2021; Collins et al., 2015). As model performance is often diminished on external data (X. Li et al., 2019; Mari et al., 2023; Siontis et al., 2015), the generalisability and utility of studies that only employ internal validation are unclear and not sufficient evidence to support clinical translation (Bleeker et al., 2003; Ramspek et al., 2021). Although significant further research is required, studies that externally validate pain prediction models are emerging (Furman et al., 2020; Mari et al., 2023).

The present study aimed to externally validate ML and EEG for the prediction of subjective pain intensity both across and within subjects. Firstly, we trained a RF model to predict subjective pain intensity (0-100) using hand-crafted time-frequency EEG features. Secondly, we developed a long short-term memory (LSTM) network to predict subjective pain intensity using the EEG time series from each electrode and frequency band for each trial. Furthermore, model performance was assessed using a multi-stage validation approach consisting of cross-validation (RF model only), internal validation, and external validation (both across and within subjects). The cross-subject validation sample consisted of a new cohort, whilst the within-subject temporal validation sample consisted of participants from the model development sample who completed a second experimental session.

To assess whether the ML models predicted pain intensity better than chance levels, we compared the ML algorithms to two additional dummy models, including a random prediction model and a baseline model that predicted the mean value of all subjective pain responses from the training set. We hypothesised that the RF and LSTM would predict subjective pain intensity (0 – 100) using EEG data more accurately than both the random and baseline models, achieving lower MAE scores for all samples. Secondly, we hypothesised that the LSTM would predict subjective pain intensity more accurately than the RF model on all samples.

6.2 Methods

6.2.1 Participants

A total of 116 EEG recordings were collected across model development, cross-subject validation, and within-subject temporal validation samples. Participants were recruited using an opportunity sampling method. The model development sample consisted of 40 participants (22 female; 7 left-handed) aged between 18 and 52 (Mean = 27.70 years, standard deviation [SD] = 7.43). The cross-subject validation sample consisted of an additional 51 participants (34 female; 6 left-handed) aged between 19 and 60 (Mean = 27.63 years, SD = 9.65). There was no participant overlap between the development and cross-subject validation samples. Moreover, a total of 25 participants aged between 21 and 53 (14 female; 4 left-handed; Mean = 28.96 years, SD = 8.01) from the development sample completed the study for a second time after a minimum of 12 weeks had elapsed from their first session

(Mean = 108.68 days, SD = 10.92), resulting in a within-subject temporal validation sample. Participants were at least 18 years old, had normal or corrected-to-normal vision, no neurological conditions, no acute pain at the time of participating, no history of chronic pain and no injuries to the left-hand index finger that may affect sensory perception (e.g., nerve damage). Participants provided written informed consent before participation and all methods were conducted in compliance with the Declaration of Helsinki. This research received ethical approval from the University of Liverpool Health and Life Sciences Research Ethics Committee. Participation was reimbursed at a rate of approximately £13.33 per hour. The raw data is available from authors on reasonable request.

6.2.2 Pneumatic Pressure Stimulator

Tonic pain stimulation was delivered to the fingernail bed of the left-hand index finger using a custom-built pneumatic pressure stimulator (Dancer Design, St. Helens, UK), as utilised in previous pain research from our lab and others (Mari et al., 2023; Watkinson et al., 2013). The pneumatic stimulator consisted of a pneumatic force controller, which directed air from an 11.1-L aluminium cylinder into the stimulator. This lowered a 1cm^2 probe to deliver the desired force. The stimulator was controlled using a LabJack U3 printed circuit board for interface. The pressure was mechanically limited to a maximum of 3.5 bar (12kg/cm^2) to reduce the risk of injury.

6.2.3 Procedure

The experiment was conducted in the EEG laboratory at the University of Liverpool between June and October 2022. On arrival at the lab, participants were seated 1 metre away from a 23-inch 1080p LCD monitor inside a Faraday cage. Participants received a verbal description of the experiment before reading the information sheet and providing written consent. A custom mould of the participant’s left-hand index finger, which correctly positioned and maintained the finger underneath the stimulator probe, was created using a two-part silicone elastomer. The stimulator probe was aligned to stimulate the fingernail bed of the left-hand index finger. Additionally, participants were offered foam earplugs (28dB) to minimise any potential noise. Following alignment, participants underwent a thresholding procedure to identify their maximum-intensity stimulus.

Before the initiation of the thresholding block, participants were instructed to rate the pain intensity of each stimulus on a 101-point (0-100) numerical rating scale by using the mouse in their right hand to select the desired point on the scale. Scale anchoring was set at 0 which represented no sensation, and 100 which reflected extreme pain. Additionally, 30 represented the pain threshold and was denoted on the rating scale with the number 30 and the term “pain threshold”. The scale rating included vertical bars which denoted increments of 10. Participants were informed that ratings below 30 represented non-painful stimulation (e.g., touch) and that a rating of 0 indicated that they did not feel the probe, or the probe did not touch their finger due to finger compression. Participants were also informed that the stimulus intensity that elicited a rating of 70 or above would be used as their maximum intensity and that this value reflected upper moderate pain intensity.

A staircase procedure was implemented for the thresholding section. Here, the pressure intensity was initialised at 0.4 bar and increased in steps of 0.2 bar whilst the participant's rating was below 40. Once the rating exceeded 40, the increment was reduced to 0.1 bar. The maximum intensity of the stimulus was limited to 3.0 bar (10.5kg/cm^2) using custom software. The participants initialised the block by pressing the space bar, which began lowering the probe. During this period, a black fixation cross was presented on the screen. For each trial, the pressure stimulus had a rise time of 1 second, which reflected the time taken for the probe to go from 0 bar to the desired intensity. Subsequently, the trial intensity was maintained for 3 seconds, released, and followed by a 4-second wait period. The rating scale was then displayed until the participant successfully rated their pain intensity. Following the rating phase, the scale was replaced with a black fixation cross, and a further two-second wait time was implemented. The thresholding block was terminated once the participant had rated one of the stimulus intensities as at least 70 representing upper moderate pain. The pressure that elicited this rating was set as the maximum intensity. Moreover, participants were informed that the intensity could be adjusted throughout the session if it was either too painful or not painful enough. Stimuli intensity changes were set at fixed values of 0.1, 0.2, or 0.3 bar depending on the judgment of the researcher (e.g., an inspection of the pain ratings) and discussions with the participant. Following the completion of the thresholding procedure, participants exited the Faraday Cage for the EEG cap fitting.

For the main experimental block, a set of 10 stimuli intensities was created for each participant. The 10 stimuli intensities were linearly spaced between the minimum pressure (0.4 bar) and the upper-pressure limit selected for the participant (e.g., the pressure that elicited a pain intensity rating $\geq$ 70 on the thresholding procedure block). In each block, a total of 40 pseudo-randomised stimuli were delivered, which consisted of four randomised repetitions of each of the 10 stimuli intensities (see Stimuli Randomisation Procedure). Before the start of the block, the participant's finger was realigned under the stimulator probe and occluded from sight. Subsequently, participants were provided with verbal and written instructions for the task.

Each trial consisted of a baseline period, stimulus delivery phase, and post-stimulus rating segment. A baseline period of 4 seconds was implemented, followed by a 1-second stimulus rise time, where the stimulus intensity increased from 0 bar to the desired stimulus intensity in 1/10th increments every 0.1 seconds. Once the stimulus had reached the desired intensity, it was maintained for a total of 3 seconds before being released. Following the end of the stimulation period, a 4-second post-stimulus phase was implemented. A black fixation cross was presented on the centre of the screen continuously during the previous segments.

Following the completion of the stimulation period in each trial, participants were required to rate their subjective pain intensity using the same 101-point scale as the thresholding procedure. However, participants were informed that they could report any subjective pain intensity that corresponded to their experience on that trial, i.e., that rating above the previous 70 threshold was permitted without the experiment terminating. The rating scale section had an infinite duration with a minimum of 2 seconds. Following the rating phase, the fixation cross was presented on the screen and a subsequent forced wait period was conducted. Each trial had a minimum inter-trial interval of 16 seconds. Following the completion of the experimental task, the EEG cap was removed and participants completed the pain catastrophising scale (M. J. L. Sullivan et al., 1995) before being debriefed and reimbursed for their time.

The experiment consisted of four blocks, resulting in a total of 160 stimuli, with each block lasting approximately 15 minutes. Each block was separated by a distractor task, resulting in a minimum inter-block interval of 5 minutes. All experimental procedures were delivered using PsychoPy (Peirce, 2007).

6.2.4 Stimuli Randomisation Procedure

Stimuli were pseudo-randomised using a custom randomisation algorithm. Pseudo-randomisation was conducted to prevent the clustering of high-intensity stimuli, which could have become too painful for the participant. Consequently, the 10 stimulus intensities were ordered from minimum to maximum and arranged into 5 pools that contained two adjacent stimuli. For example, pool one included the two lowest stimuli intensities, whilst pool 5 contained the two highest stimuli intensities. An empty array was created to store the final stimuli ordering. Subsequently, one of the 5 pools was chosen at random and one value in that pool was selected, removed, and added to the final ordering array. If the selected pool was the highest intensity pool, on the next iteration, the algorithm was forced to select from pools one, two, or three which contained stimuli at approximately low and moderate intensities. If the maximum intensity pool was not selected, the algorithm could select any of the pools that still contained values on the next iteration. This process was repeated until the 10 stimuli were shuffled and appended to the final ordering list. This process was repeated 4 times for each block, resulting in a total of 40 stimuli. To prevent high stimuli clustering at the end and beginning of each iteration, additional safety measures were implemented. On the second, third, and fourth iterations, the algorithm was prevented from selecting one of the highest-intensity pools first if the last intensity from the previous iteration was also high, which further prevented high-intensity stimuli from grouping. This process was conducted for each block and participant, meaning that all blocks across all participants had a unique ordering.

6.2.5 EEG Acquisition

EEG recordings were continuously obtained using a 129-channel EGI System (Electrical Geodesic Inc., EGI, now Magstim EGI, Eugene, Oregon, USA) and a sponge-based Geodesic sensor net. The correct net position was achieved by aligning the net with respect to three anatomical landmarks: two pre-auricular points and the nasion. Electrode-to-skin impedances were monitored and maintained below 50 kΩ for all electrodes throughout the experiment. A recording bandpass filter was set at 0.001 – 200 Hz, whilst the sampling rate was set at 1000 Hz. Finally, Cz was used as the reference electrode.

6.2.6 EEG Pre-processing

Automatic EEG pre-processing, using the Harvard Automated Processing Pipeline for Electroencephalography (HAPPE; version 3; Gabard-Durnam et al., 2018), was conducted due to the large sample size of this study. Line noise at 50 (± 2) Hz was removed using CleanLine (Mullen, 2012), high- and low-pass filters of 0.5 and 70 Hz were applied, and the data were resampled to 500 Hz. Subsequently, bad channels that did not contain useable brain data (e.g., channels affected by excessive movement) were then interpolated and the remaining data underwent wavelet thresholding artefact correction. The data were then segmented into trial epochs with a period of -4 seconds to 6 seconds relative to the stimulus onset. Any channels with remaining artefacts were interpolated and segment rejection was employed for unusable segments. Here, automated epoch rejection was conducted using two criteria: the amplitude range, which was set at -150 to 150 mV as recommended by HAPPE authors (Gabard-Durnam et al., 2018) and segment similarity. Trial epochs identified in this process were marked for rejection. Following the trial rejection, the data were re-referenced using an average reference (Lehmann, 1987). Finally, the data was visually inspected. From the visual inspection procedure, it was identified that the data still contained excessive line noise. Consequently, an additional notch filter (50 ± 2 Hz) was applied to the data to remove any remaining line noise. Table 6.1 shows the average remaining number of trials for each of the 10 stimuli intensities after pre-processing for all three samples. Following pre-processing, a total of 81.12%, 80.70%, and 81.53% of trials were retained in the model development, cross-subject validation, and within-subject validation samples, respectively. There were no significant differences between the remaining number of trials for each stimuli intensity in the model development sample (p = .153), cross-subject validation sample (p = .818), or within-subject temporal validation sample (p = .876).

Table 6.1 Average number and percentage of trials retained per subject after EEG pre-processing.

	Sample		
Stimulus Intensity	Model Development	Cross-subject	Within-subject
1	12.78 ± 1.99 (80%)	12.71 ± 1.85 (79%)	13.20 ± 1.73 (83%)
2	12.98 ± 2.04 (81%)	12.92 ± 1.84 (81%)	12.92 ± 1.75 (81%)
3	13.05 ± 1.85 (82%)	13.14 ± 1.93 (82%)	13.24 ± 1.67 (83%)
4	12.90 ± 2.25 (81%)	12.84 ± 1.67 (80%)	13.08 ± 1.44 (82%)
5	13.08 ± 1.99 (82%)	12.86 ± 1.88 (80%)	13.20 ± 2.08 (83%)
6	13.28 ± 1.84 (83%)	12.63 ± 2.16 (79%)	13.28 ± 1.74 (83%)
7	13.18 ± 1.99 (82%)	13.02 ± 1.84 (81%)	13.04 ± 1.79 (82%)
8	13.53 ± 1.99 (85%)	12.96 ± 1.77 (81%)	12.76 ± 2.09 (80%)
9	12.60 ± 1.86 (79%)	12.96 ± 1.78 (81%)	12.68 ± 1.89 (79%)
10	12.60 ± 1.98 (79%)	13.08 ± 1.72 (82%)	13.04 ± 1.77 (82%)

Spectral analysis was conducted using MATLAB 2020b (The MathWorks, Inc., Natick, Massachusetts, USA) and EEGLAB 2021.1 (Delorme & Makeig, 2004). The power spectral density (PSD) was estimated using Welch's method. The PSD was calculated for each trial from -4 seconds to 6 seconds relative to the onset stimulus onset, in 1-second segments, shifted in 0.01-second increments. The data were smoothed using 7 multi-taper Slepian sequences. The PSD was calculated between 1 and 70 Hz, with a resolution of 1 Hz. Relative band power changes were calculated across each time point and frequency in the trial epoch using the event-related desynchronisation (ERD) method (Pfurtscheller & Aranibar, 1979; See equation below). The ERD estimate at each datapoint (A in the equation) is computed by subtracting the mean PSD of the baseline period (-3.5 to -0.5; R) followed by a subsequent numerical transformation to express the relative change in power as a percentage value.

$$ERD\ (\%) = \left(\frac{A - R}{R}\right) * 100$$

Negative ERD values reflect band power decreases in the active period, relative to the baseline segment, which indicates cortical activation (Neuper & Pfurtscheller, 2001; Pfurtscheller & Aranibar, 1977; Pfurtscheller & Lopes da Silva, 1999; Pfurtscheller & Neuper, 1992). Positive ERD values represent band power increases, which generally reflect cortical inhibition and are referred to as event-related synchronisation (ERS; Pfurtscheller, 1992, 2001). The data were transformed into established frequency bands which are distinct in spatiotemporal dynamics and functional associations: theta (4 – 7 Hz), alpha (8 – 12 Hz), lower beta (16 – 24 Hz), upper beta (25 – 32 Hz) and gamma (33 – 70 Hz) for the ML analysis (Keil et al., 2022; Schomer & Lopes, 2010). This was achieved by averaging the time series within the boundaries of the five frequency bands. Topographical maps, to illustrate power changes from baseline to low and high experimental pain stimulation conditions are reported in the results section. ERD visualisation was conducted and reported following recommendations from previous research (Pfurtscheller & Aranibar, 1977, 1979) and is consistent with our previous work (Mari et al., 2023).

6.2.7 Machine Learning Procedure

The model development, cross-subject, and within-subject validation samples were processed using the same pipeline but were handled separately to prevent data leakage which could have compromised the external validation procedure (Luo et al., 2016). Firstly, we conducted feature engineering by computing candidate predictors from the single-trial time-frequency transformed data. Eighteen candidate features were calculated for ERD outputs across the five frequency bands and 128 electrodes, resulting in a total of 11,520 potential features. The features were primarily comprised of descriptive statistics of the relative band power changes in each frequency band. The features were calculated from the active segment of the trial window [0-5.5s]. Fifteen of the features were identical to our previous research (see Mari et al., 2023), with the number of peaks, number of troughs, and peak-to-trough ratio also included in this study in an attempt to provide the model with additional features to improve classification performance. All features were calculated using in-built MATLAB functions, where possible, and are consistent with previous research (Anuragi & Sisodia, 2020; Mari et al., 2023; Sai et al., 2019; Vargas-Lopez et al., 2021; Vimala et al., 2019).

Single-trial EEG is significantly hampered by noise and inter-trial variability (Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014). Consequently, an outlier interpolation procedure was employed as outliers do not follow clear patterns, which impairs ML performance (Maniruzzaman et al., 2018). Additionally, interpolation was implemented to maximise the dataset size, as smaller datasets are at an increased risk of overfitting (Vabalas et al., 2019). Outliers were identified and replaced for each participant through linear interpolation using the *filloutliers* MATLAB function, in line with our previous research (Mari et al., 2023). Outliers were identified as values that exceeded three median absolute deviations. Interpolated values were calculated from neighbouring non-outlier data points. A total of 6.46 ± 0.67%, 6.39 ± 0.76%, and 6.33 ± 0.38% of the data were interpolated for the development sample, cross-subject validation sample, and within-subjects temporal validation sample, respectively.

The data were subsequently processed for ML using Python and Scikit-learn (Pedregosa et al., 2011), with the random seed value set as 123. Firstly, the model development sample was split into a cross-validation and an internal holdout test set. Here, 10% (4 participants) of the model development sample was selected for the holdout validation set. The participants within the holdout validation set were randomly selected from a subset of participants (n = 15) from the model development sample, who did not participate in the within-subject temporal validation sample. The features for each validation set were scaled to between 0 and 1 and univariate feature selection was implemented. Once the features had been successfully ranked, a custom sequential feature selection procedure was conducted. Firstly, a baseline RF regressor was developed using only the highest-ranking feature, with no hyperparameter optimisation. This was conducted as RF models demonstrate good performance with default parameters and require minimal hyperparameter tuning (Bentéjac et al., 2021; Fernández-Delgado et al., 2014). Here, we used stratified k-fold validation (k=10) for the cross-validation procedure. Subsequent features were sequentially added to the model until a maximum of 70 features had been included. Seventy was selected as the limit for the number of features as this represented 1/70th of the total of cross-validation sample observations (4685 trials). The limit was implemented to reduce model complexity (e.g., the curse of dimensionality), whilst providing enough features to enable successful ML training. The model and features that achieved the best cross-validation performance were selected as the final feature set for the full model development procedure. Through univariate feature selection, a total of 55 features demonstrated optimal cross-validation performance.

After identifying the optimal features, the final RF regressor was developed and trained on the development sample dataset. A RF model was selected as research has shown that RFs provide optimal real-world performance and are robust to overfitting (Dong et al., 2020; Fernández-Delgado et al., 2014; T. Jiang et al., 2020; Mienye & Sun, 2022). Additionally, the RF model achieved the best performance in our previous work (Mari et al., 2023). Random search with a maximum of 50,000 iterations was conducted for hyperparameter optimisation, which evaluated a range of lower and upper bounds for hyperparameter values to identify the optimal configuration (Bergstra & Bengio, 2012; Géron, 2019; L. Yang & Shami, 2020). Stratified k-fold validation was integrated into the hyperparameter optimisation. A value of k = 10 was used for this study. Hyperparameter optimisation was only conducted in relation to cross-validation performance, the hold-out, cross-subject, and within-subject validation sets did not inform model development. Following hyperparameter optimisation, the model was refit to the entire training set. The final model was subsequently evaluated on the internal hold-out validation and external validation sets.

6.2.8 Time Series ML

In addition to the RF model, we also developed a bidirectional LSTM network for time-series prediction. LSTMs are a type of recurrent neural network that can identify and learn long-term dependencies in sequence input data through the use of memory cells, which can remember inputs across time steps (Hochreiter & Schmidhuber, 1997; LeCun et al., 2015). The regulation of the stored information (e.g., when to remove it) is managed by gating mechanisms, such as forget gates (Gers et al., 2000; Hochreiter & Schmidhuber, 1997; LeCun et al., 2015). Bidirectional LSTMs provide an extension of the original LSTM architecture, by extracting information in both forwards and backwards directions. Essentially, two LSTMs are developed and trained using the input data, with one of the networks being trained on the original time series and the second being trained backwards on the time series. This allows the network to learn relationships in either direction.

We developed a 7-layer LSTM network which had a total of 2.4 million learnable parameters. The input layer dimensions were 640x550, with 640 input time-series, which reflected the 5 frequency bands (theta, alpha, lower beta, upper beta, and gamma) for each of the 128 scalp electrodes, by 550 time points. Subsequently, a bidirectional LSTM layer with 256 hidden units was implemented. On all LSTM layers, the state activation was the hyperbolic tangent function, whilst a sigmoid was used as the gate activation function. Furthermore, a dropout layer with a probability of 0.2 was implemented to help reduce the risk of overfitting, which was followed by an additional bidirectional LSTM layer, with 128 hidden units. A final dropout layer, again with a 0.2 probability, was implemented prior to the fully connected layer. Finally, the data was passed through a regression layer to obtain a continuous prediction. The network was trained for a maximum of 200 epochs, with a learn rate of 0.001, and a minibatch size of 64 samples. Gradient clipping was implemented with a threshold of 1. Finally, the network that achieved the best validation loss was selected as the final model.

6.2.9 ML Evaluation

The primary performance metric in this study was the mean absolute error (MAE) which is consistent with previous research (Bai et al., 2016; G. Huang et al., 2013; L. Li et al., 2018; Tu et al., 2016). The MAE represents the average error between the true label and the predicted label (Mari et al., 2022; Willmott & Matsuura, 2005) and can be calculated using the following equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Where y_i is the true label and $\hat{y}_i$ is the model predicted value.

For each validation set (internal hold-out, cross-subject, and within-subject validation), we calculate the MAE for each subject and average across the validation set to obtain both individual and sample accuracies.

Unlike binary classification, there is no obvious chance level to ascertain whether the models outperform random prediction. Therefore, we developed two additional dummy models that allowed us to evaluate the effectiveness of the ML models. Dummy models provide a way of establishing a minimum expected baseline performance by using simple heuristics to provide predictions without any knowledge of the input features (Fontana et al., 2019; H. R. Johnson et al., 2016; Moon et al., 2020). For example, dummy models in classification may always predict the majority class, whilst they may predict the mean of the outcome variable in the training set for regression tasks (Fontana et al., 2019; H. R. Johnson et al., 2016; Moon et al., 2020). Simple heuristics can often outperform ML performance, so the use of dummy models provides an effective baseline to evaluate the effectiveness of ML models. This process ensures that the models are not exploiting simple rules instead of learning from the input features (Fontana et al., 2019).

Firstly, we developed a dummy model that predicted a random number between 0 and 100 for each trial and calculated the MAE for each validation set. This was termed the random model. Secondly, we developed a baseline (dummy) model that predicted the mean value of subjective response data from all trials in the training set. The baseline model provided a measure of the accuracy of predictions that can be made using only behavioural pain ratings (e.g., omitting neural data). To evaluate which models achieved the lowest MAE, we conducted three one-way ANOVAs that assessed each of the four models (random, baseline, RF, LSTM) using IBM SPSS 27 (IBM Corp., Armonk, New York, USA). Bonferroni correct post-hoc tests were conducted to investigate significant main effects.

6.3 Results

6.3.1 Subjective Pain Ratings

The average pain ratings for each stimulus intensity across all three samples is reported in Table 6.2. Three simple linear regressions were conducted using SPSS 27 (IBM Corp., Armonk, New York, USA) to assess the relationship between stimulus intensity and subjective pain intensity in the development sample, cross-subject validation, and within-subject temporal validation sample, respectively. For the development sample, the regression model was significant and predicted 77% variance (R^2 = .77, F (1,5196) = 17531.63, p < .001). Stimulus intensity was a significant positive predictor of subjective pain intensity (b = 7.40, se = 0.06, p <.001, 95% CI = 7.29 to 7.51). For the cross-subject validation sample, the regression model was also significant and predicted 75% variance (R^2 = .75, F (1,6583) = 19538.91, p < .001). Stimulus intensity was a significant positive predictor of subjective pain intensity (b = 7.38, se = 0.05, p <.001, 95% CI = 7.28 to 7.49). Finally, the regression model was significant and predicted 73% variance (R^2 = .73, F (1,3259) = 8986.29, p < .001) for the within-subjects validation sample. Again, stimulus intensity was a significant positive predictor of subjective pain intensity (b = 6.60, se = 0.07, p <.001, 95% CI = 6.46 to 6.74). Figure 6.1 illustrates the relationship between stimulus intensity and subjective pain intensity for all three samples.

Table 6.2 Descriptive statistics for each stimulus intensity and sample.

	Model Development			Cross-subject Validation			Within-subject Validation		
Stimulus Intensity	Mean	SD	N	Mean	SD	N	Mean	SD	N
1	4.52	5.62	511	4.62	7.62	648	8.35	8.60	330
2	9.17	8.46	519	9.10	9.40	659	13.46	10.05	323
3	13.87	9.50	522	14.84	10.62	670	17.92	10.62	331
4	21.82	12.42	516	22.93	12.50	655	24.55	12.14	327
5	29.93	13.15	523	30.58	13.54	656	31.33	12.25	330
6	38.05	13.22	531	39.32	14.38	644	39.17	12.97	332
7	46.14	14.03	527	47.18	14.70	664	45.36	12.82	326
8	55.06	12.62	541	55.11	14.09	661	53.50	11.82	319
9	61.11	11.56	504	61.37	12.75	661	59.55	11.56	317
10	67.82	10.53	504	67.82	10.97	667	65.97	9.99	326

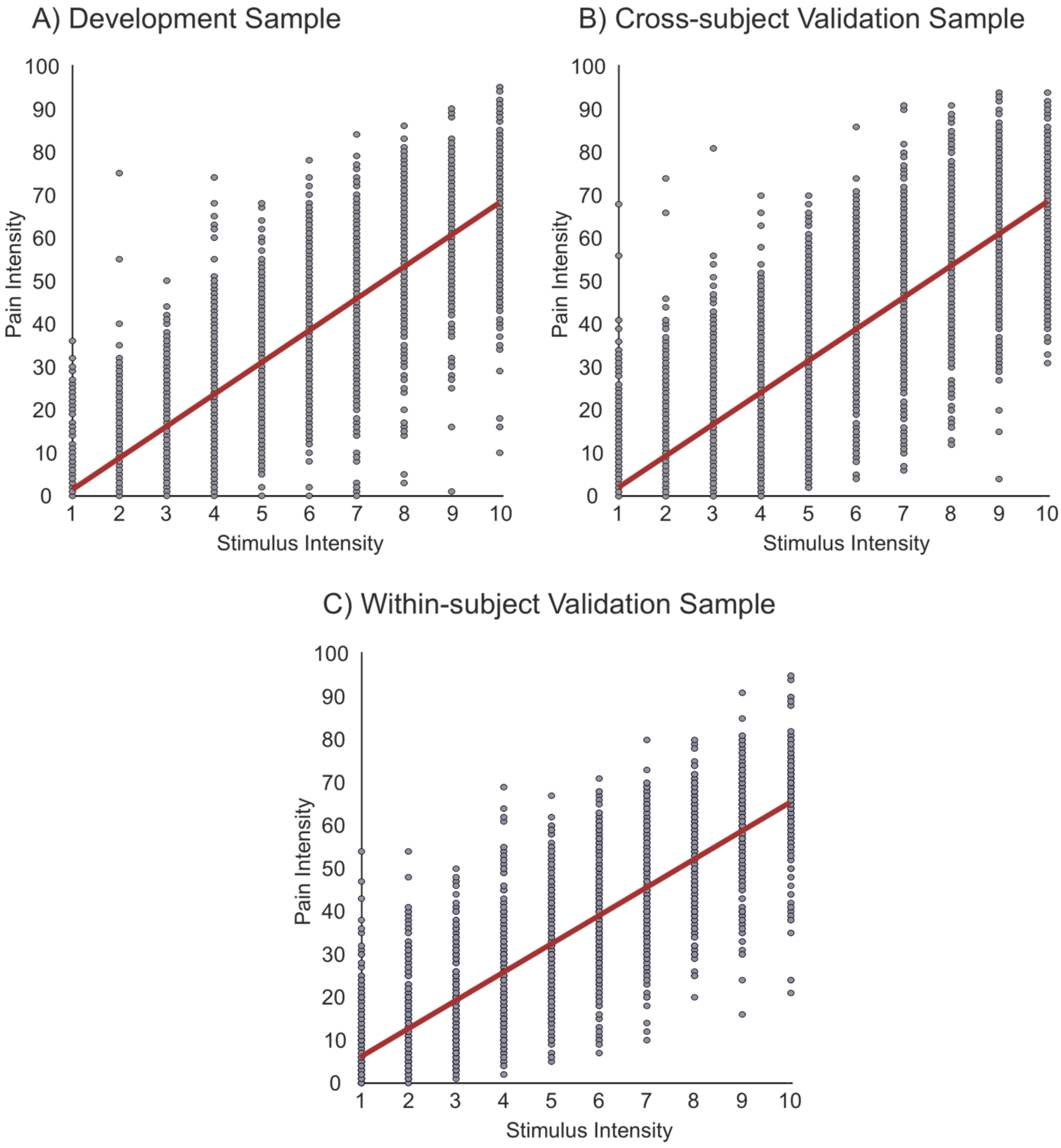


*Figure 6.1 The relationship between stimulus intensity and subjective pain intensity for the model development sample (**A**), Cross-subject validation sample (**B**), and Within-subject validation sample (**C**). The red line represents the least squares regression line.*

6.3.2 ERD/S

Topographic maps illustrating the difference between low and high pain conditions in the 91 unique participants are shown in Figure 6.2. The figure shows the time-frequency changes during the rest (-3 – -2 s relative to the onset of stimulation) and the active period (1 – 2 s relative to the onset of stimulation). The active period represents the first second of maximum pressure following the completion of the stimulation rise time. Topographic plots demonstrating relative band power changes in frequency bands Theta (4 – 7Hz), Alpha (8 – 12Hz), Lower Beta (16 – 24Hz), Upper Beta (25 –32Hz), and Gamma (33 – 70Hz) are reported. The left pair of columns represent the rest and active periods for the low pain condition, whilst the right pair of columns represent the high pain condition.

In the Theta band, we observed ERS over frontal electrodes in both low and high pain conditions (Figure 6.2A). Moreover, sensorimotor, and occipital ERD was also observed during both low and high pain conditions in theta band. There was strong bilateral ERD in the Alpha band, observed over sensorimotor regions in both low and high pain conditions (Figure 6.2B). Here, the intensity of bilateral Alpha ERD was clearly enhanced during the high pain condition, relative to low pain. Bilateral ERD was also observed in both the lower and upper Beta bands over sensorimotor regions (Figure 6.2C/D), but the pattern was visibly weaker in the upper Beta band when compared to the lower Beta band. In both lower and upper Beta bands, a similar pattern to Alpha processing was observed, with stronger ERD evident in the high pain condition. Finally, for the Gamma band power changes, we observed ERS over frontal regions in both low and high pain stimulation, which appeared to be more widespread during the high pain stimulation condition (Figure 6.2E).

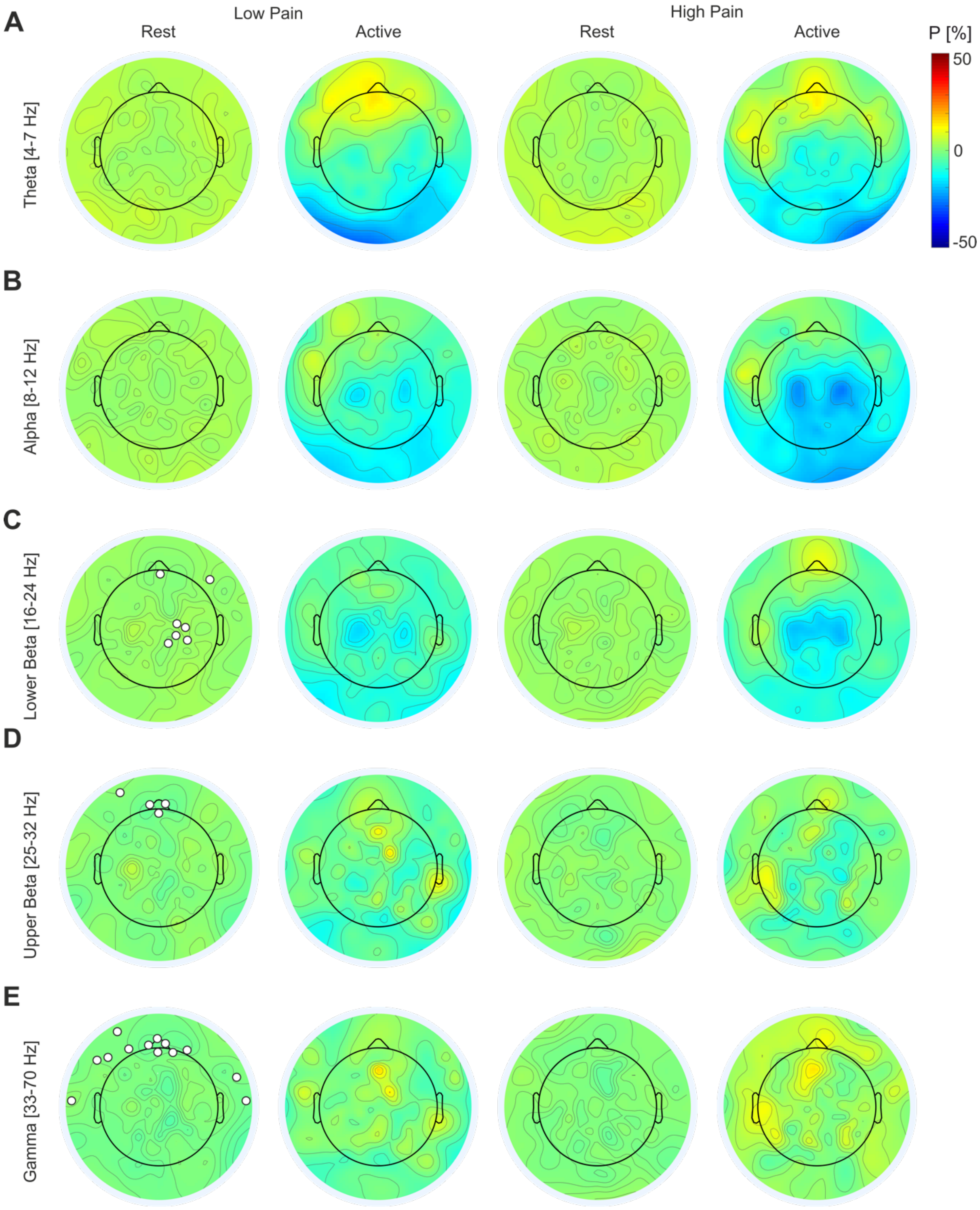


Figure 6.2 Grand average band power changes during rest (-3 – -2s) and active pressure stimulation (1 – 2 s) from all 91 unique participants (combined sample one and two). The trial period spanned from -4 s to 6 s relative to the trial onset, with a baseline period of -3.5 s to -0.5 s. The active period was selected in line with previous recommendations (Pfurtscheller & Aranibar, 1977, 1979) *and represented 1 s of continued pressure immediately after the stimulator reached the desired stimulus intensity level. The topographic maps show the band*

power changes in low and high pain intensity conditions and from rest to active periods in Theta (A), Alpha (B), Lower Beta (C), Upper Beta (D), and Gamma (E). The white circles represent the electrode locations of the features used in the ML classification of low and high pain trials. P = percentage power change from baseline.

6.3.3 Machine Learning Results

The final number of observations for each validation set/sample is presented in Table 6.3. From the feature selection procedure, a total of 55 features were optimal. The features were distributed across scalp regions and frequency bands. The theta and alpha features were predominantly located over a frontal-central electrode. Features in the beta band were located predominantly over frontal regions, with additional features calculated from peripheral electrodes. The gamma band provided most features (approximately 50%) and was distributed over frontal regions and central-parietal regions, respectively.

Table 6.3 The number of observations for each validation set.

Validation Set	Number of Trials
Training Sample (n = 36)	4685
Internal Hold-out Validation (n = 4)	513
Cross-subject (n = 51)	6585
Within-subject (n = 25)	3261
Total	15044

The MAE for all the models and validation sets are reported in Table 6.4. For the validation sample, the RF model predicted pain intensity on a 101-point scale with a MAE of 19.59 points, whilst the LSTM demonstrated similar performance with an average error of 19.97.

Whereas the random model and baseline model predicted subjective pain intensity with a MAE of 32.61 and 19.98, respectively. A one-way ANOVA demonstrated that there was a significant main effect of model on the MAE (F (3,12) = 22.74, $p < .001$, $\eta_p^2 = .85$). Bonferroni post-hoc tests showed that the random model had a significantly higher error for predicting subjective pain intensity than the RF model ($p < .001$), LSTM network ($p < .001$), and the baseline model ($p < .001$). However, there were no significant differences between the baseline model and the RF model ($p = 1.00$) or the LSTM network ($p = 1.00$) for pain intensity prediction. Finally, there was no significant difference between the RF model and the LSTM network ($p = 1.00$).

Regarding the cross-subject validation sample, the LSTM demonstrated the most accurate pain intensity prediction, achieving a MAE of 21.17. Here, the RF predicted subjective pain intensity with an error of 21.29, which was less accurate than the baseline model, which demonstrated a MAE of 21.19. Finally, the random model demonstrated an error of 33.41. A one-way ANOVA demonstrated a significant main effect of ML model on the MAE for the cross-subject validation sample (F (3,200) = 169.12, $p < .001$, $\eta_p^2 = .72$). In line with the internal validation results, the random prediction model had significantly greater error for pain predictions when compared to the RF model ($p < .001$), LSTM network ($p < .001$), and the baseline model ($p < .001$). Again, there were no significant differences between the baseline model and the RF ($p = 1.00$) or the LSTM network ($p = 1.00$). No significant difference was observed between the RF and the LSTM ($p = 1.00$).

Finally, the RF demonstrated the most accurate predictions for pain intensity for the within-subject temporal validation sample, achieving a MAE of 18.90. The LSTM was the next best-performing model, achieving an error of 19.03. Finally, the random model and baseline model demonstrated less accurate predictions, achieving MAEs of 32.24 and 19.08, respectively. The ANOVA demonstrated a significant main effect of model on the MAE (F (3,96) = 121.86, p < .001, η_p^2 = .79). The random model demonstrated significantly larger prediction error than the RF model (p < .001), LSTM network (p < .001), and the baseline model (p < .001). No further significant differences were observed between the baseline model and RF model (p = 1.00), the baseline model and LSTM network (p = 1.00), or the RF model and LSTM network (p = 1.00).

Table 6.4 The MAE for each model across all validation sets.

	Internal Validation				External Validation			
	Cross-Validation		Hold-out Validation		Cross-subject		Within-subject	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
RF	20.49	0.69	19.59	2.30	21.29	3.40	18.90	3.34
LSTM	-	-	19.97	2.50	21.17	3.53	19.03	2.73
Random Model	-	-	32.61	3.24	33.41	2.90	32.34	2.76
Baseline Model	-	-	19.98	2.58	21.19	3.51	19.08	3.19

Optimal RF hyperparameters: Number of estimators = 145, Maximum depth = 8, Minimum samples to split = 19, Minimum samples at leaf = 12, Maximum features = sqrt, Bootstrap = True.

Calibration for regression is not as clearly calculated as classification (Levi et al., 2022). Moreover, calibration assessment is only required when the models demonstrate good predictive capability (Alba et al., 2017). Therefore, as neither the LSTM nor the RF demonstrated predictive performance that surpassed the baseline model, we did not formally assess calibration. However, to provide insight into the behaviour of the model, we aimed to visualise the model's predicted values and the true pain intensity values. We only provide the visualisations for the RF as, on average, it demonstrated the most accurate predictions. Figure 6.3 illustrates the relationship between subjective pain intensity and the RF model's predicted values for all three validation sets. Briefly, the results show that the RF model tended to predict values in the middle of the scale, with most of the predictions for all samples falling between 25 and 50. Moreover, there is not a clear trend between the predicted values and true values, suggesting that the model failed to extract relevant information to provide informative predictions. Finally, the model failed to predict extreme values (e.g., < 25 and > 65), meaning that the model could not accurately predict the lowest and highest intensities.

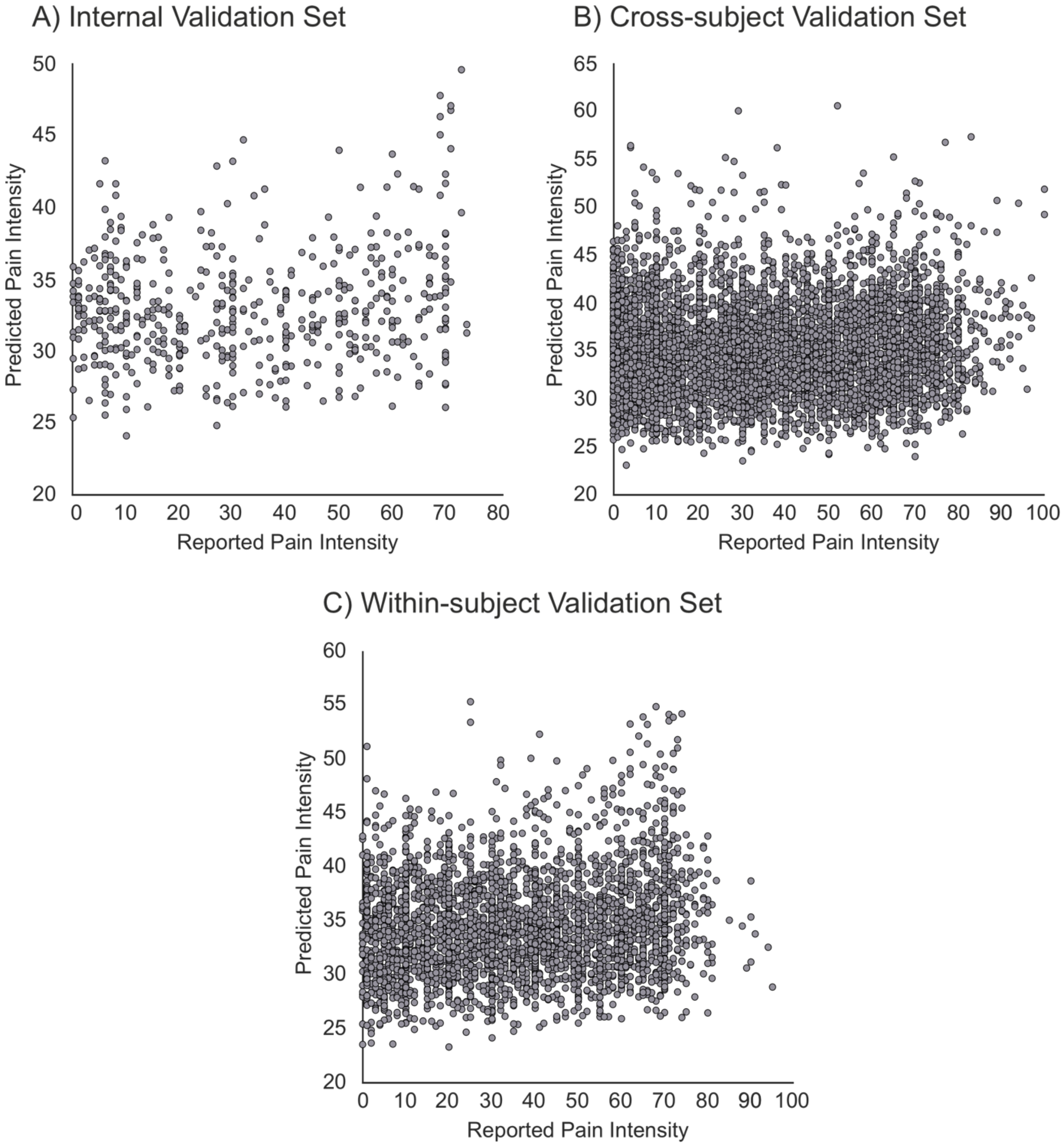


*Figure 6.3 The relationship between the RF model's predicted pain intensity and the reported subjective pain intensity for the internal validation set (**A**), cross-subject validation set (**B**), and within-subject validation set (**C**).*

6.3.4 Exploratory Analysis

As the regression model was unable to outperform the baseline model and provide informative predictions, we attempted to develop a binary classification RF model for the prediction of low and high pain trials, providing a theoretical replication of our prior work (Mari et al., 2023). This was achieved by selecting stimulus intensities that were comparable to the low and high stimuli intensities of our previous study. Stimuli intensities 4 and 5 were chosen for the low intensity, whilst 9 and 10 were chosen for the high intensity set, as these elicited comparable subjective pain ratings to our earlier model development sample (Mari et al., 2023). The RF model was developed and evaluated using a similar procedure to the regression model for feature selection and hyperparameter optimisation. Several common performance metrics were used to evaluate the performance of the model. Here, we report the accuracy, AUC, Brier score, F1 score, precision, and recall, which is consistent with existing research (Mari et al., 2022, 2023). The ML classification results are reported in Table 6.5. Moreover, the optimal feature locations for the classification are displayed in Figure 6.2.

Table 6.5 Performance metrics for binary RF model for the prediction of low and high pain intensity.

	Accuracy	SD	AUC	SD	Brier	SD	F1	SD	Precision	SD	Recall	SD
Internal Validation												
Cross-validation	0.64	0.05	0.67	0.06	0.23	0.01	0.61	0.08	0.66	0.05	0.58	0.11
Hold-out	0.62	0.11	0.68	0.18	0.22	0.05	0.58	0.12	0.61	0.15	0.57	0.15
External Validation												
Cross-subject	0.57	0.08	0.62	0.10	0.24	0.03	0.52	0.13	0.60	0.11	0.50	0.20
Within-subject	0.58	0.09	0.62	0.11	0.25	0.03	0.49	0.19	0.59	0.15	0.47	0.24

Optimal RF hyperparameters: Number of estimators = 2280, Maximum depth = 15, Minimum samples to split = 3, Minimum samples at leaf = 7, Maximum features = log2, Bootstrap = True.

The results can be separated into internal and external validation procedures. For internal validation, the classification results demonstrated that the RF model outperformed the theoretical chance level (50%) for the classification of low and high-pain intensity trials using EEG features, achieving a cross-validation accuracy and AUC of 64% and 0.67, respectively. Similar results were observed for the holdout validation sample, with the RF model exceeding chance level performance. The results demonstrated that the RF classified subjective pain intensity with an accuracy of 62% and an AUC of 0.68.

Regarding external validation, the results show that the RF model exceeded chance performance on both the cross-subject and within-subject validation samples. Here, for the cross-subject validation sample, subjective pain intensity was classified with an accuracy of 57% and an AUC of 0.62. Similar results were observed for the within-subject validation sample, with the model producing an accuracy of 58% and an AUC of 0.62. Overall, the RF demonstrated similar, above-chance performance, for both external validation datasets.

Finally, we also assessed calibration as the model performance exceeded chance performance. Calibration refers to the agreement between the model's predicted outcome value and the true outcome value (Alba et al., 2017; Luo et al., 2016; Van Calster et al., 2019). Calibration curves, which illustrate the relationship between the predicted probabilities (x-axis) and the observed probabilities (y-axis), are the preferred way to assess binary classification models (Moons et al., 2015; Van Calster et al., 2016, 2019). To achieve this, we split the data into 10 equal bins, which represented the probabilities between 0 and 1 (Y. Huang et al., 2020). The calibration curves for the holdout validation, cross-subject external validation, and within-subject external validation samples are presented in Figure 6.4. To interpret the calibration curves, when the model performance line is above the reference line (which represents perfect calibration), it suggests that the model underestimates the probability of the incidence. Whereas a model line below the reference line indicates that the model is overestimating the probability of the incidence. The calibration results suggest that the predictions from the RF model are reasonably well calibrated. There are instances of the model underestimating the incidence of the event at the lower probabilities and overestimating the higher probabilities. However, the calibration curves typically follow the expected trend, suggesting that the model provides reasonably accurate probability estimates.

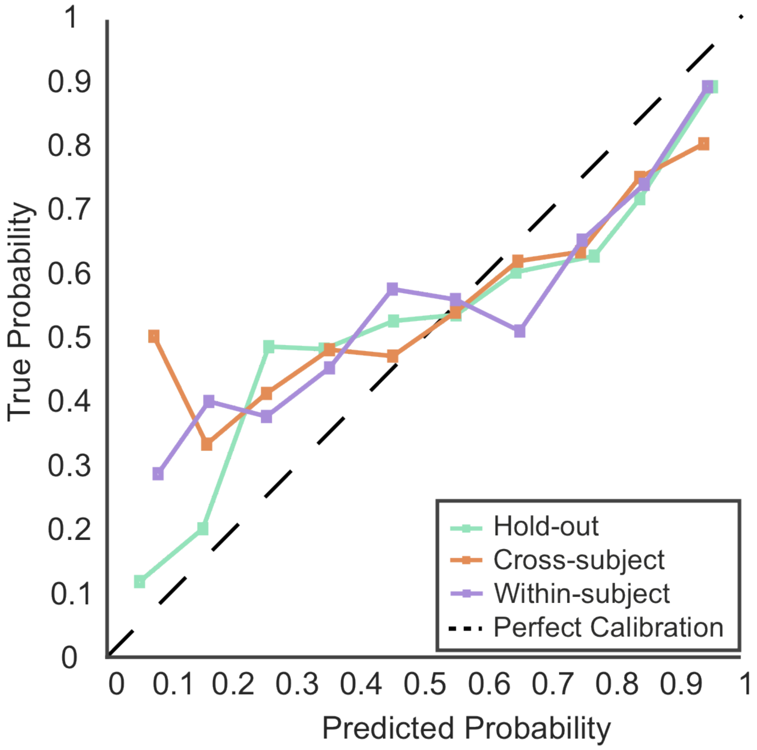


Figure 6.4 Calibration curves for the internal holdout validation sample, and two external validation samples. The black dotted line 45^{o} represents perfect calibration.

6.4 Discussion

We aimed to externally validate ML for the prediction of subjective pain intensity both across and within subjects using single-trial EEG. We hypothesised that both the RF model and LSTM network would predict pain intensity more accurately than a random prediction model and a baseline prediction model across three validation sets. The results partially support the hypotheses, with both the RF and LSTM demonstrating more accurate pain intensity predictions than the random model on all validation datasets. However, the baseline (dummy) model which utilised only behavioural data also outperformed the random model. Moreover, neither the RF nor the LSTM outperformed the baseline model on any of the validation sets. Our second hypothesis that the LSTM would outperform the RF was not supported as no differences in model performance were observed on any of the validation sets. The results suggest that regression models trained on oscillatory EEG data cannot predict subjective pain intensity more accurately than simple heuristics using the current approach.

As the regression models failed to provide meaningful predictions, we subsequently aimed to replicate our previous research by classifying low and high-pain intensity trials (Mari et al., 2023). Here, the RF model accurately classified the conditions with better-than-chance accuracies of 64%, 62%, 57%, and 58% for cross-validation, internal holdout validation, cross-subject external validation, and within-subject external validation, respectively. Moreover, the model demonstrated AUCs of 0.67, 0.68, 0.62, and 0.62 for the validation samples. Overall, the classification results are promising but require improvement to demonstrate clinically meaningful levels (e.g., AUC $\geq$ 0.75; J. Fan et al., 2006).

Despite not exceeding baseline performance, our regression metrics are comparable to previous research. Huang and colleagues (2013) demonstrated that continuous pain ratings could be predicted using LEPs with a MAE of 1.03 and 1.82 (11-point scale) for within-subject and cross-subject predictions, respectively. Further studies report MAEs of approximately 1.2 for continuous pain prediction using ML and EEG (Bai et al., 2016; L. Li et al., 2018; Tu et al., 2016). Therefore, our findings are comparable to existing literature. However, the previous results are not externally validated and should be interpreted cautiously (Cabitza et al., 2021; Siontis et al., 2015; Vabalas et al., 2019; Varma & Simon, 2006). Consequently, the results of this study provide the most robust estimates of the potential of ML and EEG for pain intensity prediction, providing realistic estimates of both model performance and clinical potential.

Our classification results support existing literature that has demonstrated that EEG and ML can classify low and high pain trials with cross-validated accuracies between 62 and 89.58% (Bai et al., 2016; G. Huang et al., 2013; Mari et al., 2023; Okolo & Omurtag, 2018; Schulz et al., 2012; Tu et al., 2016). In addition, the external validation results are comparable to our previous work which demonstrated that ML could predict low and high pain trials across subjects using two external datasets, previously achieving accuracies of 68% and 60%, respectively (Mari et al., 2023). Despite observing slightly reduced performance in the current study, the results support the potential of EEG and ML for pain intensity classification.

This study highlights the challenges associated with predicting continuous pain intensity using EEG and ML, whilst supporting the potential of classification models. Due to the large sample size and multi-stage external validation, our results provide robust performance estimates of the effectiveness of ML and EEG for pain intensity prediction. However, the increased rigour may explain the observed performance reduction compared to our previous research (Mari et al., 2023). Small samples demonstrate increased performance variability (Arbabshirani et al., 2017; Vabalas et al., 2019; Varoquaux, 2018). Varoquaux (2018) reviewed ML performance across several domains including Alzheimer's, autism, and depression, and identified that model performance decreased as a function of sample size. Furthermore, small external validation samples can also result in imprecise model performance estimates (K. I. E.

Snell et al., 2021). As much of the previous research consists of small samples, performance is likely inflated due to increased variability (Mari et al., 2022). The impact of small samples may also explain the observed minor reduction in external validation performance compared to our recent research (Mari et al., 2023), which comprised fewer participants. Therefore, this study supports our previous work with improved robustness to give increased confidence in the findings. Overall, ML and EEG remain promising for pain classification, but improved performance from robustly designed studies remains imperative. Finally, whilst continuous pain intensity prediction is desirable for finer prediction resolution which would enable improved pain assessment and treatment monitoring/recommendations (Shirvalkar et al., 2023), it appears unrealistic within the current approach.

The current study has several limitations. Firstly, EEG has a low signal-to-noise ratio (Tivadar & Murray, 2019), which likely affects the ability of the ML algorithm to extract meaningful patterns at the single trial level. Single-trial EEG is inherently noisy due to the variability and volatility of neural activity and due to the physical limitations of the apparatus (Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014; Tivadar & Murray, 2019). The diminished signal quality may have impaired model performance. In future, methods to improve the signal-to-noise ratio should be explored (e.g., spatial filtering; Miao et al., 2021; Rivet et al., 2009). Spatial filtering aims to increase the signal-to-noise ratio by maximising the differences between two classes and increasing discriminability (Blankertz et al., 2008; Y. Miao et al., 2021; Rivet et al., 2009). However, spatial filtering techniques are also prone to overfitting (Blankertz et al., 2008; Grosse-Wentrup et al., 2009), which could exacerbate an already prevalent issue in this research area. Therefore, spatial filtering techniques should be applied cautiously to minimise overfitting risk.

Moreover, whilst the RF classification model demonstrated reasonably well-calibrated predictions, there were instances where the model provided inaccurate estimates. The calibration assessment demonstrated that the model estimates were occasionally too extreme, as the models occasionally tended to underestimate the lower probabilities and overestimate the higher probabilities, which indicates a degree of overfitting (Van Calster et al., 2019). However, as single-trial EEG is an inherently noisy signal (Faisal et al., 2008; Kaplan et al., 2005; Marathe et al., 2014; Tivadar & Murray, 2019), it is expected that the model captures some random noise in the training set. Therefore, we recommend interpreting the results with caution. However, the calibration of the present study was arguably improved compared to our previous research (Mari et al., 2023), which demonstrates the improved robustness of this study. Moreover, as calibration is rarely assessed (Christodoulou et al., 2019; Mari et al., 2022), the calibration assessment of this study is an area of novelty and represents a methodical improvement over the existing literature. Nevertheless, future research should improve model calibration through techniques such as Platt scaling or isotonic regression (Y. Huang et al., 2020).

Despite promising classification results, further developments are imperative for clinical translation. Research has suggested that binary classification models with an of AUC less than 0.75 are not clinically meaningful (J. Fan et al., 2006). Therefore, future research should prioritise improving model performance on external data towards clinically meaningful results. The use of composite biological measures as predictive of subjective pain intensity may improve performance (Rockholt et al., 2023; I. Tracey et al., 2019). Methods such as heart rate variability, skin conductance, and pupillometry have shown promise for pain assessment (Cowen et al., 2015). For example, skin conductance has been shown to discriminate the presence or absence of pain in postoperative patients with promising sensitivity and specificity (Ledowski et al., 2006). Moreover, recent research has shown that ML, electrodermal activity, photoplethysmography, and respiration measures could predict the presence or absence of pain with accuracies up to approximately 94% (Fernandez Rojas et al., 2023), although this finding was not externally validated. In future, the potential clinical utility of a combination of several physiological measurements for pain classification should be explored.

Alternative neuroimaging techniques may also prove advantageous for improving model performance. Magnetencephalography (MEG) may be a potential candidate. MEG records magnetic fields as opposed to electrical fields of EEG, which are not distorted by anatomical features (e.g., differences in skull thickness), resulting in improved signal-to-noise ratio and spatial localisation of signals (S. Singh, 2014). Therefore, MEG may improve signal quality and consequently model performance. However, MEG is more expensive and less practical than EEG (S. Singh, 2014) which reduces the potential clinical applications, as EEG is easily applied and low-cost (Mackey et al., 2019; Tivadar & Murray, 2019). Although MEG is being used clinically in the surgical treatment of epilepsy (S. Singh, 2014), currently, there is limited research using MEG and ML for pain classification. One recent study demonstrated that ML and MEG could classify healthy controls and chronic migraine patients with an accuracy greater than 86% (AUC > 0.9). Moreover, the study also demonstrated that ML could correctly classify patients into either chronic migraine and episodic migraine, as well as chronic migraine or fibromyalgia patients with high accuracies (Hsiao et al., 2022). Moreover, given the ongoing development of low-cost MEG that can operate at room temperature (Boto et al., 2017), future research should investigate the utility of MEG for pain intensity prediction, perhaps combining it with EEG for multimodal imaging (S. Singh, 2014; Yoshinaga et al., 2002), which may further improve performance.

6.5 Conclusion

Our results suggest that prediction of fine-graded resolution pain intensity, such as categorising individual subjective experience on a graded scale, may not be possible using the current approach, as model performance including EEG data did not exceed a simple baseline model. Despite this, our results remain promising for predicting discrete categories of pain (e.g., low, and high pain), with the RF model exceeding chance performance for both cross-subject and within-subject predictions using external data. Due to the large sample size, the current study provides the most robust estimates for the potential of ML and EEG for pain intensity classification. Increasing the signal-to-noise ratio of EEG remains a priority, whilst composite measures should be explored and externally validated in future research to improve the performance of ML algorithms. Overall, ML and EEG can accurately predict discrete levels of subjective pain intensity, but performance levels are not currently sufficient to be considered clinically meaningful or to support the development of translation tools.

Chapter 7:

General Discussion

This thesis endeavoured to decode neural responses during noxious mechanical stimulation and pain observation using ML and EEG. We aimed to improve the understanding of the current capabilities and limitations of ML-EEG approaches and improve the understanding of the neural processes associated with empathic processing. These aims have potential to contribute to real-world applications, including the development of an objective pain assessment tool for use in vulnerable populations, or potential implications for patient and healthcare provider interactions. Initially, we systematically reviewed the literature to identify the state-of-the-art ML approaches for predicting pain-related outcomes using EEG (Chapter 3). We identified pervasive methodological limitations, resulting in overestimated performance, due to a paucity of externally validated research. Consequently, we robustly evaluated the effectiveness and generalisability of ML and EEG for pain intensity prediction using rigorous external validation procedures, calibration assessment, and greater sample sizes (Chapter 4). We found that ML and EEG could effectively classify binary pain intensity with reasonable performance and generalisability. However, model performance was considerably lower than the existing literature, suggesting that the performance of previous research (that lacks external validation) is likely inflated. In Chapter 6, we further increased methodological rigour and externally validated ML both across and within subjects. These results provided further support that the capability of ML and EEG for pain prediction is lower than reported in the literature. Moreover, we also aimed to assess whether ML and EEG could predict the observation of pain, which had not been attempted previously (Chapter 5). The approach successfully classified the observation of faces or scenes regardless of the pain component but was unable to classify pain observation, providing potential limitations of ML. Overall, this thesis contributes important knowledge to the field, by producing the most robust estimates of the effectiveness of ML and EEG for predicting subjective pain intensity and pain observation.

7.1 Summary of Findings

7.1.1 Chapter 3

- The systematic review demonstrated that the combination of EEG and ML could accurately predict subjective pain intensity, pain phenotypes, and response to treatment with accuracies between 57% and 100%.
- In previous research, time-frequency and ERP features were used to develop both classification and regression models.
- Methodological and reporting issues significantly hindered the interpretation of ML and EEG's effectiveness and clinical potential.
- Existing research was at a high risk of bias due to insufficient analysis procedures (lack of external validation) and small sample sizes, which likely contributed to exaggerated performance metrics.
- Model calibration was consistently omitted, further hindering the interpretability of the results.

7.1.2 Chapter 4

- We developed ML models that were able to classify low and high-pain-intensity trials with cross-validation accuracies of up to 77%.
- Through an extensive external validation procedure, we demonstrated that pain intensity could be predicted with accuracies of up to 69% in a novel sample.
- The ML models were also able to predict pain intensity with accuracies up to 61% during novel experimental pain stimulation in a new sample.
- Time-frequency features from frontal, central, and parietal regions produced optimal classification results.
- Our results provide the first externally validated ML performance estimates for pain intensity prediction.

7.1.3 Chapter 5

- We aimed to conduct the first ML-EEG study for the classification of neural responses during the observation of pain.
- Features calculated from single-trial ERP waveforms enabled accurate classification of the observation of faces or scenes, achieving externally validated accuracies of up to 69%.
- However, ML and EEG were unable to classify pain or neutral images for either pain scenes or faces, demonstrating the potential limits of EEG-ML approaches.
- The neutral-pain scenes classification reinforces the importance of external validation as overfitting was easily identified.

- The study demonstrates promise for decoding discrete categories of visual stimuli, but not the observation of pain.

7.1.4 Chapter 6

- We attempted to externally validate regression models for the prediction of continuous pain ratings.
- Regression models for ML-EEG outperformed a random model but failed to exceed the performance of a baseline dummy model that exploited simple heuristics.
- The regression results highlight the difficulty in predicting fine-graded pain responses.
- The classification results remained promising, with low and high pain trials classified with accuracies of 64%, 62%, 57%, and 58% for cross-validation, internal holdout validation, cross-subject external validation, and within-subject external validation, respectively.
- The study highlights greater potential clinical utility for categorical pain intensity classification, whilst also suggesting that predicting continuous scores may not be feasible.

The current thesis provides several important contributions to the existing literature. The findings provide the most robust estimates of the effectiveness of ML and EEG for predicting both pain intensity and pain observation. Our results also provide robust evidence regarding state-of-the-art and potential for practical and clinical applications, with performance across all chapters failing to reach levels sufficient to justify practical or clinical applications currently (J. Fan et al., 2006). Therefore, this thesis highlights the need for both improved performance and methodological rigour before the potential of ML-EEG approaches can be realised, e.g., for the development of practical applications or clinical tools.

Regarding the prediction of subjective pain intensity, this thesis provides a realistic framing of the current potential of ML-EEG methods and provides important methodological and practical considerations. Firstly, our results highlight that the capability of ML is likely exaggerated in previous literature. The existing cross-validation performance estimates for subjective pain intensity classification suggest that pain states can be classified with accuracies between 62 and 96%, with most of the studies demonstrating accuracies of above 80% (Mari et al., 2022). However, the results of this thesis demonstrate that the performance observed on external data may be significantly lower than these metrics, with accuracies between 57 and 69% being observed for the best models. Consequently, whilst ML and EEG for pain intensity classification remain promising, the potential of the method has likely previously been overstated. Moreover, our results also provide important practical considerations. For example, this thesis suggests that classification models are more effective for identifying subjective pain intensity than regression models. Here, the regression models were ineffective as they failed to outperform simple heuristics. Therefore, whilst the concept of fine-grade objective prediction is appealing, our results suggest that it may not be feasible, with classification methods deemed more promising. Whilst recent research has also demonstrated that regression models are ineffective for predicting continuous subjective pain intensity using intracranial electrodes (Shirvalkar et al., 2023), we are the first to identify this using scalp electrodes. Therefore, we recommend that emphasis should be placed on developing highly effective classification models for broad pain states (e.g., low, and high pain) as such methods may hold greater clinical potential.

Regarding the classification of vicarious pain/visual stimuli, this thesis provides the first externally validated results for the classification of visual stimuli both across and within subjects. Importantly, our study offers methodological advancements through significantly increased sample sizes compared to the existing research that attempted to classify discrete visual categories (Bagchi & Bathula, 2022; Cudlenco et al., 2020; Kaneshiro et al., 2015; Stewart et al., 2014; Yavandhasani & Ghaderi, 2022; Zheng et al., 2020) and through rigorous external validation procedures. The attempted classification of empathic stimuli also provided insight into the potential limits of ML. We addressed an important knowledge gap, as previous research had not attempted to classify the observation of pain using EEG and ML. Moreover, developing an objective measure of pain empathy is important for patient-clinician interactions where potential bias can result in the underestimation of pain (Hoffman et al., 2016). We were the first to attempt to develop a pain empathy classification tool using ML and EEG, and our results enhance the existing literature that demonstrated that empathic states could be classified using EMG responses or fMRI (Christov-Moore et al., 2020; Vaughn et al., 2018; Zhou et al., 2020). Our results demonstrated that ML and EEG were unable to successfully classify the observation of pain relative to visually similar neutral stimuli. Overall, this thesis generally supports the effectiveness of ML applied to EEG data, demonstrating generalisable performance within all empirical chapters of this thesis. However, limited performance metrics and unsuccessful classification attempts were observed, suggesting that ML-EEG approaches for complex inferences from single-trial data may not be as capable as previously believed. Therefore, this thesis highlights the importance of robust external validation procedures to prevent the risk of inflated, un-generalisable performance metrics, and to support the development of approaches and tools which can achieve clinically meaningful performance for practical applications.

7.2 Themes

7.2.1 Importance of External Validation

One of the key themes identified in this thesis is the importance of external validation when developing prediction models. Across all experimental chapters in this thesis, we demonstrated that ML performance generalised to novel samples, which was previously unknown in both pain intensity and visual stimuli classification domains. However, performance levels were consistently substantially lower on external validation than on internal validation. For example, in our first experimental pain experiment (Chapter 4), we identified that although the models often successfully generalised to novel samples and experimental pain stimuli, performance was often significantly lower for the external validation datasets. Whilst the best-performing models only demonstrated modest reductions of around 5%, others demonstrated significant reductions of up to 20%. Similar results were obtained during the second experimental pain intensity chapter, with the RF classifier demonstrating a reduction of performance of approximately 7% (5% from holdout validation). Therefore, based on our findings, it is reasonable to expect that the internal validation performance reported in the literature would likely reduce when assessed on external data. However, as the previous research has never been evaluated on external data, it is impossible to conclude to what extent model performance would diminish.

The importance of external validation is further exemplified by our results for the classification of pain empathy in Chapter 5. Here, performance for the faces versus scenes classification was reduced by around 10% from internal to external validation. However, the pain scene classification further reinforces the importance of external validation, as performance reduced from 80% during cross-validation to 28% and 51% during cross-subject and within-subject external validations, respectively. This represents a reduction of up to 52%, which questions the validity of studies that solely employ internal validation procedures. Without external validation, we would have concluded that the model is highly effective at discriminating the pain and neutral classes of visual stimuli, when in fact, the model was overfitting and could not effectively generalise. Therefore, external validation is imperative to assess model performance and control for overfitting. Moreover, conducting external validation also reduces the risk of over-optimisation (e.g., overfitting to the test set). As external data is easier to process separately than splitting a single dataset, overfitting to the test set and data leakage are less likely to occur during an external validation protocol. Therefore, it is possible that studies reporting high-performance metrics would not generalise and would exhibit larger reductions than those reported in this thesis. These findings further reiterate the reasons why the existing literature was deemed at a high risk of bias due to insufficient validation procedures, as the model generalisability cannot be accurately interpreted. Overall, despite many of our models demonstrating generalisable performance that outperformed chance levels, reduced metrics were consistently found when assessing model performance using external data. Consequently, this supports the argument that the performance metrics reported in the literature are likely inflated. Moreover, given that reductions in performance of up to 52% were observed, the effectiveness of ML models that have not been externally validated should be interpreted with caution, as the true performance could be dramatically lower than the performance observed during internal validation procedures.

The findings of this thesis reiterate the importance of thoroughly evaluating prediction models. Our work provides direct evidence of the potential of inflated internal validation metrics, which hinders the interpretation of the previous research. Internal validation methods are prone to overfitting, resulting in performance which does not generalise to novel data (Cabitza et al., 2021; Ramspek et al., 2021; Siontis et al., 2015; Varma & Simon, 2006). Consequently, the model performance may appear promising but cannot be considered reflective of the true error due to the risk of overfitting. Moreover, given that small samples further exacerbate the likelihood of overfitting (Combrisson & Jerbi, 2015; Vabalas et al., 2019; Varoquaux, 2018), studies that do not employ external validation, should be deemed at a high risk of bias, as the limitations jeopardise the utility of the research. We observed that the reduction in performance from internal validation to external validation varied between approximately 5% to over 52%. Previous research has further supported the reduction in performance observed during external validation (X. Li et al., 2019; Siontis et al., 2015). Consequently, we argue that by omitting external validation methods, the true effectiveness of the model cannot be established, meaning that the findings of previous research are insufficient evidence for accurate interpretation or clinical translation (Bleeker et al., 2003; Ramspek et al., 2021; Siontis et al., 2015). To successfully progress the research field, and to develop an effective prediction model, it is imperative to thoroughly evaluate model performance using independent data (Lever et al., 2016). Without improved validation protocols, the results cannot be entirely trusted, contributing to the research waste (Collins et al., 2014; Collins & Moons, 2019). Alternatively, without sufficient validation, poorly developed models may be implemented practically, resulting in inaccurate treatment recommendations for the patients including either over or undertreatment (Ramspek et al., 2021; Wilson & Pendleton, 1989; Winkler et al., 2019).

7.2.2 The Ceiling Effect

Despite promising results, we believe that we may have maximised the potential performance using the current approach, suggesting that we have reached the theoretical ceiling. Specifically, the results from Chapters 4 and 6 support the notion of a ceiling effect, whilst the empathy classification results from Chapter 5 also provide insight into this phenomenon. In our initial investigation of pain intensity prediction in Chapter 4, we identified promising external validation results with accuracies of up to 69%. However, essential methodological improvements, which we hypothesised would improve model performance, were identified, and implemented. Consequently, we increased both the sample size and number of observations significantly, as the lack of data may have hindered ML performance. Increasing the amount of high-quality data should have theoretically increased model performance (Rajput et al., 2023). From Chapters 4 to 6, we increased the sample size by 60% for the model development sample and 340% for the external validation sample. Moreover, we also assessed within-subject performance using 25 subjects, which we did not previously investigate. Furthermore, we improved the study design by increasing the number of observations. Across the entire study, the number of observations was increased fourfold compared to the earlier chapter. Consequently, given the increased robustness and data quality, it was reasonable to expect a significant increase in performance on the binary classification task. Surprisingly, we did not observe increased performance, but rather we observed a significant decrease in model performance, reducing by approximately 9% on cross-validation and 11% when comparing cross-subject validation samples.

These results initially appear counterintuitive, as improving both sample size and study rigour should have improved model performance. Simulation research has shown that performance gains of 70% can be achieved through improved data quality (Rajput et al., 2023). Moreover, larger sample sizes were associated with increased model performance, but performance plateaued once a critical sample size had been obtained (Rajput et al., 2023). Therefore, we believe that, for the current approach, our results are converging towards the theoretical limit of ML-EEG for pain intensity classification, and increasing sample sizes may not yield further performance gain.

The increased robustness of our second experimental chapter may actually account for the reduced performance. Although simulation research suggests that improved data quality is associated with greater model performance (Rajput et al., 2023), the association is less clear in empirical research. A recent review of ML prediction models across research domains such as Alzheimer's, schizophrenia, psychosis, and autism demonstrated a negative association between sample size and model performance, with smaller samples often exhibiting greater performance (Varoquaux, 2018). Smaller samples exhibited greater variability in performance, inflating the model estimates. Consequently, it is feasible that by increasing the sample size, our estimates of model performance became more precise, reducing the observed performance. Further research supports this notion as small samples often lead to positive results due to chance, and this also increases the likelihood of publication (Algermissen & Mehler, 2018; Combrisson & Jerbi, 2015). For example, in small samples, ML models can achieve accuracies greater than 70% for random data (Combrisson & Jerbi, 2015). However, this study demonstrated that as the sample size increased, model performance trended towards the theoretical chance level (Combrisson & Jerbi, 2015). Therefore, this phenomenon alone may explain the reduction in performance. Given that larger samples are less impacted by variability and produce precise estimates (Kokol et al., 2022; Varoquaux, 2018), the results of our second experimental pain chapter likely provide robust estimates of the true capability of EEG-ML approaches, with performance converging towards 60%. As the existing literature frequently consists of small samples (Mari et al., 2022), the results are more likely to be inflated or obtained by chance (Combrisson & Jerbi, 2015). Furthermore, small samples produce significantly more variable results, including both type one and two errors. However, spurious positive results are significantly more likely to be published due to publication bias (Algermissen & Mehler, 2018; Jannot et al., 2013). Therefore, the estimates currently present in the field are also likely positively inflated due to publication bias, as positive findings are significantly more likely to be published (Duyx et al., 2017; Jannot et al., 2013). Consequently, due to the increased methodological rigour, our results provide the best estimates of the current capability of ML and EEG for pain intensity prediction. Based on our results, we are confident that low and high pain trials can be classified 60% of the time, which still represents an improvement on the theoretical chance level of 50%. However, increasing the sample size and rigour is unlikely to further improve the model performance (Combrisson & Jerbi, 2015; Varoquaux, 2018) resulting in the observed ceiling effect.

7.2.3 Classifying Data with Similar Spatiotemporal Characteristics

This thesis also highlights the challenges associated with classifying data with similar spatiotemporal characteristics in terms of electrophysiology. The signal similarity between the classes (e.g., low and high pain, neutral and pain scene) and low signal-to-noise ratio could also explain the poor performance metrics. As single-trial EEG is inherently noisy, with a low signal-to-noise ratio (Cohen & Cavanagh, 2011; Kaplan et al., 2005; Marathe et al., 2014), it is feasible that the differences in the spatiotemporal pattern of the classes observed at the grand average level (Coll, 2018; Fallon, Li, Chiu, et al., 2015; Misra, Wang, et al., 2017) are not observable at the single-trial level. Low and high pain states share highly similar spatiotemporal characteristics, with increased gamma over frontal regions, and decreased alpha and beta power over sensorimotor regions often observed, with larger changes associated with high pain (Misra, Wang, et al., 2017). Therefore, the neural characteristics of low and high pain significantly overlap, with deviations in the magnitude associated with the intensity. Moreover, the ERP waveforms for neutral and pain images share a high degree of similarity in their spatiotemporal profile, with only minor differences observed as enhanced or augmented component variations (Coll, 2018; Fallon, Li, Chiu, et al., 2015). Consequently, such differences may not be easily detectable at the single-trial level, resulting in relatively poor performance metrics. The findings from the face and non-face stimuli classification support this, as the model demonstrated good discrimination and calibration. The N170 component is significantly enhanced during the observation of faces and is attenuated or missing in non-face stimuli (Bentin et al., 1996; Bötzel et al., 1995; Eimer, 2000; Itier, 2004; Itier & Taylor, 2004), the spatiotemporal characteristics were significantly different and may have led to more effective ML performance, as such distinct spatiotemporal patterns may have been detectable at the single-trial level. It is intuitive to expect ML models to perform well on data that have very different or distinct characteristics in terms of neural responses. Therefore, our results demonstrate the difficulties associated with classifying similar neural patterns, such as graded levels of subjective pain. Improving the signal-to-noise ratio through spatial filtering may result in increased ML performance (Blankertz et al., 2008; Rivet et al., 2009).

7.2.4 Clinical and Practical Utility

This thesis also provides important evidence for both practical and clinical applications of ML. Our findings demonstrate that ML and EEG for pain intensity and visual stimuli prediction do not reach the levels required to achieve practical and clinical significance. All the models developed in this thesis (apart from the within-subject face and scene visual stimuli classification) failed to demonstrate performance sufficient for clinical and practical significance. Models that exhibit AUCs of less than 0.75 are not considered clinically/practically relevant (J. Fan et al., 2006). Consequently, our findings highlight that external validation performance is not currently sufficient to warrant translation attempts to develop clinical tools or similar applications, which is an important conclusion from this thesis. Based on the existing literature, it could be concluded that ML-EEG methods demonstrate sufficient performance for practical implementation, with almost three-quarters of the pain intensity studies evaluated in our review reporting accuracies of greater than 80% (Alazrai, Momani, et al., 2019; Alazrai, Al-Rawi, et al., 2019; T. Cao et al., 2020; Elsayed et al., 2020; Hadjileontiadis, 2015; Kaur et al., 2019; Misra, Ofori, et al., 2017; Nezam et al., 2018; Okolo & Omurtag, 2018; Sai et al., 2019; Schulz et al., 2012; Tripanpitak et al., 2020; Tu et al., 2016; Vatankhah et al., 2013; Vijayakumar et al., 2017; M. Yu, Yan, et al., 2020). Similar results have been reported for visual decoding research, with several studies demonstrating accuracies of approximately 90% (Cudlenco et al., 2020; Ghosh et al., 2021; Stewart et al., 2014; Zheng et al., 2020).

The findings of previous literature would suggest that the models demonstrate adequate performance for immediate practical implementation. However, based on our findings on the reduction of performance from internal to external validation, it is likely that these metrics are inflated, resulting in an over-optimistic view of the current clinical potential. This is likely an issue for all clinical ML prediction model research, as numerous models are developed each year, but few are successfully translated (Seneviratne et al., 2020; Shah et al., 2019). Therefore, even in previous models that surpass the (arguably arbitrary) clinical significance threshold, the clinical potential of the approach is likely overstated. Beyond performance metrics, numerous barriers exist, particularly safety, legal and ethical considerations, which reduce the likelihood of successful implementation (Davis et al., 2017; Mechelli & Vieira, 2020; Seneviratne et al., 2020). Moreover, the difficulty of translating a clinical prediction tool, e.g., to objectively predict subjective pain intensity during a clinical examination, is enhanced as they are often considered medical devices and are subject to medical device regulation legislation (van Maaren et al., 2023). Overall, our findings suggest the performance level of ML and EEG is not currently sufficient to warrant the progression to practical and clinical applications. Improving external validation performance is imperative before attempts to develop practical implementations of ML-EEG-based methods are conducted. Beyond external validation, numerous imperative stages of the clinical prediction model pipeline must be carefully navigated to ensure the prediction model meets regulatory standards, otherwise, such prediction models will fail to demonstrate clinical utility, and could even lead to patient harm via over/underestimation of pain.

7.3 Limitations

The current thesis has several limitations which should be considered. Firstly, all empirical chapters lack ecological validity for consideration in a clinical context, having been conducted in a sample of healthy individuals in a laboratory environment. The principal utility of an EEG-ML pain decoding tool exists outside of the laboratory, for use in individuals with impaired ability to accurately communicate their pain (Arbour & Gélinas, 2014; Breivik et al., 2008; Herr et al., 2011; Ploner & May, 2018). Therefore, it is unknown whether our approach would work outside the laboratory in clinical populations. For example, it is unclear whether our prediction models would work in clinical populations such as dementia, which is associated with structural changes in prefrontal brain regions that reduce pain inhibition and increase pain intensity (Bunk et al., 2021). Therefore, we cannot generalise our findings beyond healthy individuals. However, whilst our research lacks ecological validity due to the laboratory environment, EEG can be used outside these conditions through dry and portable systems (C.-H. Wang et al., 2019; Zander et al., 2011). Such systems demonstrate easier application, without the requirement of specialist set-up (e.g., saline to minimise electrical impedances) or environments (e.g., faraday cages), with little performance degradation (T. J. Sullivan et al., 2008; C.-H. Wang et al., 2019; Zander et al., 2011). Dry electrodes demonstrate promise and have already been assessed for pain classification by Kimura and colleagues (2021). Therefore, such systems remain promising for improving the ecological validity of the approach. However, as the studies in this thesis were conducted in highly controlled, laboratory environments, this thesis lacks ecological validity, hindering the generalisability to real-world applications. Moreover, our findings are limited as we cannot conclude whether the approach would be effective in individuals with conditions that specifically alter brain structure and/or function.

A recurring limitation of this thesis is that many of our models demonstrated imperfect calibration. Both pain intensity prediction studies demonstrated modest calibration, with instances of overly extreme probability estimates. Whilst the calibration for the faces and scenes classification was promising, the pain empathy models are also likely to demonstrate calibration issues. However, as the model demonstrated poor discrimination metrics, we did not formally assess the calibration (Alba et al., 2017), but given the inaccurate predictions, the model can be assumed to be uncalibrated. Therefore, calibration issues are a consistent limitation across all empirical chapters. Inaccurate calibration suggests that the model provides imprecise probability estimates and is indicative of overfitting (Van Calster et al., 2019). Calibration issues are particularly problematic for pain intensity predictions as the models tend to both underestimate and overestimate the probability of pain. Such predictions may result in negative outcomes such as under or overtreatment (Ramspek et al., 2021; Wilson & Pendleton, 1989). However, as calibration is rarely assessed (Christodoulou et al., 2019; Mari et al., 2022), the calibration assessments of the thesis are an area of novelty and represent a significant advancement over previous research. Moreover, calibration estimates can be improved after ML modelling through the application of calibration models such as Platt scaling or isotonic regression (Y. Huang et al., 2020). Therefore, despite modest calibration, methods to improve model calibration exist and should be applied, especially when discrimination metrics surpass the clinically relevant threshold (J. Fan et al., 2006). Overall, continuing to assess and improve model calibration are imperative developments required for the approach to demonstrate clinical utility.

Finally, the application of mechanical stimulation in this thesis is both an area of novelty and limitation. Mechanical stimulation in the form of mechanical pressure provides an approximate model of musculoskeletal pain, which holds relevance to chronic pain conditions such as Fibromyalgia syndrome (Birnie et al., 2014; Galvez-Sánchez et al., 2018), which is characterised by widespread pain and tenderness to pressure stimulation (Kosek et al., 1995; Wolfe, 1997; Wolfe et al., 1990). However, other pain modalities may offer improved models of alternative chronic pain conditions (e.g., electrical stimulation – neuropathic pain). Consequently, we cannot generalise our findings to alternative methods of experimental pain stimulation. From our findings in Chapter 4, ML performance was slightly reduced when using alternative stimuli parameters within the same modality, suggesting the difficulty of predicting subjective pain beyond the original intention of the model. Future research should assess whether ML models trained on one pain modality would enable pain classification when using an alternative stimulation approach. A model that is effective regardless of pain modality holds greater clinical potential and would facilitate broader applications.

7.4 Suggestions for Future Research

7.4.1 Composite Measures

Further external validation paradigms are imperative to assess the true potential of EEG-ML methods for pain prediction. Specifically, future research should prioritise improving both model discrimination and calibration to demonstrate performance levels sufficient for clinical translation. As discussed, clinical prediction models should achieve an AUC of greater than 0.75 to demonstrate clinically meaningful results (J. Fan et al., 2006). Therefore, developing models that exceed this threshold when assessed using external validation paradigms remains a priority. Composite measures, e.g., combining EEG data with other physiological signals, may enable improved performance and should be thoroughly explored (Rockholt et al., 2023; I. Tracey et al., 2019). Due to the complexity of pain, single-measure approaches are unlikely to effectively decode pain, with composite approaches holding greater promise

(Cowen et al., 2015; I. Tracey et al., 2019). It is possible that through a combination of measures, performance may be extended beyond the current observed performance maximum, which may result in clinically meaningful results.

Beyond neuroimaging techniques, genetic, physiological (heart rate, skin conductance, pupillometry), and biological measures (urine metabolites) may enable pain decoding (Cowen et al., 2015; Davis et al., 2020; Eldabe et al., 2022; I. Tracey et al., 2019). The potential of these methods is enhanced when combined, as composite measures often significantly outperform individual approaches (Cowen et al., 2015). Physiological measures may demonstrate the greatest potential due to their accessibility and potential for clinical utility (Ghamari, 2018; Mackey et al., 2019; Pantelopoulos & Bourbakis, 2010; Tivadar & Murray, 2019). Moreover, physiological measures appear promising for pain identification when combined with ML (Chu et al., 2017; Fernandez Rojas et al., 2023; M. Jiang et al., 2019; Teichmann et al., 2018; F. Yang et al., 2018). For example, Chu et al. (2017) aimed to develop a composite measure of pain intensity during electrical stimulation of 6 subjects. By combining blood volume pulse measurements, electrocardiogram, and skin conductance techniques, the authors demonstrated that four levels of subjective pain intensity could be classified with accuracies of up to 75% (Chu et al., 2017). Further research has demonstrated that heart rate, breath rate, skin conductance and facial EMG assessments during thermal and electrical pain stimulation could classify three classes of pain intensity with accuracies between 71% and 83% (M. Jiang et al., 2019). Here, heart rate, respiratory rate, and skin conductance had the greatest association with pain intensity and accounted for most of the classification performance. Similar research has demonstrated that composite measures can be used to predict pain intensity with accuracies up to 94% (Fernandez Rojas et al., 2023; Teichmann et al., 2018). Despite promising results, it is important to note that the previous studies are not externally validated, which incorporates all problems for interpretation discussed earlier. However, it is certainly possible that the combination of EEG and physiological measures may improve pain decoding and should be extensively investigated in future research (Fernandez Rojas et al., 2023).

The potential of combined EEG and physiological measures is enhanced due to their potential ease of use and applicability in non-specialist environments. Whilst numerous approaches and techniques demonstrate promise for pain prediction, few are easily implementable clinically. For example, whilst fMRI may be effective for pain assessment (e.g., Wager et al., 2013), it has reduced clinical potential due to financial and infrastructure limitations (Cowen et al., 2015; Mechelli & Vieira, 2020). Measures such as EEG, photoplethysmography, and skin conductance are low-cost and easily accessible in clinical environments, which make them ideal candidates for pain assessment (Ghamari, 2018; Mackey et al., 2019; Pantelopoulos & Bourbakis, 2010; Tivadar & Murray, 2019). Therefore, future research should attempt to validate low-cost composite measures for pain assessment, which would further increase the likelihood of clinical implementation of pain prediction models.

7.4.2 Real-Time Decoding and Wearable Sensors

In addition to composite measures, the potential of a pain assessment technique could be enhanced through the development of real-time pain detection using wearable sensors. Currently, our approach requires specialist equipment and environments, with the analysis occurring offline. Real-time EEG decoding, which can often produce an outcome (e.g., pain classification) in less than 40ms, would significantly enhance the clinical potential of a pain decoding tool (Müller et al., 2008). Research within the brain-computer interface domain provides evidence for the potential of real-time decoding tools, with tasks such as speech being successfully decoded online using neural responses and ML (Moses et al., 2019). Indeed, real-time pain decoding tools are starting to be developed. Recent research has developed a real-time pain decoding tool using ML and fNIRS (X.-S. Hu et al., 2019). The previous study used a neural network to classify non-pain and pain states in real time, achieving an accuracy of 80.37%. Moreover, the authors extended the classification task to identify the location of the pain in addition to the presence or absence of pain. Here, the authors were able to classify non-pain, left-sided, and right-sided pain with accuracies of approximately 74% (X.-S. Hu et al., 2019). Again, it is important to note that these metrics are not externally validated and should be interpreted with caution. Nevertheless, the evidence suggests that varying pain states can be accurately decoded in real time, further enhancing the clinical potential of ML and EEG for pain assessment.

Wearable sensors that are low-cost and easy to use may also further enhance the translation potential of a pain assessment tool, especially when coupled with real-time decoding. Wearable devices are types of biosensors that can be placed on the body, enabling continuous, high-resolution measurements (Leroux et al., 2021). Whilst there is limited research on the use of wearable sensors for pain prediction (Leroux et al., 2021), a recent study has attempted real-time pain assessment using a wrist-worn electrodermal activity monitor (Kong et al., 2021). Kong and colleagues (2021) aimed to decode non-pain and pain states using RF models and electrodermal activity recorded using a smartphone in real time. They found that non-pain and pain states could be accurately classified with accuracies up to 82% in ten subjects (Kong et al., 2021). Given that wearable sensors are relatively inexpensive and easy to use (Blasco & Peris-Lopez, 2018), they could be readily translated into a variety of clinical settings (e.g., a doctor’s office), providing performance reaches clinically meaningful levels. Moreover, wearable sensors could be combined with mobile EEG, enabling composite measures outside of the clinic. Therefore, to further enhance the potential applications of biological measurements for pain assessment, future research should aim to develop models using data from wearable, ideally consumer-level, sensors, to increase the clinical potential of the approach. Moreover, focusing efforts on real-time decoding using such tools will further increase the feasibility of object clinical pain assessment.

7.4.3 Magnetencephalography and Electrocorticography

Alternative neuroimaging approaches with improved signal-to-noise ratio may facilitate greater predictive performance. MEG and electrocorticography (ECoG) may have sufficient signal quality to improve ML performance (N. J. Hill et al., 2012; Ploner & May, 2018; Schalk & Leuthardt, 2011; Simon et al., 2022; S. Singh, 2014). Firstly, MEG records the magnetic fields which are not distorted by anatomical differences such as skull thickness or cerebrospinal fluid (Ploner & May, 2018; S. Singh, 2014). Moreover, MEG has improved spatial resolution when compared to EEG and is advantageous for neural responses with deeper neural generators (Pizzo et al., 2019; Ploner & May, 2018; Pu et al., 2018; S. Singh, 2014). Therefore, MEG data may offer improved signal quality, making it a potential candidate for a pain assessment technique. MEG has already been implemented to assess neural correlates of pain using traditional analyses (e.g., Timmermann et al., 2001). Consequently, there is evidence to suggest the utility of MEG for pain assessment. Despite a paucity of research investigating the combination of ML and MEG for pain prediction, one study has demonstrated the potential of the approach for classifying individuals with chronic migraines (Hsiao et al., 2022). Hsiao et al. (2022) classified healthy controls and chronic migraine patients achieving an accuracy of approximately 86% and an AUC of 0.9. Further promising accuracies were obtained for the classification of individuals with different phenotypes (e.g., chronic migraine – episodic migraine, chronic migraine – Fibromyalgia). However, the previous research was not externally validated, which hinders the interpretability of the approach. Nevertheless, MEG may facilitate improved pain prediction and should be thoroughly investigated to ascertain whether the approach outperforms EEG.

Despite the improved signal quality, MEG is less practical, more expensive, and less available than EEG (Ploner & May, 2018; S. Singh, 2014). For example, MEG systems cost several millions of pounds with running costs of approximately £100,000 per year (Seki et al., 2012; Stefan & Trinka, 2017). Whereas, the most expensive EEG systems can be acquired for approximately £150,000, with cheaper systems available for less than £15,000 (Emotiv Systems approx. £600; Ledwidge et al., 2018). Moreover, EEG systems are readily available in clinical settings, increasing translation feasibility (Ploner & May, 2018). Therefore, the financial and infrastructure requirements of MEG reduce the potential utility as a pain assessment technique (Levitt & Saab, 2019). Unless MEG enables pain prediction at levels that are unobtainable using EEG, it is unlikely to displace EEG. Developments such as low-cost MEG sensors, which retain the signal quality, but do not require helium cooling (enabling use at room temperature), may increase the feasibility of MEG pain assessment (Boto et al., 2017). Therefore, the potential of MEG for pain prediction is an exciting avenue for research.

Alternatively, ECoG may also enable improved pain prediction due to an increased signal-to-noise ratio and reduced susceptibility to EMG artefacts (N. J. Hill et al., 2012; Schalk & Leuthardt, 2011; Simon et al., 2022). Intracranial electrodes are regularly used in the brain-computer interface domain, demonstrating effective neural decoding (Liang & Bougrain, 2012; Schalk & Leuthardt, 2011; Vansteensel et al., 2016). Regarding pain prediction, recent research has demonstrated promising results (Shirvalkar et al., 2023). Using ML and local field potentials (LFPs) recorded from intracranial electrodes in the ACC and OFC of 4 patients with refractory neuropathic pain, the study demonstrated that ongoing low and high pain states could be classified with an average AUC of 0.75. Moreover, the authors attempted to classify responses during thermal pain. The model only significantly predicted pain in two subjects, achieving an AUC of approximately 0.74. Finally, the authors attempted to predict continuous pain intensity during chronic pain states. They found that the regression failed to predict pain ratings across all subjects, concluding that regression models may not be feasible for pain prediction and that binary classification approaches are more pragmatic for clinical applications (Shirvalkar et al., 2023). Taken together, intracranial electrodes may enable pain prediction, but the approach cannot be widely implemented.

Whilst populations who require intracranial electrodes for adjunct functions such as communication (e.g., brain-computer interfaces for individuals with locked-in syndrome; Vansteensel et al., 2016) could be useful for pain prediction, it is not feasible to implant electrodes into most populations (e.g., infants). Therefore, even if intracranial electrodes enable accurate pain classification, they currently lack applicability and have few potential applications. Conditions such as locked-in syndrome could benefit from such approaches. However, developments in the brain-computer interface domain are allowing providing effective communicative tools for such populations (Metzger et al., 2023; Vansteensel et al., 2016), negating the potential of a pain assessment tool, as patients would regain the ability to accurately communicate their pain. Consequently, whilst intracranial electrodes may improve predictive performance, it is unclear whether the approach is necessary due to the lack of potential applications and developments from alternative research domains. Nevertheless, further research is warranted to determine the potential utility of the method.

7.4.4 Improved Signal Processing and Feature Engineering

Advancements in feature engineering and processing may also improve ML performance to clinically meaningful levels. Our feature engineering approach is relatively rudimentary, opting to focus on statistical features. Whilst we observed promising results, statistical features are the simplest approach and employing more advanced feature engineering procedures may improve ML performance (A. K. Singh & Krishnan, 2023). Spatial filtering (e.g., common spatial pattern, common spare spatio-spectral pattern) is a potential approach to enhance ML performance (A. K. Singh & Krishnan, 2023). Spatial filtering techniques increase the signal-to-noise ratio of EEG, and consequently ML performance, by maximising the differences between classes (Blankertz et al., 2008; Y. Miao et al., 2021; Rashid et al., 2020; Rivet et al., 2009). Spatial filtering enhances variance in the first class and decreases variance in the second class (Rashid et al., 2020). Spatial filtering remains promising for classification approaches, with research demonstrating improved ML performance due to the filtering (Blankertz et al., 2011). However, spatial filtering increases the overfitting likelihood, which is especially problematic when external validation approaches are not employed (Blankertz et al., 2008; Grosse-Wentrup et al., 2009). The use of spatial filters has already demonstrated promise for pain prediction (G. Huang et al., 2013). Therefore, spatial filtering techniques should be combined with external validation paradigms in future research, to provide insight into the performance gain and the effect on generalisability. Such approaches may enhance performance to a clinically meaningful level to provide a robust estimate of the potential of the approach.

7.4.5 Increased Collaboration

Finally, increased collaboration and data sharing are imperative for the successful development and translation of an AI-guided pain assessment tool (Davis et al., 2020; van der Miesen et al., 2019). Currently, prediction models are continuously developed across research groups, with few progressing to clinical translation, resulting in research waste (Collins et al., 2014; Collins & Moons, 2019; Seneviratne et al., 2020; Shah et al., 2019). The benefit of large-scale collaboration is evident. Increased transparency, external validation by other groups, and exponentially increased sample sizes are some of the many potential benefits. Furthermore, collaborative work results in enhanced productivity, leading to the acquisition of scientific knowledge that is not obtainable through siloed efforts (Katz & Martin, 1997; S. Lee & Bozeman, 2005; Wuchty et al., 2007). Consequently, effective collaboration is imperative given the number of challenges associated with developing and implementing clinical prediction models (e.g., ethical, and legal; Char et al., 2018; Vayena et al., 2018). Moreover, it is likely that even after the successful clinical implementation of a pain prediction model collaboration will be required. Handling running costs and long-term maintenance and monitoring will provide further obstacles to be overcome (Mechelli & Vieira, 2020). The creation of guidelines and committees, comprised of specialists with complementary expertise (technical, ethical, legal etc.), will be essential to ensure the appropriate maintenance of such a tool. Moreover, the development of an objective measure of pain empathy also likely requires enhanced approaches including greater collaboration. Perhaps more advanced techniques would yield improved results, necessitating collaboration with external experts. The advent of such tools would be useful for medical training and to reduce bias in patient-clinician interactions and pain assessment (Hoffman et al., 2016; Pierson et al., 2021; Preusche & Lamm, 2016). Whilst the development of prediction models for pain intensity and empathy remains in its infancy, a collaborative effort will be required throughout to overcome the many future obstacles.

7.5 Concluding Remarks

To conclude, this thesis investigated the effectiveness of ML and EEG for the classification of both subjective pain intensity and pain observation. We are the first to incorporate and report on external validation procedures, significantly advancing the research field. Our results are promising but provide realistic estimates of the current effectiveness of ML and EEG. Specifically, we demonstrated that ML and EEG can predict subjective pain intensity with above-chance levels in novel subjects and using novel experimental pain stimulation. Regarding the pain intensity prediction, we conclude that the current performance estimates in the literature are likely inflated. This thesis demonstrates the difficulty of diagnosing and preventing model overfitting without external validation, as performance generalisability estimates are often positively biased. Consequently, the lack of external validation has likely contributed to over-optimism in the research field. In addition, we identified that ML and EEG are unable to successfully classify the observation of pain, providing potential limits of the method. The potential of a ceiling effect and the difficulty in classifying data with similar spatiotemporal patterns were identified as key themes of this research. Taken together, this thesis provides foundational contributions to the field through the novel integration of external validation procedures, improved sample sizes and new experimental paradigms. Further developments and externally validated research articles are imperative before clinical translation attempts can be justified. Improving model performance is also required to demonstrate clinically meaningful results, with composite data approaches offering a key area for improvement. Our research indicates that an objective tool to measure empathic processing, e.g., for clinical training applications, may not be imminently feasible. Alternatively, through continued development, a neuroimaging-based pain assessment tool may eventually demonstrate sufficient capability to warrant clinical translation. However, further significant methodological developments are required before such a tool can exist.

References

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. https://doi.org/10.3389/fninf.2014.00014

Abraira, V. E., & Ginty, D. D. (2013). The Sensory Neurons of Touch. *Neuron*, *79*(4), 618–639. https://doi.org/10.1016/j.neuron.2013.07.051

Adrian, E. D., & Matthews, B. H. C. (1934). The Berger Rhythm: Potential Changes from The Occipital Lobes in Man. *Brain*, *57*(4), 355–385. https://doi.org/10.1093/brain/57.4.355

Ahuja, A. S. (2019). The impact of artificial intelligence in medicine on the future role of the physician. *PeerJ*, *7*, e7702. https://doi.org/10.7717/peerj.7702

Akben, S. B., Subasi, A., & Tuncel, D. (2012). Analysis of Repetitive Flash Stimulation Frequencies and Record Periods to Detect Migraine Using Artificial Neural Network. *Journal of Medical Systems*, *36*(2), 925–931. https://doi.org/10.1007/s10916-010-9556-2

Akben, S. B., Tuncel, D., & Alkan, A. (2016). Classification of multi-channel EEG signals for migraine detection. *Biomedical Research*, *27*(3), 743–748.

Akitsuki, Y., & Decety, J. (2009). Social context and perceived agency affects empathy for pain: An event-related fMRI investigation. *NeuroImage*, *47*(2), 722–734. https://doi.org/10.1016/j.neuroimage.2009.04.091

Alazrai, R., AL-Rawi, S., Alwanni, H., & Daoud, M. I. (2019). Tonic Cold Pain Detection Using Choi–Williams Time-Frequency Distribution Analysis of EEG Signals: A Feasibility Study. *Applied Sciences*, *9*(16), 3433. https://doi.org/10.3390/app9163433

Alazrai, R., Al-Rawi, S., & Daoud, M. I. (2019). A Time-Frequency Distribution Based Approach for Detecting Tonic Cold Pain using EEG Signals. *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, 589–592. https://doi.org/10.1109/BIBE.2019.00112

Alazrai, R., Momani, M., Khudair, H. A., & Daoud, M. I. (2019). EEG-based tonic cold pain recognition system using wavelet transform. *Neural Computing and Applications*, *31*(7), 3187–3200. https://doi.org/10.1007/s00521-017-3263-6

Alba, A. C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P. J., McGinn, T., & Guyatt, G. (2017). Discrimination and Calibration of Clinical Prediction Models. *JAMA*, *318*(14), 1377. https://doi.org/10.1001/jama.2017.12126

Algermissen, J., & Mehler, D. M. A. (2018). May the power be with you: are there highly powered studies in neuroscience, and how can we get more of them? *Journal of Neurophysiology*, *119*(6), 2114–2117. https://doi.org/10.1152/jn.00765.2017

Alloghani, M., Al-Jumeily, D., Mustafina, J., Hussain, A., & Aljaaf, A. J. (2020). A Systematic Review on Supervised and Unsupervised Machine Learning Algorithms for Data Science. In *Unsupervised and Semi-Supervised Learning* (pp. 3–21). https://doi.org/10.1007/978-3-030-22475-2_1

Almeida, T. F., Roizenblatt, S., & Tufik, S. (2004). Afferent pain pathways: a neuroanatomical review. *Brain Research*, *1000*(1–2), 40–56. https://doi.org/10.1016/j.brainres.2003.10.073

Anuragi, A., & Sisodia, D. S. (2020). Empirical wavelet transform based automated alcoholism detecting using EEG signal features. *Biomedical Signal Processing and Control*, *57*, 101777. https://doi.org/10.1016/j.bspc.2019.101777

Apkarian, A. V., Bushnell, M. C., Treede, R.-D., & Zubieta, J.-K. (2005). Human brain mechanisms of pain perception and regulation in health and disease. *European Journal of Pain*, *9*(4), 463–463. https://doi.org/10.1016/j.ejpain.2004.11.001

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*, *145*, 137–165. https://doi.org/10.1016/j.neuroimage.2016.02.079

Arbour, C., & Gélinas, C. (2014). Behavioral and Physiologic Indicators of Pain in Nonverbal Patients with a Traumatic Brain Injury: An Integrative Review. *Pain Management Nursing*, *15*(2), 506–518. https://doi.org/10.1016/j.pmn.2012.03.004

Assel, M., Sjoberg, D. D., & Vickers, A. J. (2017). The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models. *Diagnostic and Prognostic Research*, *1*(1), 19. https://doi.org/10.1186/s41512-017-0020-3

Ataoğlu, E., Tiftik, T., Kara, M., Tunç, H., Ersöz, M., & Akkuş, S. (2013). Effects of chronic pain on quality of life and depression in patients with spinal cord injury. *Spinal Cord*, *51*(1), 23–26. https://doi.org/10.1038/sc.2012.51

Atkins, N., & Mukhida, K. (2022). The relationship between patients' income and education and their access to pharmacological chronic pain management: A scoping review. *Canadian Journal of Pain*, *6*(1), 142–170. https://doi.org/10.1080/24740527.2022.2104699

Azevedo, F. A. C., Carvalho, L. R. B., Grinberg, L. T., Farfel, J. M., Ferretti, R. E. L., Leite, R. E. P., Filho, W. J., Lent, R., & Herculano-Houzel, S. (2009). Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *The Journal of Comparative Neurology*, *513*(5), 532–541. https://doi.org/10.1002/cne.21974

Babadi, B., & Brown, E. N. (2014). A Review of Multitaper Spectral Analysis. *IEEE*

Transactions on Biomedical Engineering, *61*(5), 1555–1564. https://doi.org/10.1109/TBME.2014.2311996

Babiloni, C., Babiloni, F., Carducci, F., Cincotti, F., Rosciarelli, F., Arendt-Nielsen, L., Chen, A. C. N., & Rossini, P. M. (2002). Human brain oscillatory activity phase-locked to painful electrical stimulations: A multi-channel EEG study. *Human Brain Mapping*, *15*(2), 112–123. https://doi.org/10.1002/hbm.10013

Backonja, M., Howland, E. W., Wang, J., Smith, J., Salinsky, M., & Cleeland, C. S. (1991). Tonic changes in alpha power during immersion of the hand in cold water. *Electroencephalography and Clinical Neurophysiology*, *79*(3), 192–203. https://doi.org/10.1016/0013-4694(91)90137-S

Bagchi, S., & Bathula, D. R. (2022). EEG-ConvTransformer for single-trial EEG-based visual stimulus classification. *Pattern Recognition*, *129*, 108757. https://doi.org/10.1016/j.patcog.2022.108757

Bai, Y., Huang, G., Tu, Y., Tan, A., Hung, Y. S., & Zhang, Z. (2016). Normalization of pain-evoked neural responses using spontaneous EEG improves the performance of EEG-based cross-individual pain prediction. *Frontiers in Computational Neuroscience*, *10*(APR). https://doi.org/10.3389/fncom.2016.00031

Banik, R. K., & Brennan, T. J. (2008). Sensitization of primary afferents to mechanical and heat stimuli after incision in a novel in vitro mouse glabrous skin-nerve preparation ☆. *Pain*, *138*(2), 380–391. https://doi.org/10.1016/j.pain.2008.01.017

Bargshady, G., Zhou, X., Deo, R. C., Soar, J., Whittaker, F., & Wang, H. (2020). Enhanced deep learning algorithm development to detect pain intensity from facial expression images. *Expert Systems with Applications*, *149*, 113305. https://doi.org/10.1016/j.eswa.2020.113305

Barnett, M. W., & Larkman, P. M. (2007). The action potential. *Practical Neurology*, *7*(3), 192–197.

Bas-Sarmiento, P., Fernández-Gutiérrez, M., Baena-Baños, M., Correro-Bermejo, A., Soler-Martins, P. S., & de la Torre-Moyano, S. (2020). Empathy training in health sciences: A systematic review. *Nurse Education in Practice*, *44*, 102739. https://doi.org/10.1016/j.nepr.2020.102739

Basbaum, A. I., Bautista, D. M., Scherrer, G., & Julius, D. (2009). Cellular and Molecular Mechanisms of Pain. *Cell*, *139*(2), 267–284. https://doi.org/10.1016/j.cell.2009.09.028

Batson, C. D. (2009). These Things Called Empathy: Eight Related but Distinct Phenomena. In *The Social Neuroscience of Empathy* (pp. 3–16). The MIT Press. https://doi.org/10.7551/mitpress/9780262012973.003.0002

Bauder, R. A., & Khoshgoftaar, T. M. (2018). The effects of varying class distribution on learner behavior for medicare fraud detection with imbalanced big data. *Health Information Science and Systems*, *6*(1), 9. https://doi.org/10.1007/s13755-018-0051-3

Bendat, J. S., & Piersol, A. G. (2011). *Random data: analysis and measurement procedures* (4th ed.). John Wiley & Sons Ltd.

Bendinger, T., & Plunkett, N. (2016). Measurement in pain medicine. *BJA Education*, *16*(9), 310–315. https://doi.org/10.1093/bjaed/mkw014

Bennington, J. Y., & Polich, J. (1999). Comparison of P300 from passive and active tasks for auditory and visual stimuli. *International Journal of Psychophysiology*, *34*(2), 171–177. https://doi.org/10.1016/S0167-8760(99)00070-7

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5

Bentin, S., Allison, T., Puce, A., Perez, E., & McCarthy, G. (1996). Electrophysiological Studies of Face Perception in Humans. *Journal of Cognitive Neuroscience*, *8*(6), 551–565. https://doi.org/10.1162/jocn.1996.8.6.551

Berg, P., & Scherg, M. (1994). A multiple source approach to the correction of eye artifacts. *Electroencephalography and Clinical Neurophysiology*, *90*(3), 229–241. https://doi.org/10.1016/0013-4694(94)90094-9

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, *13*(2), 281–305.

Bertrand, O., Perrin, F., & Pernier, J. (1985). A theoretical justification of the average reference in topographic evoked potential studies. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, *62*(6), 462–464. https://doi.org/10.1016/0168-5597(85)90058-9

Betts, J. G., Desaix, P., Johnson, E., Johnson, J. E., Korol, O., Kruse, D., Poe, B., Wise, J. A., Womble, M., & Young, K. A. (2013). *Anatomy and Physiology* (1st ed.). OpenStax.

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00016

Birnie, K. A., Caes, L., Wilson, A. C., Williams, S. E., & Chambers, C. T. (2014). A practical guide and perspectives on the use of experimental pain modalities with children and adolescents. *Pain Management*, *4*(2), 97–111. https://doi.org/10.2217/pmt.13.72

Bishop, G. H., & Landau, W. M. (1958). Evidence for a Double Peripheral Pathway for Pain. *Science*, *128*(3326), 712–713. https://doi.org/10.1126/science.128.3326.712

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components — A tutorial. *NeuroImage*, *56*(2), 814–825.

https://doi.org/10.1016/j.neuroimage.2010.06.048

Blankertz, B., Tomioka, R., Lemm, S., Kawanabe, M., & Muller, K. (2008). Optimizing Spatial filters for Robust EEG Single-Trial Analysis. *IEEE Signal Processing Magazine*, *25*(1), 41–56. https://doi.org/10.1109/MSP.2008.4408441

Blasco, J., & Peris-Lopez, P. (2018). On the Feasibility of Low-Cost Wearable Sensors for Multi-Modal Biometric Verification. *Sensors*, *18*(9), 2782. https://doi.org/10.3390/s18092782

Bleeker, S. ., Moll, H. ., Steyerberg, E. ., Donders, A. R. ., Derksen-Lubsen, G., Grobbee, D. ., & Moons, K. G. . (2003). External validation is necessary in prediction research: *Journal of Clinical Epidemiology*, *56*(9), 826–832. https://doi.org/10.1016/S0895-4356(03)00207-5

Boord, P., Siddall, P. J., Tran, Y., Herbert, D., Middleton, J., & Craig, A. (2008). Electroencephalographic slowing and reduced reactivity in neuropathic pain following spinal cord injury. *Spinal Cord*, *46*(2), 118–123. https://doi.org/10.1038/sj.sc.3102077

Bornhövd, K., Quante, M., Glauche, V., Bromm, B., Weiller, C., & Büchel, C. (2002). Painful stimuli evoke different stimulus–response functions in the amygdala, prefrontal, insula and somatosensory cortex: a single-trial fMRI study. *Brain*, *125*(6), 1326–1336. https://doi.org/10.1093/brain/awf137

Boto, E., Meyer, S. S., Shah, V., Alem, O., Knappe, S., Kruger, P., Fromhold, T. M., Lim, M., Glover, P. M., Morris, P. G., Bowtell, R., Barnes, G. R., & Brookes, M. J. (2017). A new generation of magnetoencephalography: Room temperature measurements using optically-pumped magnetometers. *NeuroImage*, *149*, 404–414. https://doi.org/10.1016/j.neuroimage.2017.01.034

Bötzel, K., Schulze, S., & Stodieck, S. R. G. (1995). Scalp topography and analysis of intracranial sources of face-evoked potentials. *Experimental Brain Research*, *104*(1).

https://doi.org/10.1007/BF00229863

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 2005–2032. https://doi.org/10.1023/A:1010933404324

Breivik, H., Borchgrevink, P. C., Allen, S. M., Rosseland, L. A., Romundstad, L., Breivik Hals, E. K., Kvarstein, G., & Stubhaug, A. (2008). Assessment of pain. *British Journal of Anaesthesia*, *101*(1), 17–24. https://doi.org/10.1093/bja/aen103

Breivik, H., Collett, B., Ventafridda, V., Cohen, R., & Gallacher, D. (2006). Survey of chronic pain in Europe: Prevalence, impact on daily life, and treatment. *European Journal of Pain*, *10*(4), 287–287. https://doi.org/10.1016/j.ejpain.2005.06.009

Brooks, J., & Tracey, I. (2005). REVIEW: From nociception to pain perception: imaging the spinal and supraspinal pathways. *Journal of Anatomy*, *207*(1), 19–33. https://doi.org/10.1111/j.1469-7580.2005.00428.x

Browne, M. W. (2000). Cross-Validation Methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Bunk, S. F., Lautenbacher, S., Rüsseler, J., Müller, K., Schultz, J., & Kunz, M. (2018). Does EEG activity during painful stimulation mirror more closely the noxious stimulus intensity or the subjective pain sensation? *Somatosensory & Motor Research*, *35*(3–4), 192–198. https://doi.org/10.1080/08990220.2018.1521790

Bunk, S. F., Zuidema, S., Koch, K., Lautenbacher, S., De Deyn, P. P., & Kunz, M. (2021). Pain processing in older adults with dementia-related cognitive impairment is associated with frontal neurodegeneration. *Neurobiology of Aging*, *106*, 139–152.

https://doi.org/10.1016/j.neurobiolaging.2021.06.009

Bushnell, M. C., Čeko, M., & Low, L. A. (2013). Cognitive and emotional control of pain and its disruption in chronic pain. *Nature Reviews Neuroscience*, *14*(7), 502–511. https://doi.org/10.1038/nrn3516

Bussone, G., & Grazzi, L. (2013). Understanding the relationship between pain and emotion in idiopathic headaches. *Neurological Sciences*, *34*(S1), 29–31. https://doi.org/10.1007/s10072-013-1362-4

Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. *Nature Reviews Neuroscience*, *13*(6), 407–420. https://doi.org/10.1038/nrn3241

Cabitza, F., Campagner, A., Soares, F., García de Guadiana-Romualdo, L., Challa, F., Sulejmani, A., Seghezzi, M., & Carobene, A. (2021). The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Computer Methods and Programs in Biomedicine*, *208*, 106288. https://doi.org/10.1016/j.cmpb.2021.106288

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. *BMJ*, l6890. https://doi.org/10.1136/bmj.l6890

Cao, T., Wang, Q., Liu, D., Sun, J., & Bai, O. (2020). Resting state EEG-based sudden pain recognition method and experimental study. *Biomedical Signal Processing and Control*, *59*, 101925. https://doi.org/10.1016/j.bspc.2020.101925

Cao, Y., Contreras-Huerta, L. S., McFadyen, J., & Cunnington, R. (2015). Racial bias in neural response to others' pain is reduced with other-race contact. *Cortex*, *70*, 68–78. https://doi.org/10.1016/j.cortex.2015.02.010

Cao, Z., Lai, K.-L., Lin, C.-T., Chuang, C.-H., Chou, C.-C., & Wang, S.-J. (2018). Exploring resting-state EEG complexity before migraine attacks. *Cephalalgia*, *38*(7), 1296–1306. https://doi.org/10.1177/0333102417733953

Carmon, A., Mor, J., & Goldberg, J. (1976). Evoked cerebral responses to noxious thermal stimuli in humans. *Experimental Brain Research*, *25*(1). https://doi.org/10.1007/BF00237330

Caune, V., Ranta, R., Le Cam, S., Hofmanis, J., Maillard, L., Koessler, L., & Louis-Dorr, V. (2014). Evaluating dipolar source localization feasibility from intracerebral SEEG recordings. *NeuroImage*, *98*, 118–133. https://doi.org/10.1016/j.neuroimage.2014.04.058

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *The Journal of Machine Learning Research*, *11*, 2079–2107.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, P. F., Arendt-Nielsen, L., & Chen, A. C. . (2002). Dynamic changes and spatial correlation of EEG activities during cold pressor test in man. *Brain Research Bulletin*, *57*(5), 667–675. https://doi.org/10.1016/S0361-9230(01)00763-8

Chang, P. F., Arendt-Nielsen, L., Graven-Nielsen, T., & Chen, A. C. . (2003). Psychophysical and EEG responses to repeated experimental muscle pain in humans: Pain intensity encodes EEG activity. *Brain Research Bulletin*, *59*(6), 533–543. https://doi.org/10.1016/S0361-9230(02)00950-4

Chang, P. F., Arendt-Nielsen, L., Graven-Nielsen, T., Svensson, P., & Chen, A. (2001). Different EEG topographic effects of painful and non-painful intramuscular stimulation in man. *Experimental Brain Research*, *141*(2), 195–203. https://doi.org/10.1007/s002210100864

Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing Machine Learning in Health Care — Addressing Ethical Challenges. *New England Journal of Medicine*, *378*(11), 981–983. https://doi.org/10.1056/NEJMp1714229

Chatrian, G. E., Canfield, R. C., Knauss, T. A., & Lettich, E. (1975). Cerebral responses to electrical tooth pulp stimulation in man: An objective correlate of acute experimental pain. *Neurology*, *25*(8), 745–745. https://doi.org/10.1212/WNL.25.8.745

Chen, A. C. N., & Rappelsberger, P. (1994). Brain and Human pain: Topographic EEG amplitude and coherence mapping. *Brain Topography*, *7*(2), 129–140. https://doi.org/10.1007/BF01186771

Chen, C., Yang, C.-Y., & Cheng, Y. (2012). Sensorimotor resonance is an outcome but not a platform to anticipating harm to others. *Social Neuroscience*, *7*(6), 578–590. https://doi.org/10.1080/17470919.2012.686924

Chen, Q., & Heinricher, M. M. (2022). Shifting the Balance: How Top-Down and Bottom-Up Input Modulate Pain via the Rostral Ventromedial Medulla. *Frontiers in Pain Research*, *3*. https://doi.org/10.3389/fpain.2022.932476

Cheng, Y., Chen, C., & Decety, J. (2014). An EEG/ERP investigation of the development of empathy in early and middle childhood. *Developmental Cognitive Neuroscience*, *10*, 160–169. https://doi.org/10.1016/j.dcn.2014.08.012

Cheng, Y., Hung, A.-Y., & Decety, J. (2012). Dissociation between affective sharing and emotion understanding in juvenile psychopaths. *Development and Psychopathology*, *24*(2), 623–636. https://doi.org/10.1017/S095457941200020X

Choo, E. K., Magruder, W., Montgomery, C. J., Lim, J., Brant, R., & Ansermino, J. M. (2010). Skin Conductance Fluctuations Correlate Poorly with Postoperative Self-report Pain Measures in School-aged Children. *Anesthesiology*, *113*(1), 175–182. https://doi.org/10.1097/ALN.0b013e3181de6ce9

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

Christov-Moore, L., Reggente, N., Douglas, P. K., Feusner, J. D., & Iacoboni, M. (2020). Predicting Empathy From Resting State Brain Connectivity: A Multivariate Approach. *Frontiers in Integrative Neuroscience*, *14*. https://doi.org/10.3389/fnint.2020.00003

Chu, Y., Zhao, X., Han, J., & Su, Y. (2017). Physiological Signal-Based Method for Measurement of Pain Intensity. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00279

Coghill, R. C., Sang, C. N., Maisog, J. M., & Iadarola, M. J. (1999). Pain Intensity Processing Within the Human Brain: A Bilateral, Distributed Mechanism. *Journal of Neurophysiology*, *82*(4), 1934–1943. https://doi.org/10.1152/jn.1999.82.4.1934

Coghill, R. C., Talbot, J., Evans, A., Meyer, E., Gjedde, A., Bushnell, M., & Duncan, G. (1994). Distributed processing of pain and vibration by the human brain. *The Journal of Neuroscience*, *14*(7), 4095–4108. https://doi.org/10.1523/JNEUROSCI.14-07-04095.1994

Cohen, M. X. (2014). *Analyzing neural time series data: theory and practice.* MIT Press.

Cohen, M. X. (2017). Where Does EEG Come From and What Does It Mean? *Trends in Neurosciences*, *40*(4), 208–218. https://doi.org/10.1016/j.tins.2017.02.004

Cohen, M. X., & Cavanagh, J. F. (2011). Single-Trial Regression Elucidates the Role of Prefrontal Theta Oscillations in Response Conflict. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00030

Coll, M.-P. (2018). Meta-analysis of ERP investigations of pain empathy underlines methodological issues in ERP research. *Social Cognitive and Affective Neuroscience*, *13*(10), 1003–1017. https://doi.org/10.1093/scan/nsy072

Collins, G. S., de Groot, J. A., Dutton, S., Omar, O., Shanyinde, M., Tajar, A., Voysey, M., Wharton, R., Yu, L.-M., Moons, K. G., & Altman, D. G. (2014). External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. *BMC Medical Research Methodology*, *14*(1), 40. https://doi.org/10.1186/1471-2288-14-40

Collins, G. S., & Moons, K. G. M. (2019). Reporting of artificial intelligence prediction models. *The Lancet*, *393*(10181), 1577–1579. https://doi.org/10.1016/S0140-6736(19)30037-6

Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement. *BMC Medicine*, *13*(1), 1. https://doi.org/10.1186/s12916-014-0241-z

Combrisson, E., & Jerbi, K. (2015). Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *Journal of Neuroscience Methods*, *250*, 126–136. https://doi.org/10.1016/j.jneumeth.2015.01.010

Cooley, J. W., & Tukey, J. W. (1965). An Algorithm for the Machine Calculation of Complex Fourier Series. *Mathematics of Computation*, *19*(90), 297–301. https://doi.org/10.2307/2003354

Cowen, R., Stasiowska, M. K., Laycock, H., & Bantel, C. (2015). Assessing pain objectively: the use of physiological markers. *Anaesthesia*, *70*(7), 828–847. https://doi.org/10.1111/anae.13018

Craig, A. D. (2002). How do you feel? Interoception: the sense of the physiological condition of the body. *Nature Reviews Neuroscience*, *3*(8), 655–666. https://doi.org/10.1038/nrn894

Crawford, L. K., & Caterina, M. J. (2020). Functional Anatomy of the Sensory Nervous System: Updates From the Neuroscience Bench. *Toxicologic Pathology*, *48*(1), 174–189. https://doi.org/10.1177/0192623319869011

Crofford, L. J. (2015). Chronic Pain: Where the Body Meets the Brain. *Transactions of the American Clinical and Climatological Association*, *126*, 167–183. https://doi.org/26330672

Cruz Rivera, S., Liu, X., Chan, A.-W., Denniston, A. K., & Calvert, M. J. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nature Medicine*, *26*(9), 1351–1363. https://doi.org/10.1038/s41591-020-1037-7

Cudlenco, N., Popescu, N., & Leordeanu, M. (2020). Reading into the mind’s eye: Boosting automatic visual recognition with EEG signals. *Neurocomputing*, *386*, 281–292. https://doi.org/10.1016/j.neucom.2019.12.076

Cuff, B. M. P., Brown, S. J., Taylor, L., & Howat, D. J. (2016). Empathy: A Review of the Concept. *Emotion Review*, *8*(2), 144–153. https://doi.org/10.1177/1754073914558466

Cui, F., Zhu, X., Duan, F., & Luo, Y. (2016). Instructions of cooperation and competition influence the neural responses to others' pain: An ERP study. *Social Neuroscience*, *11*(3), 289–296. https://doi.org/10.1080/17470919.2015.1078258

D'Mello, R., & Dickenson, A. H. (2008). Spinal cord mechanisms of pain. *British Journal of Anaesthesia*, *101*(1), 8–16. https://doi.org/10.1093/bja/aen088

Dahlhamer, J., Lucas, J., Zelaya, C., Nahin, R., Mackey, S., DeBar, L., Kerns, R., Von Korff, M., Porter, L., & Helmick, C. (2018). Prevalence of Chronic Pain and High-Impact Chronic Pain Among Adults — United States, 2016. *MMWR. Morbidity and Mortality Weekly Report*, *67*(36), 1001–1006. https://doi.org/10.15585/mmwr.mm6736a2

Dansie, E. J., & Turk, D. C. (2013). Assessment of patients with chronic pain. *British Journal of Anaesthesia*, *111*(1), 19–25. https://doi.org/10.1093/bja/aet124

Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, *6*(2), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Davis, K. D., Aghaeepour, N., Ahn, A. H., Angst, M. S., Borsook, D., Brenton, A., Burczynski, M. E., Crean, C., Edwards, R., Gaudilliere, B., Hergenroeder, G. W., Iadarola, M. J., Iyengar, S., Jiang, Y., Kong, J.-T., Mackey, S., Saab, C. Y., Sang, C. N., Scholz, J., … Pelleymounter, M. A. (2020). Discovery and validation of biomarkers to aid the development of safe and effective pain therapeutics: challenges and opportunities. *Nature Reviews Neurology*, *16*(7), 381–400. https://doi.org/10.1038/s41582-020-0362-2

Davis, K. D., Flor, H., Greely, H. T., Iannetti, G. D., Mackey, S., Ploner, M., Pustilnik, A., Tracey, I., Treede, R.-D., & Wager, T. D. (2017). Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. *Nature Reviews Neurology*, *13*(10), 624–638. https://doi.org/10.1038/nrneurol.2017.122

Davis, K. D., Racine, E., & Collett, B. (2012). 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*(8), 1555–1559. https://doi.org/10.1016/j.pain.2012.02.037

De Tommaso, M., Sciruicchio, V., Bellotti, R., Guido, M., Sasanelli, G., Specchio, L. M., & Puca, F. (1999). Photic driving response in primary headache: diagnostic value tested by discriminant analysis and artificial neural network classifiers. *The Italian Journal of Neurological Sciences*, *20*(1), 23–28. https://doi.org/10.1007/s100720050006

Decety, J., Bartal, I. B.-A., Uzefovsky, F., & Knafo-Noam, A. (2016). Empathy as a driver of prosocial behaviour: highly conserved neurobehavioural mechanisms across species. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686), 20150077. https://doi.org/10.1098/rstb.2015.0077

Decety, J., & Holvoet, C. (2021). The emergence of empathy: A developmental neuroscience perspective. *Developmental Review*, *62*, 100999. https://doi.org/10.1016/j.dr.2021.100999

Decety, J., & Jackson, P. L. (2004). The Functional Architecture of Human Empathy. *Behavioral and Cognitive Neuroscience Reviews*, *3*(2), 71–100. https://doi.org/10.1177/1534582304267187

Decety, J., Norman, G. J., Berntson, G. G., & Cacioppo, J. T. (2012). A neurobehavioral evolutionary perspective on the mechanisms underlying empathy. *Progress in Neurobiology*, *98*(1), 38–48. https://doi.org/10.1016/j.pneurobio.2012.05.001

Decety, J., Yang, C.-Y., & Cheng, Y. (2010). Physicians down-regulate their pain empathy response: An event-related brain potential study. *NeuroImage*, *50*(4), 1676–1682. https://doi.org/10.1016/j.neuroimage.2010.01.025

DeFelipe, J., & Fariñas, I. (1992). The pyramidal neuron of the cerebral cortex: Morphological and chemical characteristics of the synaptic inputs. *Progress in Neurobiology*, *39*(6), 563–607. https://doi.org/10.1016/0301-0082(92)90015-7

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Demyttenaere, K., Bruffaerts, R., Lee, S., Posada-Villa, J., Kovess, V., Angermeyer, M. C., Levinson, D., de Girolamo, G., Nakane, H., Mneimneh, Z., Lara, C., de Graaf, R., Scott, K. M., Gureje, O., Stein, D. J., Haro, J. M., Bromet, E. J., Kessler, R. C., Alonso, J., & Von Korff, M. (2007). Mental disorders among persons with chronic back or neck pain: Results from the world mental health surveys. *Pain*, *129*(3), 332–342. https://doi.org/10.1016/j.pain.2007.01.022

Dey, A. (2016). Machine learning algorithms: a review. *International Journal of Computer Science and Information Technologies, 7*(3), 1174–1179.

Diaz, E., & Morales, H. (2016). Spinal Cord Anatomy and Clinical Syndromes. *Seminars in Ultrasound, CT and MRI*, *37*(5), 360–371. https://doi.org/10.1053/j.sult.2016.05.002

Djouhri, L., & Lawson, S. N. (2004). Aβ-fiber nociceptive primary afferent neurons: a review of incidence and properties in relation to other afferent A-fiber neurons in mammals. *Brain Research Reviews*, *46*(2), 131–145. https://doi.org/10.1016/j.brainresrev.2004.07.015

Dong, X., Yu, Z., Cao, W., Shi, Y., & Ma, Q. (2020). A survey on ensemble learning. *Frontiers* *of Computer Science*, *14*(2), 241–258. https://doi.org/10.1007/s11704-019-8208-z

Dowman, R., Rissacher, D., & Schuckers, S. (2008). EEG indices of tonic pain-related activity in the somatosensory cortices. *Clinical Neurophysiology*, *119*(5), 1201–1212. https://doi.org/10.1016/j.clinph.2008.01.019

Drimalla, H., Landwehr, N., Hess, U., & Dziobek, I. (2019). From face to face: the contribution of facial mimicry to cognitive and emotional empathy. *Cognition and Emotion*, *33*(8), 1672–1686. https://doi.org/10.1080/02699931.2019.1596068

Dubin, A. E., & Patapoutian, A. (2010). Nociceptors: the sensors of the pain pathway. *Journal of Clinical Investigation*, *120*(11), 3760–3772. https://doi.org/10.1172/JCI42843

Duerden, E. G., & Albanese, M.-C. (2013). Localization of pain-related brain activation: A meta-analysis of neuroimaging data. *Human Brain Mapping*, *34*(1), 109–149. https://doi.org/10.1002/hbm.21416

Dunning, J. P., & Hajcak, G. (2009). See no evil: Directing visual attention within unpleasant images modulates the electrocortical response. *Psychophysiology*, *46*(1), 28–33. https://doi.org/10.1111/j.1469-8986.2008.00723.x

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: a systematic review and meta-analysis. *Journal of Clinical Epidemiology*, *88*, 92–101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Eimer, M. (2000). Effects of face inversion on the structural encoding and recognition of faces. *Cognitive Brain Research*, *10*(1–2), 145–158. https://doi.org/10.1016/S0926-6410(00)00038-0

Eldabe, S., Obara, I., Panwar, C., & Caraway, D. (2022). Biomarkers for Chronic Pain: Significance and Summary of Recent Advances. *Pain Research and Management*, *2022*, 1–6. https://doi.org/10.1155/2022/1940906

Elsayed, M., Sim, K. S., & Tan, S. C. (2020). A Novel Approach to Objectively Quantify the Subjective Perception of Pain Through Electroencephalogram Signal Analysis. *IEEE Access*, *8*, 199920–199930. https://doi.org/10.1109/ACCESS.2020.3032153

Elston. (2011). Pyramidal cells in prefrontal cortex of primates: marked differences in neuronal structure among species. *Frontiers in Neuroanatomy*. https://doi.org/10.3389/fnana.2011.00002

Eroğlu, K., Kayıkçıoğlu, T., & Osman, O. (2020). Effect of brightness of visual stimuli on EEG signals. *Behavioural Brain Research*, *382*, 112486. https://doi.org/10.1016/j.bbr.2020.112486

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, *542*(7639), 115–118. https://doi.org/10.1038/nature21056

Fabi, S., & Leuthold, H. (2016). Empathy for pain influences perceptual and motor processing: Evidence from response force, ERPs, and EEG oscillations. *Social Neuroscience*, 1–16. https://doi.org/10.1080/17470919.2016.1238009

Faisal, A. A., Selen, L. P. J., & Wolpert, D. M. (2008). Noise in the nervous system. *Nature Reviews Neuroscience*, *9*(4), 292–303. https://doi.org/10.1038/nrn2258

Fallon, N., Chiu, Y., Nurmikko, T., & Stancak, A. (2018). Altered theta oscillations in resting EEG of fibromyalgia syndrome patients. *European Journal of Pain*, *22*(1), 49–57. https://doi.org/10.1002/ejp.1076

Fallon, N., Li, X., Chiu, Y., Nurmikko, T., & Stancak, A. (2015). Altered Cortical Processing of Observed Pain in Patients With Fibromyalgia Syndrome. *The Journal of Pain*, *16*(8), 717–726. https://doi.org/10.1016/j.jpain.2015.04.008

Fallon, N., Li, X., & Stancak, A. (2015). Pain Catastrophising Affects Cortical Responses to

Viewing Pain in Others. *PLOS ONE*, *10*(7), e0133504. https://doi.org/10.1371/journal.pone.0133504

Fallon, N., Roberts, C., & Stancak, A. (2020). Shared and distinct functional networks for empathy and pain processing: a systematic review and meta-analysis of fMRI studies. *Social Cognitive and Affective Neuroscience*, *15*(7), 709–723. https://doi.org/10.1093/scan/nsaa090

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *CJEM*, *8*(01), 19–20. https://doi.org/10.1017/S1481803500013336

Fan, Y.-T., Chen, C., Chen, S.-C., Decety, J., & Cheng, Y. (2014). Empathic arousal and social understanding in individuals with autism: evidence from fMRI and ERP measurements. *Social Cognitive and Affective Neuroscience*, *9*(8), 1203–1213. https://doi.org/10.1093/scan/nst101

Fan, Y., Duncan, N. W., de Greck, M., & Northoff, G. (2011). Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. *Neuroscience & Biobehavioral Reviews*, *35*(3), 903–911. https://doi.org/10.1016/j.neubiorev.2010.10.009

Fan, Y., & Han, S. (2008). Temporal dynamic of neural mechanisms involved in empathy for pain: An event-related brain potential study. *Neuropsychologia*, *46*(1), 160–173. https://doi.org/10.1016/j.neuropsychologia.2007.07.023

Faraggi, D., & Reiser, B. (2002). Estimation of the area under the ROC curve. *Statistics in Medicine*, *21*(20), 3093–3106. https://doi.org/10.1002/sim.1228

Fatourechi, M., Bashashati, A., Ward, R. K., & Birch, G. E. (2007). EMG and EOG artifacts in brain computer interface systems: A survey. *Clinical Neurophysiology*, *118*(3), 480–494. https://doi.org/10.1016/j.clinph.2006.10.019

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–

874. https://doi.org/10.1016/j.patrec.2005.10.010

Fayaz, A., Croft, P., Langford, R. M., Donaldson, L. J., & Jones, G. T. (2016). Prevalence of chronic pain in the UK: a systematic review and meta-analysis of population studies. *BMJ Open*, *6*(6), e010364. https://doi.org/10.1136/bmjopen-2015-010364

Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *The Journal of Machine Learning Research*, *15*(1), 3133–3181.

Fernandez Rojas, R., Hirachan, N., Brown, N., Waddington, G., Murtagh, L., Seymour, B., & Goecke, R. (2023). Multimodal physiological sensing for the assessment of acute pain. *Frontiers in Pain Research*, *4*. https://doi.org/10.3389/fpain.2023.1150264

Ferracuti, S., Seri, S., Mattia, D., & Cruccu, G. (1994). Quantitative EEG modifications during the cold water pressor test: hemispheric and hand differences. *International Journal of Psychophysiology*, *17*(3), 261–268. https://doi.org/10.1016/0167-8760(94)90068-X

Fillingim, R. B. (2005). Individual differences in pain responses. *Current Rheumatology Reports*, *7*(5), 342–347. https://doi.org/10.1007/s11926-005-0018-7

Fillingim, R. B. (2017). Individual differences in pain. *PAIN*, *158*, S11–S18. https://doi.org/10.1097/j.pain.0000000000000775

Fillingim, R. B., Loeser, J. D., Baron, R., & Edwards, R. R. (2016). Assessment of Chronic Pain: Domains, Methods, and Mechanisms. *The Journal of Pain*, *17*(9), T10–T20. https://doi.org/10.1016/j.jpain.2015.08.010

Fine, P. G. (2011). Long-Term Consequences of Chronic Pain: Mounting Evidence for Pain as a Neurological Disease and Parallels with Other Chronic Disease States. *Pain Medicine*, *12*(7), 996–1004. https://doi.org/10.1111/j.1526-4637.2011.01187.x

Fishbain, D. A., Cutler, R., Rosomoff, H. L., & Rosomoff, R. S. (1997). Chronic Pain-Associated

Depression: Antecedent or Consequence of Chronic Pain? A Review. *The Clinical Journal of Pain*, *13*(2), 116–137. https://doi.org/10.1097/00002508-199706000-00006

Fitzcharles, M.-A., Cohen, S. P., Clauw, D. J., Littlejohn, G., Usui, C., & Häuser, W. (2021). Nociplastic pain: towards an understanding of prevalent pain conditions. *The Lancet*, *397*(10289), 2098–2110. https://doi.org/10.1016/S0140-6736(21)00392-5

Fontana, M. A., Lyman, S., Sarker, G. K., Padgett, D. E., & MacLean, C. H. (2019). Can Machine Learning Algorithms Predict Which Patients Will Achieve Minimally Clinically Important Differences From Total Joint Arthroplasty? *Clinical Orthopaedics & Related Research*, *477*(6), 1267–1279. https://doi.org/10.1097/CORR.0000000000000687

Frid, A., Shor, M., Shifrin, A., Yarnitsky, D., & Granovsky, Y. (2020). A Biomarker for Discriminating Between Migraine With and Without Aura: Machine Learning on Functional Connectivity on Resting-State EEGs. *Annals of Biomedical Engineering*, *48*(1), 403–412. https://doi.org/10.1007/s10439-019-02357-3

Fuchs-Lacelle, S., & Hadjistavropoulos, T. (2004). Development and preliminary validation of the pain assessment checklist for seniors with limited ability to communicate (PACSLAC). *Pain Management Nursing*, *5*(1), 37–49. https://doi.org/10.1016/j.pmn.2003.10.001

Furman, A. J., Meeker, T. J., Rietschel, J. C., Yoo, S., Muthulingam, J., Prokhorenko, M., Keaser, M. L., Goodman, R. N., Mazaheri, A., & Seminowicz, D. A. (2018). Cerebral peak alpha frequency predicts individual differences in pain sensitivity. *NeuroImage*, *167*, 203–210. https://doi.org/10.1016/j.neuroimage.2017.11.042

Furman, A. J., Prokhorenko, M., Keaser, M. L., Zhang, J., Chen, S., Mazaheri, A., & Seminowicz, D. A. (2020). Sensorimotor Peak Alpha Frequency Is a Reliable Biomarker of Prolonged Pain Sensitivity. *Cerebral Cortex*, *30*(12), 6069–6082.

https://doi.org/10.1093/cercor/bhaa124

Furman, A. J., Thapa, T., Summers, S. J., Cavaleri, R., Fogarty, J. S., Steiner, G. Z., Schabrun, S. M., & Seminowicz, D. A. (2019). Cerebral peak alpha frequency reflects average pain severity in a human model of sustained, musculoskeletal pain. *Journal of Neurophysiology*, *122*(4), 1784–1793. https://doi.org/10.1152/jn.00279.2019

Fushiki, T. (2011). Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, *21*(2), 137–146. https://doi.org/10.1007/s11222-009-9153-8

Gabard-Durnam, L. J., Mendez Leal, A. S., Wilkinson, C. L., & Levin, A. R. (2018). The Harvard Automated Processing Pipeline for Electroencephalography (HAPPE): Standardized Processing Software for Developmental and High-Artifact Data. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00097

Galang, C. M., Jenkins, M., & Obhi, S. S. (2020). Exploring the effects of visual perspective on the ERP components of empathy for pain. *Social Neuroscience*, *15*(2), 186–198. https://doi.org/10.1080/17470919.2019.1674686

Galvez-Sánchez, C. M., Muñoz Ladrón de Guevara, C., Montoro, C. I., Fernández-Serrano, M. J., Duschek, S., & Reyes del Paso, G. A. (2018). Cognitive deficits in fibromyalgia syndrome are associated with pain responses to low intensity pressure stimulation. *PLOS ONE*, *13*(8), e0201488. https://doi.org/10.1371/journal.pone.0201488

Garcia-Larrea, L., & Peyron, R. (2013). Pain matrices and neuropathic pain matrices: A review. *Pain*, *154*, S29–S43. https://doi.org/10.1016/j.pain.2013.09.001

Garland, E. L. (2012). Pain Processing in the Human Nervous System. A Selective Review of Nociceptive and Biobehavioral Pathways. In *Primary Care - Clinics in Office Practice* (Vol. 39, Issue 3, pp. 561–571). https://doi.org/10.1016/j.pop.2012.06.013

Gaskin, D. J., & Richard, P. (2012). The Economic Costs of Pain in the United States. *The*

Journal of Pain, *13*(8), 715–724. https://doi.org/10.1016/j.jpain.2012.03.009

Gatchel, R. J., Peng, Y. B., Peters, M. L., Fuchs, P. N., & Turk, D. C. (2007). The biopsychosocial approach to chronic pain: Scientific advances and future directions. *Psychological Bulletin*, *133*(4), 581–624. https://doi.org/10.1037/0033-2909.133.4.581

Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent* (2nd ed.). O'Reilly.

Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural Computation*, *12*(10), 2451–2471. https://doi.org/10.1162/089976600300015015

Ghamari, M. (2018). A review on wearable photoplethysmography sensors and their potential future applications in health care. *International Journal of Biosensors & Bioelectronics*, *4*(4). https://doi.org/10.15406/ijbsbe.2018.04.00125

Ghosh, L., Dewan, D., Chowdhury, A., & Konar, A. (2021). Exploration of face-perceptual ability by EEG induced deep learning algorithm. *Biomedical Signal Processing and Control*, *66*, 102368. https://doi.org/10.1016/j.bspc.2020.102368

Giehl, J., Meyer-Brandis, G., Kunz, M., & Lautenbacher, S. (2014). Responses to tonic heat pain in the ongoing EEG under conditions of controlled attention. *Somatosensory & Motor Research*, *31*(1), 40–48. https://doi.org/10.3109/08990220.2013.837045

Ginsburg, G., & McCarthy, J. J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*, *19*(12), 491–496. https://doi.org/10.1016/S0167-7799(01)01814-5

Glover, G. H. (2011). Overview of Functional Magnetic Resonance Imaging. *Neurosurgery Clinics of North America*, *22*(2), 133–139. https://doi.org/10.1016/j.nec.2010.11.001

Goncharova, I. ., McFarland, D. ., Vaughan, T. ., & Wolpaw, J. . (2003). EMG contamination of

EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, *114*(9), 1580–1593. https://doi.org/10.1016/S1388-2457(03)00093-2

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Gopalakrishnan, S., & Ganeshkumar, P. (2013). Systematic Reviews and Meta-analysis: Understanding the Best Evidence in Primary Healthcare. *Journal of Family Medicine and Primary Care*, *2*(1), 9–14. https://doi.org/10.4103/2249-4863.109934

Górecka, J., & Makiewicz, P. (2019). The Dependence of Electrode Impedance on the Number of Performed EEG Examinations. *Sensors*, *19*(11), 2608. https://doi.org/10.3390/s19112608

Goubert, L., Craig, K. D., Vervoort, T., Morley, S., Sullivan, M. J. L., Williams, de C. A. C., Cano, A., & Crombez, G. (2005). Facing others in pain: the effects of empathy. *Pain*, *118*(3), 285–288. https://doi.org/10.1016/j.pain.2005.10.025

Govindan, R. B., Massaro, A., Vezina, G., Tsuchida, T., Cristante, C., & du Plessis, A. (2017). Does relative or absolute EEG power have prognostic value in HIE setting? *Clinical Neurophysiology*, *128*(1), 14–15. https://doi.org/10.1016/j.clinph.2016.10.094

Gram, M., Erlenwein, J., Petzke, F., Falla, D., Przemeck, M., Emons, M. I., Reuster, M., Olesen, S. S., & Drewes, A. M. (2017). Prediction of postoperative opioid analgesia using clinical-experimental parameters and electroencephalography. *European Journal of Pain (United Kingdom)*, *21*(2), 264–277. https://doi.org/10.1002/ejp.921

Gram, M., Graversen, C., Olesen, S. S., & Drewes, A. M. (2015). Dynamic spectral indices of the electroencephalogram provide new insights into tonic pain. *Clinical Neurophysiology*, *126*(4), 763–771. https://doi.org/10.1016/j.clinph.2014.07.027

Granovsky, Y., Matre, D., Sokolik, A., Lorenz, J., & Casey, K. L. (2005). Thermoreceptive innervation of human glabrous and hairy skin: a contact heat evoked potential analysis. *Pain*, *115*(3), 238–247. https://doi.org/10.1016/j.pain.2005.02.017

Graversen, C., Brock, C., Drewes, A. M., & Farina, D. (2011). Biomarkers for visceral hypersensitivity identified by classification of electroencephalographic frequency alterations. *Journal of Neural Engineering*, *8*(5), 056014. https://doi.org/10.1088/1741-2560/8/5/056014

Graversen, C., Malver, L. P., Kurita, G. P., Staahl, C., Christrup, L. L., Sjøgren, P., & Drewes, A. M. (2015). Altered Frequency Distribution in the Electroencephalogram is Correlated to the Analgesic Effect of Remifentanil. *Basic & Clinical Pharmacology & Toxicology*, *116*(5), 414–422. https://doi.org/10.1111/bcpt.12330

Graversen, C., Olesen, S. S., Olesen, A. E., Steimle, K., Farina, D., Wilder-Smith, O. H. G., Bouwense, S. A. W., van Goor, H., & Drewes, A. M. (2012). The analgesic effect of pregabalin in patients with chronic pain is reflected by changes in pharmaco-EEG spectral indices. *British Journal of Clinical Pharmacology*, *73*(3), 363–372. https://doi.org/10.1111/j.1365-2125.2011.04104.x

Grech, R., Cassar, T., Muscat, J., Camilleri, K. P., Fabri, S. G., Zervakis, M., Xanthopoulos, P., Sakkalis, V., & Vanrumste, B. (2008). Review on solving the inverse problem in EEG source analysis. *Journal of NeuroEngineering and Rehabilitation*, *5*(1), 25. https://doi.org/10.1186/1743-0003-5-25

Grosen, K., Olesen, A. E., Gram, M., Jonsson, T., Kamp-Jensen, M., Andresen, T., Nielsen, C., Pozlep, G., Pfeiffer-Jensen, M., Morlion, B., & Drewes, A. M. (2017). Predictors of opioid efficacy in patients with chronic pain: A prospective multicenter observational cohort study. *PLoS ONE*, *12*(2). https://doi.org/10.1371/journal.pone.0171723

Gross, J., Schnitzler, A., Timmermann, L., & Ploner, M. (2007). Gamma Oscillations in Human Primary Somatosensory Cortex Reflect Pain Perception. *PLoS Biology*, *5*(5), e133. https://doi.org/10.1371/journal.pbio.0050133

Grosse-Wentrup, M., Liefhold, C., Gramann, K., & Buss, M. (2009). Beamforming in Noninvasive Brain–Computer Interfaces. *IEEE Transactions on Biomedical Engineering*, *56*(4), 1209–1219. https://doi.org/10.1109/TBME.2008.2009768

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, *3*, 1157–1182.

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia Medica*, 297–307. https://doi.org/10.11613/BM.2016.034

Hadi, M. A., McHugh, G. A., & Closs, S. J. (2019). Impact of Chronic Pain on Patients' Quality of Life: A Comparative Mixed-Methods Study. *Journal of Patient Experience*, *6*(2), 133–141. https://doi.org/10.1177/2374373518786013

Hadjileontiadis, L. J. (2015). EEG-Based Tonic Cold Pain Characterization Using Wavelet Higher Order Spectral Features. *IEEE Transactions on Biomedical Engineering*, *62*(8), 1981–1991. https://doi.org/10.1109/TBME.2015.2409133

Hadjistavropoulos, T., Baeyer, C. v., & Craig, K. D. (2001). Pain assessment in persons with limited ability to communicate. In D. C. Turk & R. Melzack (Eds.), *Handbook of pain assessment* (pp. 134–149). The Guilford Press.

Haefeli, M., & Elfering, A. (2006). Pain assessment. *European Spine Journal*, *15*(S1), S17–S24. https://doi.org/10.1007/s00586-005-1044-x

Hajcak, G., MacNamara, A., Foti, D., Ferri, J., & Keil, A. (2013). The dynamic allocation of attention to emotion: Simultaneous and independent evidence from the late positive potential and steady state visual evoked potentials. *Biological Psychology*, *92*(3), 447–455. https://doi.org/10.1016/j.biopsycho.2011.11.012

Hajian-Tilaki, K. (2013). Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. *Caspian Journal of Internal Medicine*, *4*(2), 627–635. http://www.ncbi.nlm.nih.gov/pubmed/24009950

Han, S., Fan, Y., & Mao, L. (2008). Gender difference in empathy for pain: An electrophysiological investigation. *Brain Research*, *1196*, 85–93. https://doi.org/10.1016/j.brainres.2007.12.062

Handwerker, H. O., Kilo, S., & Reeh, P. W. (1991). Unresponsive afferent nerve fibres in the sural nerve of the rat. *The Journal of Physiology*, *435*(1), 229–242. https://doi.org/10.1113/jphysiol.1991.sp018507

Harmonya, T., Fernández, T., Rodríguez, M., Reyes, A., Marosi, E., & Bernal, J. (1993). Test-Retest Reliability of EEG Spectral Parameters During Cognitive Tasks: II Coherence. *International Journal of Neuroscience*, *68*(3–4), 263–271. https://doi.org/10.3109/00207459308994281

Hauck, M., Domnick, C., Lorenz, J., Gerloff, C., & Engel, A. K. (2015). Top-down and bottom-up modulation of pain-induced oscillations. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00375

Hein, G., Silani, G., Preuschoff, K., Batson, C. D., & Singer, T. (2010). Neural Responses to Ingroup and Outgroup Members' Suffering Predict Individual Differences in Costly Helping. *Neuron*, *68*(1), 149–160. https://doi.org/10.1016/j.neuron.2010.09.003

Henmar, S., Simonsen, E. B., & Berg, R. W. (2020). What are the gray and white matter volumes of the human spinal cord? *Journal of Neurophysiology*, *124*(6), 1792–1797. https://doi.org/10.1152/jn.00413.2020

Herr, K., Coyne, P. J., McCaffery, M., Manworren, R., & Merkel, S. (2011). Pain Assessment in the Patient Unable to Self-Report: Position Statement with Clinical Practice Recommendations. *Pain Management Nursing*, *12*(4), 230–250. https://doi.org/10.1016/j.pmn.2011.10.002

Herrero, J. F., Laird, J. M. ., & Lopez-Garcia, J. A. (2000). Wind-up of spinal cord neurones and pain sensation: much ado about something? *Progress in Neurobiology*, *61*(2), 169–203. https://doi.org/10.1016/S0301-0082(99)00051-9

Heus, P., Damen, J. A. A. G., Pajouheshnia, R., Scholten, R. J. P. M., Reitsma, J. B., Collins, G. S., Altman, D. G., Moons, K. G. M., & Hooft, L. (2018). Poor reporting of multivariable prediction model studies: towards a targeted implementation strategy of the TRIPOD statement. *BMC Medicine*, *16*(1), 120. https://doi.org/10.1186/s12916-018-1099-2

Hill, N. J., Gupta, D., Brunner, P., Gunduz, A., Adamo, M. A., Ritaccio, A., & Schalk, G. (2012). Recording Human Electrocorticographic (ECoG) Signals for Neuroscientific Research and Real-time Functional Cortical Mapping. *Journal of Visualized Experiments*, *64*. https://doi.org/10.3791/3993

Hill, R. Z., & Bautista, D. M. (2020). Getting in Touch with Mechanical Pain Mechanisms. *Trends in Neurosciences*, *43*(5), 311–325. https://doi.org/10.1016/j.tins.2020.03.004

Hinrichs, H., Scholz, M., Baum, A. K., Kam, J. W. Y., Knight, R. T., & Heinze, H.-J. (2020). Comparison between a wireless dry electrode EEG system with a conventional wired wet electrode EEG system for clinical applications. *Scientific Reports*, *10*(1), 5218. https://doi.org/10.1038/s41598-020-62154-0

Ho, S. Y., Phua, K., Wong, L., & Bin Goh, W. W. (2020). Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. *Patterns*, *1*(8), 100129. https://doi.org/10.1016/j.patter.2020.100129

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, *117*(4), 500–544. https://doi.org/10.1113/jphysiol.1952.sp004764

Hoffman, K. M., Trawalter, S., Axt, J. R., & Oliver, M. N. (2016). Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences*, *113*(16), 4296–4301. https://doi.org/10.1073/pnas.1516047113

Holder, L. B., Haque, M. M., & Skinner, M. K. (2017). Machine learning for epigenetics and future medical applications. *Epigenetics*, *12*(7), 505–514. https://doi.org/10.1080/15592294.2017.1329068

Hoo, Z. H., Candlish, J., & Teare, D. (2017). What is an ROC curve? *Emergency Medicine Journal*, *34*(6), 357–359. https://doi.org/10.1136/emermed-2017-206735

Hossin, M., & Sulaiman, M. . (2015). A Review on Evaluation Metrics for Data Classification Evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 01–11. https://doi.org/10.5121/ijdkp.2015.5201

Hsiao, F.-J., Chen, W.-T., Pan, L.-L. H., Liu, H.-Y., Wang, Y.-F., Chen, S.-P., Lai, K.-L., Coppola, G., & Wang, S.-J. (2022). Resting-state magnetoencephalographic oscillatory connectivity to identify patients with chronic migraine using machine learning. *The Journal of Headache and Pain*, *23*(1), 130. https://doi.org/10.1186/s10194-022-01500-1

Hsu, H.-H., Hsieh, C.-W., & Lu, M.-D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications*, *38*(7), 8144–8150. https://doi.org/10.1016/j.eswa.2010.12.156

Hu, L., Peng, W., Valentini, E., Zhang, Z., & Hu, Y. (2013). Functional Features of Nociceptive-Induced Suppression of Alpha Band Electroencephalographic Oscillations. *The Journal of Pain*, *14*(1), 89–99. https://doi.org/10.1016/j.jpain.2012.10.008

Hu, X.-S., Nascimento, T. D., Bender, M. C., Hall, T., Petty, S., O'Malley, S., Ellwood, R. P., Kaciroti, N., Maslowski, E., & DaSilva, A. F. (2019). Feasibility of a Real-Time Clinical Augmented Reality and Artificial Intelligence Framework for Pain Detection and Localization From the Brain. *Journal of Medical Internet Research*, *21*(6), e13594. https://doi.org/10.2196/13594

Huang, G., Xiao, P., Hung, Y. S., Iannetti, G. D., Zhang, Z. G., & Hu, L. (2013). A novel approach to predict subjective pain perception from single-trial laser-evoked potentials. *NeuroImage*, *81*, 283–293. https://doi.org/10.1016/J.NEUROIMAGE.2013.05.017

Huang, Y., Li, W., Macheret, F., Gabriel, R. A., & Ohno-Machado, L. (2020). A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association*, *27*(4), 621–633. https://doi.org/10.1093/jamia/ocz228

Huber, M. T., Bartling, J., Pachur, D., Woikowsky-Biedau, S. v., & Lautenbacher, S. (2006). EEG responses to tonic heat pain. *Experimental Brain Research*, *173*(1), 14–24. https://doi.org/10.1007/s00221-006-0366-1

Huishi Zhang, C., Sohrabpour, A., Lu, Y., & He, B. (2016). Spectral and spatial changes of brain rhythmic activity in response to the sustained thermal pain stimulation. *Human*

Brain Mapping, *37*(8), 2976–2991. https://doi.org/10.1002/hbm.23220

Hunter, A. M., Leuchter, A. F., Cook, I. A., Abrams, M., Siegman, B. E., Furst, D. E., & Chappell, A. S. (2009). Brain Functional Changes and Duloxetine Treatment Response in Fibromyalgia: A Pilot Study. *Pain Medicine*, *10*(4), 730–738. https://doi.org/10.1111/j.1526-4637.2009.00614.x

Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725–726. https://doi.org/10.1126/science.359.6377.725

Hylands-White, N., Duarte, R. V., & Raphael, J. H. (2017). An overview of treatment approaches for chronic pain management. *Rheumatology International*, *37*(1), 29–42. https://doi.org/10.1007/s00296-016-3481-8

Iannetti, G. D., Hughes, N. P., Lee, M. C., & Mouraux, A. (2008). Determinants of Laser-Evoked EEG Responses: Pain Perception or Stimulus Saliency? *Journal of Neurophysiology*, *100*(2), 815–828. https://doi.org/10.1152/jn.00097.2008

Iannetti, G. D., & Mouraux, A. (2010). From the neuromatrix to the pain matrix (and back). *Experimental Brain Research*, *205*(1), 1–12. https://doi.org/10.1007/s00221-010-2340-1

IASP. (2017). *IASP Terminology*.

Ibáñez, A., Hurtado, E., Lobos, A., Escobar, J., Trujillo, N., Baez, S., Huepe, D., Manes, F., & Decety, J. (2011). Subliminal presentation of other faces (but not own face) primes behavioral and evoked cortical processing of empathy for pain. *Brain Research*, *1398*, 72–85. https://doi.org/10.1016/j.brainres.2011.05.014

Ille, N., Berg, P., & Scherg, M. (2002). Artifact Correction of the Ongoing EEG Using Spatial Filters Based on Artifact and Brain Signal Topographies. *Journal of Clinical Neurophysiology*, *19*(2), 113–124. https://doi.org/10.1097/00004691-200203000-

00002

İnce, R., Adanır, S. S., & Sevmez, F. (2021). The inventor of electroencephalography (EEG): Hans Berger (1873–1941). *Child's Nervous System*, *37*(9), 2723–2724. https://doi.org/10.1007/s00381-020-04564-z

Itier, R. J. (2004). N170 or N1? Spatiotemporal Differences between Object and Face Processing Using ERPs. *Cerebral Cortex*, *14*(2), 132–142. https://doi.org/10.1093/cercor/bhg111

Itier, R. J., & Taylor, M. J. (2002). Inversion and Contrast Polarity Reversal Affect both Encoding and Recognition Processes of Unfamiliar Faces: A Repetition Study Using ERPs. *NeuroImage*, *15*(2), 353–372. https://doi.org/10.1006/nimg.2001.0982

Itier, R. J., & Taylor, M. J. (2004). Source analysis of the N170 to faces and objects. *NeuroReport*, *15*(8), 1261–1265. https://doi.org/10.1097/01.wnr.0000127827.73576.d8

Jackson, A. F., & Bolger, D. J. (2014). The neurophysiological bases of EEG and EEG measurement: A review for the rest of us. *Psychophysiology*, *51*(11), 1061–1071. https://doi.org/10.1111/psyp.12283

Jannot, A.-S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology*, *66*(3), 296–301. https://doi.org/10.1016/j.jclinepi.2012.09.015

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalography and Clinical Neurophysiology*, *10*, 371–375.

Jauniaux, J., Khatibi, A., Rainville, P., & Jackson, P. L. (2019). A meta-analysis of neuroimaging studies on pain empathy: investigating the role of visual information and observers' perspective. *Social Cognitive and Affective Neuroscience*, *14*(8), 789–813.

https://doi.org/10.1093/scan/nsz055

Jaworska, N., de la Salle, S., Ibrahim, M.-H., Blier, P., & Knott, V. (2019). Leveraging Machine Learning Approaches for Predicting Antidepressant Treatment Response Using Electroencephalography (EEG) and Clinical Data. *Frontiers in Psychiatry*, *9*. https://doi.org/10.3389/fpsyt.2018.00768

Jeffreys, D. A. (1989). A face-responsive potential recorded from the human scalp. *Experimental Brain Research*, *78*(1). https://doi.org/10.1007/BF00230699

Jeffreys, D. A. (1996). Evoked Potential Studies of Face and Object Processing. *Visual Cognition*, *3*(1), 1–38. https://doi.org/10.1080/713756729

Jensen, K. B., Regenbogen, C., Ohse, M. C., Frasnelli, J., Freiherr, J., & Lundström, J. N. (2016). Brain activations during pain. *Pain*, *157*(6), 1279–1286. https://doi.org/10.1097/j.pain.0000000000000517

Jiang, M., Mieronkoski, R., Syrjälä, E., Anzanpour, A., Terävä, V., Rahmani, A. M., Salanterä, S., Aantaa, R., Hagelberg, N., & Liljeberg, P. (2019). Acute pain intensity monitoring with the classification of multiple physiological parameters. *Journal of Clinical Monitoring and Computing*, *33*(3), 493–507. https://doi.org/10.1007/s10877-018-0174-8

Jiang, T., Gradus, J. L., & Rosellini, A. J. (2020). Supervised Machine Learning: A Brief Primer. *Behavior Therapy*, *51*(5), 675–687. https://doi.org/10.1016/j.beth.2020.05.002

Johnson, H. R., Trinidad, D. D., Guzman, S., Khan, Z., Parziale, J. V., DeBruyn, J. M., & Lents, N. H. (2016). A Machine Learning Approach for Using the Postmortem Skin Microbiome to Estimate the Postmortem Interval. *PLOS ONE*, *11*(12), e0167370. https://doi.org/10.1371/journal.pone.0167370

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance.

Journal of Big Data, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Johnston, P., Molyneux, R., & Young, A. W. (2015). The N170 observed 'in the wild': robust event-related potentials to faces in cluttered dynamic visual scenes. *Social Cognitive and Affective Neuroscience*, *10*(7), 938–944. https://doi.org/10.1093/scan/nsu136

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415

Joyce, C., & Rossion, B. (2005). The face-sensitive N170 and VPP components manifest the same brain processes: The effect of reference electrode site. *Clinical Neurophysiology*, *116*(11), 2613–2631. https://doi.org/10.1016/j.clinph.2005.07.005

Julius, D., & Basbaum, A. I. (2001). Molecular mechanisms of nociception. *Nature*, *413*(6852), 203–210. https://doi.org/10.1038/35093019

Kam, J. W. Y., Xu, J., & Handy, T. C. (2014). I don't feel your pain (as much): The desensitizing effect of mind wandering on the perception of others' discomfort. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(1), 286–296. https://doi.org/10.3758/s13415-013-0197-z

Kamarudin, A. N., Cox, T., & Kolamunnage-Dona, R. (2017). Time-dependent ROC curve analysis in medical research: current methods and applications. *BMC Medical Research Methodology*, *17*(1), 53. https://doi.org/10.1186/s12874-017-0332-6

Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., & Hudspeth, A. J. (2012). *Principles of Neural Science* (5th ed.). McGraw-Hill Publishing.

Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M., & Suppes, P. (2015). A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *PLOS ONE*, *10*(8), e0135697. https://doi.org/10.1371/journal.pone.0135697

Kaplan, A. Y., Fingelkurts, A. A., Fingelkurts, A. A., Borisov, S. V., & Darkhovsky, B. S. (2005). Nonstationary nature of the brain activity as revealed by EEG/MEG: Methodological, practical and conceptual challenges. *Signal Processing*, *85*(11), 2190–2212. https://doi.org/10.1016/j.sigpro.2005.07.010

Kappesser, J., Williams, A. C. de C., & Prkachin, K. M. (2006). Testing two accounts of pain underestimation. *Pain*, *124*(1), 109–116. https://doi.org/10.1016/j.pain.2006.04.003

Katz, J. S., & Martin, B. R. (1997). What is research collaboration? *Research Policy*, *26*(1), 1–18. https://doi.org/10.1016/S0048-7333(96)00917-1

Kaur, M., Prakash, N. R., Kalra, P., & Puri, G. D. (2019). Electroencephalogram-Based Pain Classification Using Artificial Neural Networks. *IETE Journal of Research*, 1–14. https://doi.org/10.1080/03772063.2019.1702903

Keil, A., Bernat, E. M., Cohen, M. X., Ding, M., Fabiani, M., Gratton, G., Kappenman, E. S., Maris, E., Mathewson, K. E., Ward, R. T., & Weisz, N. (2022). Recommendations and publication guidelines for studies using frequency domain and time-frequency domain analyses of neural time series. *Psychophysiology*, *59*(5). https://doi.org/10.1111/psyp.14052

Kelley, A. S., Siegler, E. L., & Reid, M. C. (2008). Pitfalls and Recommendations Regarding the Management of Acute Pain Among Hospitalized Patients with Dementia. *Pain Medicine*, *9*(5), 581–586. https://doi.org/10.1111/j.1526-4637.2008.00472.x

Khalid, S., Khalil, T., & Nasreen, S. (2014). A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning. *Science and Information Conference*, 1–8.

Kim, J. A., & Davis, K. D. (2021). Neural Oscillations: Understanding a Neural Code of Pain. *The Neuroscientist*, *27*(5), 544–570. https://doi.org/10.1177/1073858420958629

Kim, S. E., Behr, M. K., Ba, D., & Brown, E. N. (2018). State-space multitaper time-frequency analysis. *Proceedings of the National Academy of Sciences*, *115*(1). https://doi.org/10.1073/pnas.1702877115

Kimura, A., Mitsukura, Y., Oya, A., Matsumoto, M., Nakamura, M., Kanaji, A., & Miyamoto, T. (2021). Objective characterization of hip pain levels during walking by combining quantitative electroencephalography with machine learning. *Scientific Reports*, *11*(1), 3192. https://doi.org/10.1038/s41598-021-82696-1

King, N. B., & Fraser, V. (2013). Untreated Pain, Narcotics Regulation, and Global Health Ideologies. *PLoS Medicine*, *10*(4), e1001411. https://doi.org/10.1371/journal.pmed.1001411

Kirschfeld, K. (2005). The physical basis of alpha waves in the electroencephalogram and the origin of the ?Berger effect? *Biological Cybernetics*, *92*(3), 177–185. https://doi.org/10.1007/s00422-005-0547-1

Kirschstein, T., & Köhling, R. (2009). What is the Source of the EEG? *Clinical EEG and Neuroscience*, *40*(3), 146–149. https://doi.org/10.1177/155005940904000305

Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C. (1999). The ten-twenty electrode system of the International Federation. The International Federation of Clinical Neurophysiology. *Electroencephalography and Clinical Neurophysiology. Supplement*, *52*, 3–6. http://www.ncbi.nlm.nih.gov/pubmed/10590970

Kokol, P., Kokol, M., & Zagoranski, S. (2022). Machine learning on small size samples: A synthetic knowledge synthesis. *Science Progress*, *105*(1), 003685042110297. https://doi.org/10.1177/00368504211029777

Kong, Y., Posada-Quintero, H. F., & Chon, K. H. (2021). Real-Time High-Level Acute Pain Detection Using a Smartphone and a Wrist-Worn Electrodermal Activity Sensor. *Sensors*, *21*(12), 3956. https://doi.org/10.3390/s21123956

Kosek, E., Cohen, M., Baron, R., Gebhart, G. F., Mico, J.-A., Rice, A. S. C., Rief, W., & Sluka, A. K. (2016). Do we need a third mechanistic descriptor for chronic pain states? *Pain*, *157*(7), 1382–1386. https://doi.org/10.1097/j.pain.0000000000000507

Kosek, E., Ekholm, J., & Hansson, P. (1995). Increased pressure pain sensibility in fibromyalgia patients is located deep to the skin but not restricted to muscle tissue. *Pain*, *63*(3), 335–339. https://doi.org/10.1016/0304-3959(95)00061-5

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*, 249–268. http://www.informatica.si/index.php/informatica/article/viewFile/148/140

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01117

Kress, M., Koltzenburg, M., Reeh, P. W., & Handwerker, H. O. (1992). Responsiveness and functional attributes of electrically localized terminals of cutaneous C-fibers in vivo and in vitro. *Journal of Neurophysiology*, *68*(2), 581–595. https://doi.org/10.1152/jn.1992.68.2.581

Kropotov, J. D. (2009). *Quantitative EEG, Event-Related Potentials and Neurotherapy* (1st ed.). Elsevier. https://doi.org/10.1016/B978-0-12-374512-5.X0001-1

Kunz, M., Mylius, V., Scharmann, S., Schepelman, K., & Lautenbacher, S. (2009). Influence of dementia on multiple components of pain. *European Journal of Pain*, *13*(3), 317–325. https://doi.org/10.1016/j.ejpain.2008.05.001

Lamm, C., Decety, J., & Singer, T. (2011). Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. *NeuroImage*, *54*(3), 2492–2502. https://doi.org/10.1016/j.neuroimage.2010.10.014

Larochette, A.-C., Chambers, C. T., & Craig, K. D. (2006). Genuine, suppressed and faked facial expressions of pain in children. *Pain*, *126*(1), 64–71. https://doi.org/10.1016/j.pain.2006.06.013

Larrañaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., Lozano, J. A., Armañanzas, R., Santafé, G., Pérez, A., & Robles, V. (2006). Machine learning in bioinformatics. *Briefings in Bioinformatics*, *7*(1), 86–112. https://doi.org/10.1093/bib/bbk007

Lau-Zhu, A., Lau, M. P. H., & McLoughlin, G. (2019). Mobile EEG in research on neurodevelopmental disorders: Opportunities and challenges. *Developmental Cognitive Neuroscience*, *36*, 100635. https://doi.org/10.1016/j.dcn.2019.100635

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Ledowski, T., Bromilow, J., Paech, M. J., Storm, H., Hacking, R., & Schug, S. A. (2006). Monitoring of skin conductance to assess postoperative pain intensity. *British Journal of Anaesthesia*, *97*(6), 862–865. https://doi.org/10.1093/bja/ael280

Ledwidge, P., Foust, J., & Ramsey, A. (2018). Recommendations for Developing an EEG Laboratory at a Primarily Undergraduate Institution. *Journal of Undergraduate Neuroscience Education : JUNE : A Publication of FUN, Faculty for Undergraduate Neuroscience*, *17*(1), A10–A19. https://doi.org/30618494

Lee, J. H., Shin, J., & Realff, M. J. (2018). Machine learning: Overview of the recent progresses and implications for the process systems engineering field. *Computers & Chemical Engineering*, *114*, 111–121. https://doi.org/10.1016/j.compchemeng.2017.10.008

Lee, S., & Bozeman, B. (2005). The Impact of Research Collaboration on Scientific Productivity. *Social Studies of Science*, *35*(5), 673–702.

https://doi.org/10.1177/0306312705052359

Legrain, V., Bruyer, R., Guérit, J.-M., & Plaghki, L. (2005). Involuntary orientation of attention to unattended deviant nociceptive stimuli is modulated by concomitant visual task difficulty. Evidence from laser evoked potentials. *Clinical Neurophysiology*, *116*(9), 2165–2174. https://doi.org/10.1016/j.clinph.2005.05.019

Legrain, V., Damme, S. Van, Eccleston, C., Davis, K. D., Seminowicz, D. A., & Crombez, G. (2009). A neurocognitive model of attention to pain: Behavioral and neuroimaging evidence. *Pain*, *144*(3), 230–232. https://doi.org/10.1016/j.pain.2009.03.020

Lehmann, D. (1987). Principles of spatial analysis. In A. S. Gevins & A. Remond (Eds.), *Handbook of electroencephalography and clinical neurophysiology: Methods of analysis of brain electrical and magnetic signals* (pp. 309–354). Elsevier.

Lei, X., & Liao, K. (2017). Understanding the Influences of EEG Reference: A Large-Scale Brain Network Perspective. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00205

Leroux, A., Rzasa-Lynn, R., Crainiceanu, C., & Sharma, T. (2021). Wearable Devices: Current Status and Opportunities in Pain Assessment and Management. *Digital Biomarkers*, *5*(1), 89–102. https://doi.org/10.1159/000515576

Lever, J., Krzywinski, M., & Altman, N. (2016). Model selection and overfitting. *Nature Methods*, *13*(9), 703–704. https://doi.org/10.1038/nmeth.3968

Levi, D., Gispan, L., Giladi, N., & Fetaya, E. (2022). Evaluating and Calibrating Uncertainty Prediction in Regression Tasks. *Sensors*, *22*(15), 5540. https://doi.org/10.3390/s22155540

Levitt, J., Edhi, M. M., Thorpe, R. V., Leung, J. W., Michishita, M., Koyama, S., Yoshikawa, S., Scarfo, K. A., Carayannopoulos, A. G., Gu, W., Srivastava, K. H., Clark, B. A., Esteller, R., Borton, D. A., Jones, S. R., & Saab, C. Y. (2020). Pain phenotypes classified by machine learning using electroencephalography features. *NeuroImage*, *223*, 117256. https://doi.org/10.1016/j.neuroimage.2020.117256

Levitt, J., & Saab, C. Y. (2019). What does a pain 'biomarker' mean and can a machine be taught to measure pain? *Neuroscience Letters*, *702*, 40–43. https://doi.org/10.1016/j.neulet.2018.11.038

Lewin, G. R., & Moshourab, R. (2004). Mechanosensation and pain. *Journal of Neurobiology*, *61*(1), 30–44. https://doi.org/10.1002/neu.20078

Li, A., Wolfe, J. M., & Chen, Z. (2020). Implicitly and explicitly encoded features can guide attention in free viewing. *Journal of Vision*, *20*(6), 8. https://doi.org/10.1167/jov.20.6.8

Li, D., Puntillo, K., & Miaskowski, C. (2008). A Review of Objective Pain Measures for Use With Critical Care Adult Patients Unable to Self-Report. *The Journal of Pain*, *9*(1), 2–10. https://doi.org/10.1016/j.jpain.2007.08.009

Li, L., Huang, G., Lin, Q., Liu, J., Zhang, S., & Zhang, Z. (2018). Magnitude and temporal variability of inter-stimulus EEG modulate the linear relationship between laser-evoked potentials and fast-pain perception. *Frontiers in Neuroscience*, *12*(MAY). https://doi.org/10.3389/fnins.2018.00340

Li, L., Liu, X., Cai, C., Yang, Y., Li, D., Xiao, L., Xiong, D., Hu, L., & Qiu, Y. (2016). Changes of gamma-band oscillatory activity to tonic muscle pain. *Neuroscience Letters*, *627*, 126–131. https://doi.org/10.1016/j.neulet.2016.05.067

Li, X., Zhang, S., Zhang, Q., Wei, X., Pan, Y., Zhao, J., Xin, X., Qin, C., Wang, X., Li, J., Yang, F., Zhao, Y., Yang, M., Wang, Q., Zheng, Z., Zheng, X., Yang, X., Whitlow, C. T., Gurcan, M. N., … Chen, K. (2019). Diagnosis of thyroid cancer using deep convolutional neural network models applied to sonographic images: a retrospective, multicohort, diagnostic study. *The Lancet Oncology*, *20*(2), 193–201. https://doi.org/10.1016/S1470-2045(18)30762-9

Li, Z., Zhang, L., Zeng, Y., Zhao, Q., & Hu, L. (2023). Gamma-band oscillations of pain and nociception: A systematic review and meta-analysis of human and rodent studies. *Neuroscience & Biobehavioral Reviews*, *146*, 105062. https://doi.org/10.1016/j.neubiorev.2023.105062

Liang, N., & Bougrain, L. (2012). Decoding Finger Flexion from Band-Specific ECoG Signals in Humans. *Frontiers in Neuroscience*, *6*. https://doi.org/10.3389/fnins.2012.00091

Liao, W., Zhang, Y., Huang, X., Xu, X., & Peng, X. (2021). “Emoji, I can feel your pain” – Neural responses to facial and emoji expressions of pain. *Biological Psychology*, *163*, 108134. https://doi.org/10.1016/j.biopsycho.2021.108134

Littlewort, G. C., Bartlett, M. S., & Lee, K. (2009). Automatic coding of facial expressions displayed during posed and genuine pain. *Image and Vision Computing*, *27*(12), 1797–1803. https://doi.org/10.1016/j.imavis.2008.12.010

Liu, X., Cruz Rivera, S., Moher, D., Calvert, M. J., & Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nature Medicine*, *26*(9), 1364–1374. https://doi.org/10.1038/s41591-020-1034-x

Liu, X., Faes, L., Calvert, M. J., & Denniston, A. K. (2019). Extension of the CONSORT and SPIRIT statements. *The Lancet*, *394*(10205), 1225. https://doi.org/10.1016/S0140-6736(19)31819-7

Logothetis, N. K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*(7197), 869–878. https://doi.org/10.1038/nature06976

Löken, L. S., Wessberg, J., Morrison, I., McGlone, F., & Olausson, H. (2009). Coding of pleasant touch by unmyelinated afferents in humans. *Nature Neuroscience*, *12*(5), 547–548. https://doi.org/10.1038/nn.2312

Lopes da Silva, F. H. (2011). Event-related potentials: General aspects of methodology and quantification. In *Niedermeyer's Electroencephalography: Basic Principles, Clinical Applications, and Related Fields* (6th ed., pp. 923–934). Wolters Kluwer/Lippincott Williams & Wilkins Health.

Lötsch, J., & Ultsch, A. (2018). Machine learning in pain research. *PAIN*, *159*(4), 623–630. https://doi.org/10.1097/j.pain.0000000000001118

Lovinger, D. M. (2008). Communication networks in the brain: neurons, receptors, neurotransmitters, and alcohol. *Alcohol Research & Health : The Journal of the National Institute on Alcohol Abuse and Alcoholism*, *31*(3), 196–214. http://www.ncbi.nlm.nih.gov/pubmed/23584863

Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT Press.

Lundervold, A. S., & Lundervold, A. (2019). An overview of deep learning in medical imaging focusing on MRI. *Zeitschrift Für Medizinische Physik*, *29*(2), 102–127. https://doi.org/10.1016/j.zemedi.2018.11.002

Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., Shilton, A., Yearwood, J., Dimitrova, N., Ho, T. B., Venkatesh, S., & Berk, M. (2016). Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View. *Journal of Medical Internet Research*, *18*(12), e323. https://doi.org/10.2196/jmir.5870

Luu, P., & Ferree, T. (2005). *Determination of the HydroCel Geodesic Sensor Nets' Average Electrode Positions and Their 10 – 10 International Equivalents*.

Mackey, S., Greely, H. T., & Martucci, K. T. (2019). Neuroimaging-based pain biomarkers:

definitions, clinical and research applications, and evaluation frameworks to achieve personalized pain medicine. *PAIN Reports*, *4*(4), e762. https://doi.org/10.1097/PR9.0000000000000762

Mahesh, B. (2020). Machine Learning Algorithms - A Review. *International Journal of Science and Research (ISJR)*, *9*(1), 381–386.

Maleki, F., Muthukrishnan, N., Ovens, K., Reinhold, C., & Forghani, R. (2020). Machine Learning Algorithm Validation. *Neuroimaging Clinics of North America*, *30*(4), 433–445. https://doi.org/10.1016/j.nic.2020.08.004

Malviya, S., Voepel-Lewis, T., Burke, C., Merkel, S., & Tait, A. R. (2006). The revised FLACC observational pain tool: improved reliability and validity for pain assessment in children with cognitive impairment. *Pediatric Anesthesia*, *16*(3), 258–265. https://doi.org/10.1111/j.1460-9592.2005.01773.x

Mandrekar, J. N. (2010). Receiver Operating Characteristic Curve in Diagnostic Test Assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316. https://doi.org/10.1097/JTO.0b013e3181ec173d

Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate Diabetes Risk Stratification Using Machine Learning: Role of Missing Value and Outliers. *Journal of Medical Systems*, *42*(5), 92. https://doi.org/10.1007/s10916-018-0940-7

Marathe, A. R., Ries, A. J., & McDowell, K. (2014). Sliding HDCA: Single-Trial EEG Classification to Overcome and Quantify Temporal Variability. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *22*(2), 201–211. https://doi.org/10.1109/TNSRE.2014.2304884

Mari, T., Asgard, O., Henderson, J., Hewitt, D., Brown, C., Stancak, A., & Fallon, N. (2023).

External validation of binary machine learning models for pain intensity perception classification from EEG in healthy individuals. *Scientific Reports*, *13*(1), 242. https://doi.org/10.1038/s41598-022-27298-1

Mari, T., Henderson, J., Maden, M., Nevitt, S., Duarte, R., & Fallon, N. (2022). Systematic Review of the Effectiveness of Machine Learning Algorithms for Classifying Pain Intensity, Phenotype or Treatment Outcomes Using Electroencephalogram Data. *The Journal of Pain*, *23*(3), 349–369. https://doi.org/10.1016/j.jpain.2021.07.011

Mateen, B. A., Liley, J., Denniston, A. K., Holmes, C. C., & Vollmer, S. J. (2020). Improving the quality of machine learning in health applications and clinical research. *Nature Machine Intelligence*, *2*(10), 554–556. https://doi.org/10.1038/s42256-020-00239-1

Mathewson, K. J., Hashemi, A., Sheng, B., Sekuler, A. B., Bennett, P. J., & Schmidt, L. A. (2015). Regional electroencephalogram (EEG) alpha power and asymmetry in older adults: a study of short-term test–retest reliability. *Frontiers in Aging Neuroscience*, *7*. https://doi.org/10.3389/fnagi.2015.00177

McGlone, F., & Reilly, D. (2010). The cutaneous sensory system. *Neuroscience and Biobehavioral Reviews*, *34*(2), 148–159. https://doi.org/10.1016/j.neubiorev.2009.08.004

McGuire, D. B., Kaiser, K. S., Haisfield-Wolfe, M. E., & Iyamu, F. (2016). Pain Assessment in Noncommunicative Adult Palliative Care Patients. *Nursing Clinics of North America*, *51*(3), 397–431. https://doi.org/10.1016/j.cnur.2016.05.009

McLain, N. J., Yani, M. S., & Kutch, J. J. (2022). Analytic consistency and neural correlates of peak alpha frequency in the study of pain. *Journal of Neuroscience Methods*, *368*, 109460. https://doi.org/10.1016/j.jneumeth.2021.109460

Mechelli, A., & Vieira, S. (2020). From models to tools: clinical translation of machine learning studies in psychosis. *Npj Schizophrenia*, *6*(1), 4. https://doi.org/10.1038/s41537-020-0094-8

Melzack, R. (1999). From the gate to the neuromatrix. *Pain*, *82*(Supplement 1), S121–S126. https://doi.org/10.1016/S0304-3959(99)00145-1

Melzack, R. (2001). Pain and the neuromatrix in the brain. *Journal of Dental Education*, *65*(12), 1378–1382. http://www.ncbi.nlm.nih.gov/pubmed/11780656

Melzack, R., & Katz, J. (2013). Pain Measurement in Adult Patients. In S. B. McMahon, M. Koltzenburg, I. Tracey, & D. Turk (Eds.), *Wall and Melzack's Textbook of Pain* (pp. 301–314). Saunders.

Mende-Siedlecki, P., Lin, J., Ferron, S., Gibbons, C., Drain, A., & Goharzad, A. (2021). Seeing no pain: Assessing the generalizability of racial bias in pain perception. *Emotion*, *21*(5), 932–950. https://doi.org/10.1037/emo0000953

Mende-Siedlecki, P., Qu-Lee, J., Lin, J., Drain, A., & Goharzad, A. (2020). The Delaware Pain Database: a set of painful expressions and corresponding norming data. *PAIN Reports*, *5*(6), e853. https://doi.org/10.1097/PR9.0000000000000853

Metzger, S. L., Littlejohn, K. T., Silva, A. B., Moses, D. A., Seaton, M. P., Wang, R., Dougherty, M. E., Liu, J. R., Wu, P., Berger, M. A., Zhuravleva, I., Tu-Chan, A., Ganguly, K., Anumanchipalli, G. K., & Chang, E. F. (2023). A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, *620*(7976), 1037–1046. https://doi.org/10.1038/s41586-023-06443-4

Miao, J., & Niu, L. (2016). A Survey on Feature Selection. *Procedia Computer Science*, *91*, 919–926. https://doi.org/10.1016/j.procs.2016.07.111

Miao, Y., Jin, J., Daly, I., Zuo, C., Wang, X., Cichocki, A., & Jung, T.-P. (2021). Learning Common Time-Frequency-Spatial Patterns for Motor Imagery Classification. *IEEE*

Transactions on Neural Systems and Rehabilitation Engineering, *29*, 699–707. https://doi.org/10.1109/TNSRE.2021.3071140

Michail, G., Dresel, C., Witkovský, V., Stankewitz, A., & Schulz, E. (2016). Neuronal Oscillations in Various Frequency Bands Differ between Pain and Touch. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00182

Michel, C. M., & Brunet, D. (2019). EEG Source Imaging: A Practical Review of the Analysis Steps. *Frontiers in Neurology*, *10*. https://doi.org/10.3389/fneur.2019.00325

Michel, C. M., & Murray, M. M. (2012). Towards the utilization of EEG as a brain imaging tool. *NeuroImage*, *61*(2), 371–385. https://doi.org/10.1016/j.neuroimage.2011.12.039

Mienye, I. D., & Sun, Y. (2022). A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects. *IEEE Access*, *10*, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3207287

Millard, S. K., Furman, A. J., Kerr, A., Seminowicz, D. A., Gao, F., Naidu, B. V., & Mazaheri, A. (2022). Predicting postoperative pain in lung cancer patients using preoperative peak alpha frequency. *British Journal of Anaesthesia*, *128*(6), e346–e348. https://doi.org/10.1016/j.bja.2022.03.006

Mischkowski, D., Palacios-Barrios, E. E., Banker, L., Dildine, T. C., & Atlas, L. Y. (2018). Pain or nociception? Subjective experience mediates the effects of acute noxious heat on autonomic responses. *Pain*, *159*(4), 699–711. https://doi.org/10.1097/j.pain.0000000000001132

Misra, G., Ofori, E., Chung, J. W., & Coombes, S. A. (2017). Pain-Related Suppression of Beta Oscillations Facilitates Voluntary Movement. *Cerebral Cortex (New York, N.Y. : 1991)*, *27*(4), 2592–2606. https://doi.org/10.1093/cercor/bhw061

Misra, G., Wang, W., Archer, D. B., Roy, A., & Coombes, S. A. (2017). Automated classification of pain perception using high-density electroencephalography data. *Journal of Neurophysiology*, *117*(2), 786–795. https://doi.org/10.1152/jn.00650.2016

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Monachino, A. D., Lopez, K. L., Pierce, L. J., & Gabard-Durnam, L. J. (2022). The HAPPE plus Event-Related (HAPPE+ER) software: A standardized preprocessing pipeline for event-related potential analyses. *Developmental Cognitive Neuroscience*, *57*, 101140. https://doi.org/10.1016/j.dcn.2022.101140

Moon, S., Song, H.-J., Sharma, V. D., Lyons, K. E., Pahwa, R., Akinwuntan, A. E., & Devos, H. (2020). Classification of Parkinson's disease and essential tremor based on balance and gait characteristics from wearable motion sensors via machine learning techniques: a data-driven approach. *Journal of NeuroEngineering and Rehabilitation*, *17*(1), 125. https://doi.org/10.1186/s12984-020-00756-5

Moons, K. G. M., Altman, D. G., Reitsma, J. B., Ioannidis, J. P. A., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F., & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine*, *162*(1), W1–W73. https://doi.org/10.7326/M14-0698

Moons, K. G. M., Wolff, R. F., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess Risk of Bias and Applicability of Prediction Model Studies: Explanation and Elaboration. *Annals of Internal Medicine*, *170*(1), W1. https://doi.org/10.7326/M18-1377

Moses, D. A., Leonard, M. K., Makin, J. G., & Chang, E. F. (2019). Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nature Communications*, *10*(1), 3096. https://doi.org/10.1038/s41467-019-10994-4

Mouraux, A., Guérit, J. ., & Plaghki, L. (2003). Non-phase locked electroencephalogram (EEG) responses to CO2 laser skin stimulations may reflect central interactions between A∂- and C-fibre afferent volleys. *Clinical Neurophysiology*, *114*(4), 710–722. https://doi.org/10.1016/S1388-2457(03)00027-0

Mouraux, A., & Iannetti, G. D. (2018). The search for pain biomarkers in the human brain. *Brain*, *141*(12), 3290–3307. https://doi.org/10.1093/brain/awy281

Mullen, T. (2012). *CleanLine EEGLAB Plugin*. San Diego, CA: Neuroimaging Informatic Tools and Resources Clearinghouse (NITRC).

Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., & Blankertz, B. (2008). Machine learning for real-time single-trial EEG-analysis: From brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods*, *167*(1), 82–90. https://doi.org/10.1016/j.jneumeth.2007.09.022

Nagendran, M., Chen, Y., Lovejoy, C. A., Gordon, A. C., Komorowski, M., Harvey, H., Topol, E. J., Ioannidis, J. P. A., Collins, G. S., & Maruthappu, M. (2020). Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, m689. https://doi.org/10.1136/bmj.m689

Neuper, C., & Pfurtscheller, G. (2001). Event-related dynamics of cortical rhythms: frequency-specific features and functional correlates. *International Journal of Psychophysiology*, *43*(1), 41–58. https://doi.org/10.1016/S0167-8760(01)00178-7

Nezam, T., Boostani, R., Abootalebi, V., & Rastegar, K. (2018). A Novel Classification Strategy to Distinguish Five Levels of Pain using the EEG Signal Features. *IEEE Transactions on Affective Computing*, 1–1. https://doi.org/10.1109/TAFFC.2018.2851236

Nezam, T., Boostani, R., Abootalebi, V., & Rastegar, K. (2021). A Novel Classification Strategy to Distinguish Five Levels of Pain Using the EEG Signal Features. *IEEE Transactions on Affective Computing*, *12*(1), 131–140. https://doi.org/10.1109/TAFFC.2018.2851236

Nickel, M. M., May, E. S., Tiemann, L., Schmidt, P., Postorino, M., Ta Dinh, S., Gross, J., & Ploner, M. (2017). Brain oscillations differentially encode noxious stimulus intensity and pain intensity. *NeuroImage*, *148*, 141–147. https://doi.org/10.1016/j.neuroimage.2017.01.011

Nielsen, C. S., Staud, R., & Price, D. D. (2009). Individual Differences in Pain Sensitivity: Measurement, Causation, and Consequences. *The Journal of Pain*, *10*(3), 231–237. https://doi.org/10.1016/j.jpain.2008.09.010

Nir, R.-R., Sinai, A., Moont, R., Harari, E., & Yarnitsky, D. (2012). Tonic pain and continuous EEG: Prediction of subjective pain perception by alpha-1 power during stimulation and at rest. *Clinical Neurophysiology*, *123*(3), 605–612. https://doi.org/10.1016/j.clinph.2011.08.006

Nolan, H., Whelan, R., & Reilly, R. B. (2010). FASTER: Fully Automated Statistical Thresholding for EEG artifact Rejection. *Journal of Neuroscience Methods*, *192*(1), 152–162. https://doi.org/10.1016/j.jneumeth.2010.07.015

Nunez, P. L., & Srinivasan, R. (2006). *Electric Fields of the Brain: The neurophysics of EEG* (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195050387.001.0001

Nunez, P. L., Srinivasan, R., Westdorp, A. F., Wijesinghe, R. S., Tucker, D. M., Silberstein, R. B., & Cadusch, P. J. (1997). EEG coherency. I: Statistics, reference electrode, volume conduction, Laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalography and Clinical Neurophysiology*, *103*(5), 499–515.

https://doi.org/10.1016/S0013-4694(97)00066-7

Nuwer, M. R. (1988). Quantitative EEG: I. Techniques and problems of frequency analysis and topographic mapping. *Journal of Clinical Neurophysiology*, *5*(1), 1–44.

Oberman, L. M., Hubbard, E. M., McCleery, J. P., Altschuler, E. L., Ramachandran, V. S., & Pineda, J. A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Cognitive Brain Research*, *24*(2), 190–198. https://doi.org/10.1016/j.cogbrainres.2005.01.014

Office for National Statistics, (ONS). (2021). *Ethnic group, England and Wales: Census 2021*.

Okolo, C., & Omurtag, A. (2018). Research : Use of Dry Electroencephalogram and Support Vector for Objective Pain Assessment. *Biomedical Instrumentation & Technology*, *52*(5), 372–378. https://doi.org/10.2345/0899-8205-52.5.372

Olausson, H., Cole, J., Rylander, K., McGlone, F., Lamarre, Y., Wallin, B. G., Krämer, H., Wessberg, J., Elam, M., Bushnell, M. C., & Vallbo, Å. (2007). Functional role of unmyelinated tactile afferents in human hairy skin: sympathetic response and perceptual localization. *Experimental Brain Research*, *184*(1), 135–140. https://doi.org/10.1007/s00221-007-1175-x

Olejniczak, P. (2006). Neurophysiologic Basis of EEG. *Journal of Clinical Neurophysiology*, *23*(3), 186–189. https://doi.org/10.1097/01.wnp.0000220079.61973.6c

Oosterman, J. M., Zwakhalen, S., Sampson, E. L., & Kunz, M. (2016). The use of facial expressions for pain assessment purposes in dementia: a narrative review. *Neurodegenerative Disease Management*, *6*(2), 119–131. https://doi.org/10.2217/nmt-2015-0006

Osborn, J., & Derbyshire, S. W. G. (2010). Pain sensation evoked by observing injury in others. *Pain*, *148*(2), 268–274. https://doi.org/10.1016/j.pain.2009.11.007

Osisanwo, F. ., Akinsola, J. E. ., Awodele, O., Hinmikaiye, J. O., Olakanmi, O., & Akinjobi, J. (2017). Supervised Machine Learning Algorithms: Classification and Comparison. *International Journal of Computer Trends and Technology*, *48*(3), 128–138. https://doi.org/10.14445/22312803/IJCTT-V48P126

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, n71. https://doi.org/10.1136/bmj.n71

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … McKenzie, J. E. (2021). PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ*, n160. https://doi.org/10.1136/bmj.n160

Pantelopoulos, A., & Bourbakis, N. G. (2010). A Survey on Wearable Sensor-Based Systems for Health Monitoring and Prognosis. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *40*(1), 1–12. https://doi.org/10.1109/TSMCC.2009.2032660

Pascual-Marqui, R. D. (1999). Review of Methods for Solving the EEG Inverse Problem. *International Journal of Bioelectromagnetism*, *1*(1), 75–86.

Paul, J. K., Iype, T., R, D., Hagiwara, Y., Koh, J. E. W., & Acharya, U. R. (2019). Characterization of fibromyalgia using sleep EEG signals with nonlinear dynamical features. *Computers in Biology and Medicine*, *111*.

https://doi.org/10.1016/j.compbiomed.2019.103331

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825--2830.

Peirce, J. W. (2007). PsychoPy—Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1–2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Pereira, F. C., & Borysov, S. S. (2019). Machine Learning Fundamentals. In *Mobility Patterns, Big Data and Transport Analytics* (pp. 9–29). Elsevier. https://doi.org/10.1016/B978-0-12-812970-8.00002-6

Perry, A., Bentin, S., Bartal, I. B.-A., Lamm, C., & Decety, J. (2010). “Feeling” the pain of those who are different from us: Modulation of EEG in the mu/alpha range. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(4), 493–504. https://doi.org/10.3758/CABN.10.4.493

Petre, B., Kragel, P., Atlas, L. Y., Geuter, S., Jepma, M., Koban, L., Krishnan, A., Lopez-Sola, M., Losin, E. A. R., Roy, M., Woo, C.-W., & Wager, T. D. (2022). A multistudy analysis reveals that evoked pain intensity representation is distributed across brain systems. *PLOS Biology*, *20*(5), e3001620. https://doi.org/10.1371/journal.pbio.3001620

Peyron, R., Laurent, B., & García-Larrea, L. (2000). Functional imaging of brain responses to pain. A review and meta-analysis (2000). *Neurophysiologie Clinique/Clinical Neurophysiology*, *30*(5), 263–288. https://doi.org/10.1016/S0987-7053(00)00227-6

Pfurtscheller, G. (1992). Event-related synchronization (ERS): an electrophysiological correlate of cortical areas at rest. *Electroencephalography and Clinical Neurophysiology*, *83*(1), 62–69. https://doi.org/10.1016/0013-4694(92)90133-3

Pfurtscheller, G. (2001). Functional brain imaging based on ERD/ERS. *Vision Research*, *41*(10–11), 1257–1260. https://doi.org/10.1016/S0042-6989(00)00235-2

Pfurtscheller, G., & Aranibar, A. (1977). Event-related cortical desynchronization detected by power measurements of scalp EEG. *Electroencephalography and Clinical Neurophysiology*, *42*(6), 817–826. https://doi.org/10.1016/0013-4694(77)90235-8

Pfurtscheller, G., & Aranibar, A. (1979). Evaluation of event-related desynchronization (ERD) preceding and following voluntary self-paced movement. *Electroencephalography and Clinical Neurophysiology*, *46*(2), 138–146. https://doi.org/10.1016/0013-4694(79)90063-4

Pfurtscheller, G., & Lopes da Silva, F. H. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology*, *110*(11), 1842–1857. https://doi.org/10.1016/S1388-2457(99)00141-8

Pfurtscheller, G., & Neuper, C. (1992). Simultaneous EEG 10 Hz desynchronization and 40 Hz synchronization during finger movements. *NeuroReport*, *3*(12), 1057–1060. https://doi.org/10.1097/00001756-199212000-00006

Pierson, E., Cutler, D. M., Leskovec, J., Mullainathan, S., & Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. *Nature Medicine*, *27*(1), 136–140. https://doi.org/10.1038/s41591-020-01192-7

Pistoia, F., Conson, M., Trojano, L., Grossi, D., Ponari, M., Colonnese, C., Pistoia, M. L., Carducci, F., & Sara, M. (2010). Impaired Conscious Recognition of Negative Facial Expressions in Patients with Locked-in Syndrome. *Journal of Neuroscience*, *30*(23), 7838–7844. https://doi.org/10.1523/JNEUROSCI.6300-09.2010

Pizzo, F., Roehri, N., Medina Villalon, S., Trébuchon, A., Chen, S., Lagarde, S., Carron, R., Gavaret, M., Giusiano, B., McGonigal, A., Bartolomei, F., Badier, J. M., & Bénar, C. G. (2019). Deep brain activities can be detected with magnetoencephalography. *Nature Communications*, *10*(1), 971. https://doi.org/10.1038/s41467-019-08665-5

Ploner, M., Gross, J., Timmermann, L., Pollok, B., & Schnitzler, A. (2006). Pain Suppresses Spontaneous Brain Rhythms. *Cerebral Cortex*, *16*(4), 537–540. https://doi.org/10.1093/cercor/bhj001

Ploner, M., Gross, J., Timmermann, L., & Schnitzler, A. (2002). Cortical representation of first and second pain sensation in humans. *Proceedings of the National Academy of Sciences*, *99*(19), 12444–12448. https://doi.org/10.1073/pnas.182272899

Ploner, M., & May, E. S. (2018). Electroencephalography and magnetoencephalography in pain research—current state and future perspectives. *Pain*, *159*(2), 206–211. https://doi.org/10.1097/j.pain.0000000000001087

Ploner, M., Sorg, C., & Gross, J. (2017). Brain Rhythms of Pain. *Trends in Cognitive Sciences*, *21*(2), 100–110. https://doi.org/10.1016/j.tics.2016.12.001

Powers, D. M. (2011). Evaluation: from Precision, Recall and F-measure to ROC, Informedness, Markedness and Correlation. *Journal of Machine Learning Technologies*, *2*(1), 37–63.

Prerau, M. J., Brown, R. E., Bianchi, M. T., Ellenbogen, J. M., & Purdon, P. L. (2017). Sleep Neurophysiological Dynamics Through the Lens of Multitaper Spectral Analysis. *Physiology*, *32*(1), 60–92. https://doi.org/10.1152/physiol.00062.2015

Preusche, I., & Lamm, C. (2016). Reflections on empathy in medical education: What can we learn from social neurosciences? *Advances in Health Sciences Education*, *21*(1), 235–249. https://doi.org/10.1007/s10459-015-9581-5

Prichep, L. S., Shah, J., Merkin, H., & Hiesiger, E. M. (2018). Exploration of the

Pathophysiology of Chronic Pain Using Quantitative EEG Source Localization. *Clinical EEG and Neuroscience*, *49*(2), 103–113. https://doi.org/10.1177/1550059417736444

Priebe, J. A., Kunz, M., Morcinek, C., Rieckmann, P., & Lautenbacher, S. (2015). Does Parkinson's disease lead to alterations in the facial expression of pain? *Journal of the Neurological Sciences*, *359*(1–2), 226–235. https://doi.org/10.1016/j.jns.2015.10.056

Pritchard, W. S. (1992). The Brain in Fractal Time: 1/F-Like Power Spectrum Scaling of the Human Electroencephalogram. *International Journal of Neuroscience*, *66*(1–2), 119–129. https://doi.org/10.3109/00207459208999796

Prkachin, K. M. (2009). Assessing Pain by Facial Expression: Facial Expression as Nexus. *Pain Research and Management*, *14*(1), 53–58. https://doi.org/10.1155/2009/542964

Pryse-Phillips, W. E. M., Dodick, D. W., Edmeads, J. G., Gawel, M. J., Nelson, R. F., Allan Purdy, R., Robinson, G., Stirling, D., & Worthington, I. (1997). Guidelines for the diagnosis and management of migraine in clinical practice. *Cmaj*, *156*(9), 1273–1287.

Pu, Y., Cheyne, D. O., Cornwell, B. R., & Johnson, B. W. (2018). Non-invasive Investigation of Human Hippocampal Rhythms Using Magnetoencephalography: A Review. *Frontiers in Neuroscience*, *12*. https://doi.org/10.3389/fnins.2018.00273

Purves, D., Augustine, G. J., Fitzpatrick, D., Hall, W. C., LaMantia, A.-S., Mooney, R. D., Platt, M. L., & White, L. E. (2017). *Neuroscience* (6th ed.). Oxford University Press.

Quiton, R. L., & Greenspan, J. D. (2008). Across- and within-session variability of ratings of painful contact heat stimuli. *Pain*, *137*(2), 245–256. https://doi.org/10.1016/j.pain.2007.08.034

Raja, S. N., Carr, D. B., Cohen, M., Finnerup, N. B., Flor, H., Gibson, S., Keefe, F. J., Mogil, J. S., Ringkamp, M., Sluka, K. A., Song, X.-J., Stevens, B., Sullivan, M. D., Tutelman, P. R., Ushida, T., & Vader, K. (2020). The revised International Association for the Study of

Pain definition of pain: concepts, challenges, and compromises. *Pain*, *161*(9), 1976–1982. https://doi.org/10.1097/j.pain.0000000000001939

Rajput, D., Wang, W.-J., & Chen, C.-C. (2023). Evaluation of a decided sample size in machine learning applications. *BMC Bioinformatics*, *24*(1), 48. https://doi.org/10.1186/s12859-023-05156-9

Ramspek, C. L., Jager, K. J., Dekker, F. W., Zoccali, C., & van Diepen, M. (2021). External validation of prognostic models: what, why, how, when and where? *Clinical Kidney Journal*, *14*(1), 49–58. https://doi.org/10.1093/ckj/sfaa188

Rashid, M., Sulaiman, N., P. P. Abdul Majeed, A., Musa, R. M., Ab. Nasir, A. F., Bari, B. S., & Khatun, S. (2020). Current Status, Challenges, and Possible Solutions of EEG-Based Brain-Computer Interface: A Comprehensive Review. *Frontiers in Neurorobotics*, *14*. https://doi.org/10.3389/fnbot.2020.00025

Ratcliffe, G. E., Enns, M. W., Belik, S.-L., & Sareen, J. (2008). Chronic Pain Conditions and Suicidal Ideation and Suicide Attempts: An Epidemiologic Perspective. *The Clinical Journal of Pain*, *24*(3), 204–210. https://doi.org/10.1097/AJP.0b013e31815ca2a3

Rexed, B. (1952). The cytoarchitectonic organization of the spinal cord in the cat. *The Journal of Comparative Neurology*, *96*(3), 415–495. https://doi.org/10.1002/cne.900960303

Riečanský, I., Paul, N., Kölble, S., Stieger, S., & Lamm, C. (2015). Beta oscillations reveal ethnicity ingroup bias in sensorimotor resonance to pain of others. *Social Cognitive and Affective Neuroscience*, *10*(7), 893–901. https://doi.org/10.1093/scan/nsu139

Rivet, B., Souloumiac, A., Attina, V., & Gibert, G. (2009). xDAWN Algorithm to Enhance Evoked Potentials: Application to Brain–Computer Interface. *IEEE Transactions on Biomedical Engineering*, *56*(8), 2035–2043.

https://doi.org/10.1109/TBME.2009.2012869

Rockholt, M. M., Kenefati, G., Doan, L. V., Chen, Z. S., & Wang, J. (2023). In search of a composite biomarker for chronic pain by way of EEG and machine learning: where do we currently stand? *Frontiers in Neuroscience*, *17*. https://doi.org/10.3389/fnins.2023.1186418

Roy, S. D., Bhowmik, M. K., Saha, P., & Ghosh, A. K. (2016). An Approach for Automatic Pain Detection through Facial Expression. *Procedia Computer Science*, *84*, 99–106. https://doi.org/10.1016/j.procs.2016.04.072

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery*, *8*(4), 1–18. https://doi.org/10.1002/widm.1249

Sai, C. Y., Mokhtar, N., Yip, H. W., Bak, L. L. M., Hasan, M. S., Arof, H., Cumming, P., & Mat Adenan, N. A. (2019). Objective identification of pain due to uterine contraction during the first stage of labour using continuous EEG signals and SVM. *Sādhanā*, *44*(4), 87. https://doi.org/10.1007/s12046-019-1058-4

Saif, M. G. M., Hassan, M. A., & Vuckovic, A. (2021). Efficacy evaluation of neurofeedback applied for treatment of central neuropathic pain using machine learning. *SN Applied Sciences*, *3*(1), 58. https://doi.org/10.1007/s42452-020-04035-9

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Salehinejad, H., Kitamura, J., Ditkofsky, N., Lin, A., Bharatha, A., Suthiphosuwan, S., Lin, H.-

M., Wilson, J. R., Mamdani, M., & Colak, E. (2021). A real-world demonstration of machine learning generalizability in the detection of intracranial hemorrhage on head computerized tomography. *Scientific Reports*, *11*(1), 17051. https://doi.org/10.1038/s41598-021-95533-2

Samuel, A. L. (1959). Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. https://doi.org/10.1147/rd.33.0210

Sapir-Pichhadze, R., & Kaplan, B. (2020). Seeing the Forest for the Trees: Random Forest Models for Predicting Survival in Kidney Transplant Recipients. *Transplantation*, *104*(5), 905–906. https://doi.org/10.1097/TP.0000000000002923

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x

Sarnthein, J., Stern, J., Aufenberg, C., Rousson, V., & Jeanmonod, D. (2006). Increased EEG power and slowed dominant frequency in patients with neurogenic pain. *Brain*, *129*(1), 55–64. https://doi.org/10.1093/brain/awh631

Sawa, T., Yamada, T., & Obata, Y. (2022). Power spectrum and spectrogram of EEG analysis during general anesthesia: Python-based computer programming analysis. *Journal of Clinical Monitoring and Computing*, *36*(3), 609–621. https://doi.org/10.1007/s10877-021-00771-4

Schaible, H.-G., Ebersberger, A., & Natura, G. (2011). Update on peripheral mechanisms of pain: beyond prostaglandins and cytokines. *Arthritis Research & Therapy*, *13*(2), 210. https://doi.org/10.1186/ar3305

Schalk, G., & Leuthardt, E. C. (2011). Brain-Computer Interfaces Using Electrocorticographic

Signals. *IEEE Reviews in Biomedical Engineering*, *4*, 140–154. https://doi.org/10.1109/RBME.2011.2172408

Schiavenato, M., & Craig, K. D. (2010). Pain Assessment as a Social Transaction. *The Clinical Journal of Pain*, *26*(8), 667–676. https://doi.org/10.1097/AJP.0b013e3181e72507

Schmidt, R., Schmelz, M., Forster, C., Ringkamp, M., Torebjork, E., & Handwerker, H. (1995). Novel classes of responsive and unresponsive C nociceptors in human skin. *The Journal of Neuroscience*, *15*(1), 333–341. https://doi.org/10.1523/JNEUROSCI.15-01-00333.1995

Schmidt, R., Schmelz, M., Ringkamp, M., Handwerker, H. O., & Torebjörk, H. E. (1997). Innervation Territories of Mechanically Activated C Nociceptor Units in Human Skin. *Journal of Neurophysiology*, *78*(5), 2641–2648. https://doi.org/10.1152/jn.1997.78.5.2641

Schnakers, C., & Zasler, N. D. (2007). Pain assessment and management in disorders of consciousness. *Current Opinion in Neurology*, *20*(6), 620–626. https://doi.org/10.1097/WCO.0b013e3282f169d9

Schomer, D. L., & Lopes, D. S. F. (2010). *Niedermeyer's electroencephalography : Basic principles, clinical applications, and related fields*. Wolters Kluwer Health.

Schulz, E., May, E. S., Postorino, M., Tiemann, L., Nickel, M. M., Witkovsky, V., Schmidt, P., Gross, J., & Ploner, M. (2015). Prefrontal Gamma Oscillations Encode Tonic Pain in Humans. *Cerebral Cortex*, *25*(11), 4407–4414. https://doi.org/10.1093/cercor/bhv043

Schulz, E., Zherdin, A., Tiemann, L., Plant, C., & Ploner, M. (2012). Decoding an Individual's Sensitivity to Pain from the Multivariate Analysis of EEG Data. *Cerebral Cortex*, *22*(5), 1118–1123. https://doi.org/10.1093/cercor/bhr186

Schwiening, C. J. (2012). A brief historical perspective: Hodgkin and Huxley. *The Journal of*

Physiology, *590*(11), 2571–2575. https://doi.org/10.1113/jphysiol.2012.230458

Seers, T., Derry, S., Seers, K., & Moore, R. A. (2018). Professionals underestimate patients' pain: a comprehensive review. *Pain*, *159*(5), 811–818. https://doi.org/10.1097/j.pain.0000000000001165

Seki, Y., Miyashita, T., Kandori, A., Maki, A., & Koizumi, H. (2012). Simultaneous measurement of neuronal activity and cortical hemodynamics by unshielded magnetoencephalography and near-infrared spectroscopy. *Journal of Biomedical Optics*, *17*(10), 1070011. https://doi.org/10.1117/1.JBO.17.10.107001

Seneviratne, M. G., Shah, N. H., & Chu, L. (2020). Bridging the implementation gap of machine learning in healthcare. *BMJ Innovations*, *6*(2), 45–47. https://doi.org/10.1136/bmjinnov-2019-000359

Senn, S. (2003). Disappointing dichotomies. *Pharmaceutical Statistics*, *2*(4), 239–240. https://doi.org/10.1002/pst.90

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., Ringel, M., & Schork, N. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. *Npj Digital Medicine*, *2*(1), 69. https://doi.org/10.1038/s41746-019-0148-3

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., & Stewart, L. A. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ*, *349*(jan02 1), g7647–g7647. https://doi.org/10.1136/bmj.g7647

Shao, S., Shen, K., Yu, K., Wilder-Smith, E. P. V., & Li, X. (2012). Frequency-domain EEG source analysis for acute tonic cold pain perception. *Clinical Neurophysiology*, *123*(10), 2042–2049. https://doi.org/10.1016/j.clinph.2012.02.084

Shirvalkar, P., Prosky, J., Chin, G., Ahmadipour, P., Sani, O. G., Desai, M., Schmitgen, A., Dawes, H., Shanechi, M. M., Starr, P. A., & Chang, E. F. (2023). First-in-human prediction of chronic pain state using intracranial neural biomarkers. *Nature Neuroscience*, *26*(6), 1090–1099. https://doi.org/10.1038/s41593-023-01338-z

Simon, M. V., Nuwer, M. R., & Szelényi, A. (2022). Electroencephalography, electrocorticography, and cortical stimulation techniques. In *Handbook of clinical neurology* (pp. 11–38). https://doi.org/10.1016/B978-0-12-819826-1.00001-6

Simons, L. E., Elman, I., & Borsook, D. (2014). Psychological processing in chronic pain: A neural systems approach. *Neuroscience & Biobehavioral Reviews*, *39*, 61–78. https://doi.org/10.1016/j.neubiorev.2013.12.006

Singer, T., & Lamm, C. (2009). The Social Neuroscience of Empathy. *Annals of the New York Academy of Sciences*, *1156*(1), 81–96. https://doi.org/10.1111/j.1749-6632.2009.04418.x

Singer, T., Seymour, B., O’Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. *Science*, *303*(5661), 1157–1162. https://doi.org/10.1126/science.1093535

Singh, A. K., & Krishnan, S. (2023). Trends in EEG signal feature extraction applications. *Frontiers in Artificial Intelligence*, *5*. https://doi.org/10.3389/frai.2022.1072801

Singh, S. (2014). Magnetoencephalography: Basic principles. *Annals of Indian Academy of Neurology*, *17*(5), 107. https://doi.org/10.4103/0972-2327.128676

Siontis, G. C. M., Tzoulaki, I., Castaldi, P. J., & Ioannidis, J. P. A. (2015). External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *Journal of Clinical Epidemiology*, *68*(1), 25–34. https://doi.org/10.1016/j.jclinepi.2014.09.007

Slater, R., Worley, A., Fabrizi, L., Roberts, S., Meek, J., Boyd, S., & Fitzgerald, M. (2010). Evoked potentials generated by noxious stimulation in the human infant brain. *European Journal of Pain*, *14*(3), 321–326. https://doi.org/10.1016/j.ejpain.2009.05.005

Slepian, D., & Pollak, H. O. (1961). Prolate Spheroidal Wave Functions, Fourier Analysis and Uncertainty - I. *Bell System Technical Journal*, *40*(1), 43–63. https://doi.org/10.1002/j.1538-7305.1961.tb03976.x

Smith, E. S. J., & Lewin, G. R. (2009). Nociceptors: a phylogenetic view. *Journal of Comparative Physiology A*, *195*(12), 1089–1106. https://doi.org/10.1007/s00359-009-0482-z

Snapinn, S. M., & Jiang, Q. (2007). Responder analyses and the assessment of a clinically relevant treatment effect. *Trials*, *8*(1), 31. https://doi.org/10.1186/1745-6215-8-31

Sneddon, L. U. (2018). Comparative Physiology of Nociception and Pain. *Physiology*, *33*(1), 63–73. https://doi.org/10.1152/physiol.00022.2017

Snell, K. I. E., Archer, L., Ensor, J., Bonnett, L. J., Debray, T. P. A., Phillips, B., Collins, G. S., & Riley, R. D. (2021). External validation of clinical prediction models: simulation-based sample size calculations were more reliable than rules-of-thumb. *Journal of Clinical Epidemiology*, *135*, 79–89. https://doi.org/10.1016/j.jclinepi.2021.02.011

Snell, R. S. (2009). *Clinical Neuroanatomy* (7th ed.). Lippincott Williams and Wilkins.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Soto, V., Tyson-Carr, J., Kokmotou, K., Roberts, H., Cook, S., Fallon, N., Giesbrecht, T., & Stancak, A. (2018). Brain Responses to Emotional Faces in Natural Settings: A Wireless

Mobile EEG Recording Study. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.02003

Spruston, N. (2008). Pyramidal neurons: dendritic structure and synaptic integration. *Nature Reviews Neuroscience*, *9*(3), 206–221. https://doi.org/10.1038/nrn2286

Srinivasan, R., Nunez, P. L., Tucker, D. M., Silberstein, R. B., & Cadusch, P. J. (1996). Spatial sampling and filtering of EEG with spline Laplacians to estimate cortical potentials. *Brain Topography*, *8*(4), 355–366. https://doi.org/10.1007/BF01186911

Steeds, C. E. (2009). The anatomy and physiology of pain. *Surgery (Oxford)*, *27*(12), 507–511. https://doi.org/10.1016/j.mpsur.2009.10.013

Stefan, H., & Trinka, E. (2017). Magnetoencephalography (MEG): Past, current and future perspectives for improved differentiation and treatment of epilepsies. *Seizure*, *44*, 121–124. https://doi.org/10.1016/j.seizure.2016.10.028

Stern, J., Jeanmonod, D., & Sarnthein, J. (2006). Persistent EEG overactivation in the cortical pain matrix of neurogenic pain patients. *NeuroImage*, *31*(2), 721–731. https://doi.org/10.1016/j.neuroimage.2005.12.042

Stevens, A., Batra, A., Kötter, I., Bartels, M., & Schwarz, J. (2000). Both pain and EEG response to cold pressor stimulation occurs faster in fibromyalgia patients than in control subjects. *Psychiatry Research*, *97*(2–3), 237–247. https://doi.org/10.1016/S0165-1781(00)00223-7

Stewart, A. X., Nuthmann, A., & Sanguinetti, G. (2014). Single-trial classification of EEG in a visual object task using ICA and machine learning. *Journal of Neuroscience Methods*, *228*, 1–14. https://doi.org/10.1016/j.jneumeth.2014.02.014

Steyerberg, E. W., & Harrell, F. E. (2016). Prediction models need appropriate internal, internal–external, and external validation. *Journal of Clinical Epidemiology*, *69*, 245–

247. https://doi.org/10.1016/j.jclinepi.2015.04.005

Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., Pencina, M. J., & Kattan, M. W. (2010). Assessing the Performance of Prediction Models: a framework for some traditional and novel measures. *Epidemiology*, *21*(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2

Stuart, G., Spruston, N., Sakmann, B., & Häusser, M. (1997). Action potential initiation and backpropagation in neurons of the mammalian CNS. *Trends in Neurosciences*, *20*(3), 125–131. https://doi.org/10.1016/S0166-2236(96)10075-8

Su, Q., Song, Y., Zhao, R., & Liang, M. (2019). A review on the ongoing quest for a pain signature in the human brain. *Brain Science Advances*, *5*(4), 274–287. https://doi.org/10.26599/BSA.2019.9050024

Subasi, A., Ahmed, A., Aličković, E., & Rashik Hassan, A. (2019). Effect of photic stimulation for migraine detection using random forest and discrete wavelet transform. *Biomedical Signal Processing and Control*, *49*, 231–239. https://doi.org/10.1016/j.bspc.2018.12.011

Subha, D. P., Joseph, P. K., Acharya U, R., & Lim, C. M. (2010). EEG Signal Analysis: A Survey. *Journal of Medical Systems*, *34*(2), 195–212. https://doi.org/10.1007/s10916-008-9231-z

Sullivan, M. J. L., Bishop, S. R., & Pivik, J. (1995). The Pain Catastrophizing Scale: Development and validation. *Psychological Assessment*, *7*(4), 524–532. https://doi.org/10.1037/1040-3590.7.4.524

Sullivan, T. J., Deiss, S. R., Tzyy-Ping Jung, & Cauwenberghs, G. (2008). A brain-machine interface using dry-contact, low-noise EEG sensors. *2008 IEEE International Symposium on Circuits and Systems*, 1986–1989. https://doi.org/10.1109/ISCAS.2008.4541835

Sur, S., & Sinha, V. (2009). Event-related potential: An overview. *Industrial Psychiatry Journal*, *18*(1), 70. https://doi.org/10.4103/0972-6748.57865

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, *5*(1), 15924. https://doi.org/10.1038/srep15924

Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM Parameter Optimization using Grid Search and Genetic Algorithm to Improve Classification Performance. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, *14*(4), 1502. https://doi.org/10.12928/telkomnika.v14i4.3956

Ta Dinh, S., Nickel, M. M., Tiemann, L., May, E. S., Heitmann, H., Hohn, V. D., Edenharter, G., Utpadel-Fischler, D., Tölle, T. R., Sauseng, P., Gross, J., & Ploner, M. (2019). Brain dysfunction in chronic pain patients assessed by resting-state electroencephalography. *PAIN*, *160*(12), 2751–2765. https://doi.org/10.1097/j.pain.0000000000001666

Tallgren, P., Vanhatalo, S., Kaila, K., & Voipio, J. (2005). Evaluation of commercially available electrodes and gels for recording of slow EEG potentials. *Clinical Neurophysiology*, *116*(4), 799–806. https://doi.org/10.1016/j.clinph.2004.10.001

Tanasescu, R., Cottam, W. J., Condon, L., Tench, C. R., & Auer, D. P. (2016). Functional reorganisation in chronic pain and neural correlates of pain sensitisation: A coordinate based meta-analysis of 266 cutaneous pain fMRI studies. *Neuroscience & Biobehavioral Reviews*, *68*, 120–133. https://doi.org/10.1016/j.neubiorev.2016.04.001

Tarca, A. L., Carey, V. J., Chen, X., Romero, R., & Drăghici, S. (2007). Machine Learning and Its Applications to Biology. *PLoS Computational Biology*, *3*(6), e116. https://doi.org/10.1371/journal.pcbi.0030116

Teichmann, D., Klopp, J., Hallmann, A., Schuett, K., Wolfart, S., & Teichmann, M. (2018).

Detection of acute periodontal pain from physiological signals. *Physiological Measurement*, *39*(9), 095007. https://doi.org/10.1088/1361-6579/aadf0c

Teplan, M. (2002). Fundamentals of EEG measurement. *Measurement Science Review*, *2*(2), 1–11.

Tharwat, A. (2020). Classification assessment methods. *New England Journal of Entrepreneurship*, *17*, 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J., & van Lier, R. (2019). Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience. *Scientific Reports*, *9*(1), 17456. https://doi.org/10.1038/s41598-019-54018-z

Tiemann, L., May, E. S., Postorino, M., Schulz, E., Nickel, M. M., Bingel, U., & Ploner, M. (2015). Differential neurophysiological correlates of bottom-up and top-down modulations of pain. *Pain*, *156*(2), 289–296. https://doi.org/10.1097/01.j.pain.0000460309.94442.44

Timmermann, L., Ploner, M., Haucke, K., Schmitz, F., Baltissen, R., & Schnitzler, A. (2001). Differential Coding of Pain Intensity in the Human Primary and Secondary Somatosensory Cortex. *Journal of Neurophysiology*, *86*(3), 1499–1503. https://doi.org/10.1152/jn.2001.86.3.1499

Timmers, I., Park, A. L., Fischer, M. D., Kronman, C. A., Heathcote, L. C., Hernandez, J. M., & Simons, L. E. (2018). Is Empathy for Pain Unique in Its Neural Correlates? A Meta-Analysis of Neuroimaging Studies of Empathy. *Frontiers in Behavioral Neuroscience*, *12*. https://doi.org/10.3389/fnbeh.2018.00289

Tivadar, R. I., & Murray, M. M. (2019). A Primer on Electroencephalography and Event-Related Potentials for Organizational Neuroscience. *Organizational Research Methods*, *22*(1), 69–94. https://doi.org/10.1177/1094428118804657

Todd, A. J. (2010). Neuronal circuitry for pain processing in the dorsal horn. *Nature Reviews Neuroscience*, *11*(12), 823–836. https://doi.org/10.1038/nrn2947

Torebjörk, H. E., Lundberg, L. E., & LaMotte, R. H. (1992). Central changes in processing of mechanoreceptive input in capsaicin-induced secondary hyperalgesia in humans. *The Journal of Physiology*, *448*(1), 765–780. https://doi.org/10.1113/jphysiol.1992.sp019069

Torta, D. M., Legrain, V., Mouraux, A., & Valentini, E. (2017). Attention to pain! A neurocognitive perspective on attentional modulation of pain in neuroimaging studies. *Cortex*, *89*, 120–134. https://doi.org/10.1016/j.cortex.2017.01.010

Tracey, I., & Mantyh, P. W. (2007). The Cerebral Signature for Pain Perception and Its Modulation. *Neuron*, *55*(3), 377–391. https://doi.org/10.1016/j.neuron.2007.07.012

Tracey, I., Woolf, C. J., & Andrews, N. A. (2019). Composite Pain Biomarker Signatures for Objective Assessment and Effective Treatment. *Neuron*, *101*(5), 783–800. https://doi.org/10.1016/j.neuron.2019.02.019

Tracey, W. D. (2017). Nociception. *Current Biology*, *27*(4), R129–R133. https://doi.org/10.1016/j.cub.2017.01.037

Treede, R.-D., Meyer, R. A., & Campbell, J. N. (1998). Myelinated Mechanically Insensitive Afferents From Monkey Hairy Skin: Heat-Response Properties. *Journal of Neurophysiology*, *80*(3), 1082–1093. https://doi.org/10.1152/jn.1998.80.3.1082

Treede, R.-D., Rief, W., Barke, A., Aziz, Q., Bennett, M. I., Benoliel, R., Cohen, M., Evers, S., Finnerup, N. B., First, M. B., Giamberardino, M. A., Kaasa, S., Kosek, E., Lavand'homme, P., Nicholas, M., Perrot, S., Scholz, J., Schug, S., Smith, B. H., … Wang, S.-J. (2015). A classification of chronic pain for ICD-11. *Pain*, *156*(6), 1003–1007. https://doi.org/10.1097/j.pain.0000000000000160

Tripanpitak, K., Viriyavit, W., Huang, S. Y., & Yu, W. (2020). Classification of Pain Event Related Potential for Evaluation of Pain Perception Induced by Electrical Stimulation. *Sensors*, *20*(5), 1491. https://doi.org/10.3390/s20051491

Tu, Y., Tan, A., Bai, Y., Hung, Y. S., & Zhang, Z. (2016). Decoding subjective intensity of nociceptive pain from pre-stimulus and post-stimulus brain activities. *Frontiers in Computational Neuroscience*, *10*(APR). https://doi.org/10.3389/fncom.2016.00032

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

Uman, L. S. (2011). Systematic reviews and meta-analyses. *Journal of the Canadian Academy of Child and Adolescent Psychiatry*, *20*(1), 57–59. https://doi.org/https://pubmed.ncbi.nlm.nih.gov/21286370

Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, 167–176.

https://doi.org/10.1016/j.jclinepi.2015.12.005

van der Miesen, M. M., Lindquist, M. A., & Wager, T. D. (2019). Neuroimaging-based biomarkers for pain. *PAIN Reports*, *4*(4), e751. https://doi.org/10.1097/pr9.0000000000000751

van Maaren, M. C., Hueting, T. A., Völkel, V., van Hezewijk, M., Strobbe, L. J., & Siesling, S. (2023). The use and misuse of risk prediction tools for clinical decision-making. *The Breast*, *69*, 428–430. https://doi.org/10.1016/j.breast.2023.01.006

Vanneste, S., Song, J.-J., & De Ridder, D. (2018). Thalamocortical dysrhythmia detected by machine learning. *Nature Communications*, *9*(1), 1103. https://doi.org/10.1038/s41467-018-02820-0

Vansteensel, M. J., Pels, E. G. M., Bleichner, M. G., Branco, M. P., Denison, T., Freudenburg, Z. V., Gosselaar, P., Leinders, S., Ottens, T. H., Van Den Boom, M. A., Van Rijen, P. C., Aarnoutse, E. J., & Ramsey, N. F. (2016). Fully Implanted Brain–Computer Interface in a Locked-In Patient with ALS. *New England Journal of Medicine*, *375*(21), 2060–2066. https://doi.org/10.1056/NEJMoa1608085

Vargas-Lopez, O., Perez-Ramirez, C. A., Valtierra-Rodriguez, M., Yanez-Borjas, J. J., & Amezquita-Sanchez, J. P. (2021). An Explainable Machine Learning Approach Based on Statistical Indexes and SVM for Stress Detection in Automobile Drivers Using Electromyographic Signals. *Sensors*, *21*(9), 3155. https://doi.org/10.3390/s21093155

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, *7*(1), 91. https://doi.org/10.1186/1471-2105-7-91

Varoquaux, G. (2018). Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage*, *180*, 68–77. https://doi.org/10.1016/j.neuroimage.2017.06.061

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. (2010). Pharmacological treatment of chronic pain – the need for CHANGE. *Current Medical Research and Opinion*, *26*(5), 1231–1245. https://doi.org/10.1185/03007991003689175

Vatankhah, M., Asadpour, V., & Fazel-Rezai, R. (2013). Perceptual pain classification using ANFIS adapted RBF kernel support vector machine for therapeutic usage. *Applied Soft Computing*, *13*(5), 2537–2546. https://doi.org/10.1016/j.asoc.2012.11.032

Vaughn, D. A., Savjani, R. R., Cohen, M. S., & Eagleman, D. M. (2018). Empathic Neural Responses Predict Group Allegiance. *Frontiers in Human Neuroscience*, *12*. https://doi.org/10.3389/fnhum.2018.00302

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLOS Medicine*, *15*(11), e1002689. https://doi.org/10.1371/journal.pmed.1002689

Venkatasubramaniam, A., Wolfson, J., Mitchell, N., Barnes, T., JaKa, M., & French, S. (2017). Decision trees in epidemiological research. *Emerging Themes in Epidemiology*, *14*(1), 11. https://doi.org/10.1186/s12982-017-0064-4

Vijayakumar, V., Case, M., Shirinpour, S., & He, B. (2017). Quantifying and Characterizing Tonic Thermal Pain Across Subjects From EEG Data Using Random Forest Models. *IEEE Transactions on Biomedical Engineering*, *64*(12), 2988–2996. https://doi.org/10.1109/TBME.2017.2756870

Vimala, V., Ramar, K., & Ettappan, M. (2019). An Intelligent Sleep Apnea Classification System Based on EEG Signals. *Journal of Medical Systems*, *43*(2), 36. https://doi.org/10.1007/s10916-018-1146-8

Voepel-Lewis, T., Merkel, S., Tait, A. R., Trzcinka, A., & Malviya, S. (2002). The Reliability and Validity of the Face, Legs, Activity, Cry, Consolability Observational Tool as a Measure of Pain in Children with Cognitive Impairment. *Anesthesia & Analgesia*, *95*(5), 1224–1229. https://doi.org/10.1097/00000539-200211000-00020

Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., Cumbers, S., Jonas, A., McAllister, K. S. L., Myles, P., Grainger, D., Birse, M., Branson, R., Moons, K. G. M., Collins, G. S., Ioannidis, J. P. A., Holmes, C., & Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*, l6927. https://doi.org/10.1136/bmj.l6927

Voscopoulos, C., & Lema, M. (2010). When does acute pain become chronic? *British Journal of Anaesthesia*, *105*, i69–i85. https://doi.org/10.1093/bja/aeq323

Vu, M.-A. T., Adalı, T., Ba, D., Buzsáki, G., Carlson, D., Heller, K., Liston, C., Rudin, C., Sohal, V. S., Widge, A. S., Mayberg, H. S., Sapiro, G., & Dzirasa, K. (2018). A Shared Vision for Machine Learning in Neuroscience. *The Journal of Neuroscience*, *38*(7), 1601–1607. https://doi.org/10.1523/JNEUROSCI.0508-17.2018

Vuckovic, A., Gallardo, V. J. F., Jarjees, M., Fraser, M., & Purcell, M. (2018). Prediction of central neuropathic pain in spinal cord injury based on EEG classifier. *Clinical Neurophysiology*, *129*(8), 1605–1617. https://doi.org/10.1016/j.clinph.2018.04.750

Vuckovic, A., Hasan, M. A., Fraser, M., Conway, B. A., Nasseroleslami, B., & Allan, D. B. (2014). Dynamic Oscillatory Signatures of Central Neuropathic Pain in Spinal Cord Injury. *The Journal of Pain*, *15*(6), 645–655. https://doi.org/10.1016/j.jpain.2014.02.005

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of Medicine*, *368*(15), 1388–1397. https://doi.org/10.1056/NEJMoa1204471

Wandishin, M. S., & Mullen, S. J. (2009). Multiclass ROC Analysis. *Weather and Forecasting*, *24*(2), 530–547. https://doi.org/10.1175/2008WAF2222119.1

Wang, C.-H., Moreau, D., & Kao, S.-C. (2019). From the Lab to the Field: Potential Applications of Dry EEG Systems to Understand the Brain-Behavior Relationship in Sports. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00893

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, *9*(2), 187–212. https://doi.org/10.1007/s40745-020-00253-5

Wang, W., Kiik, M., Peek, N., Curcin, V., Marshall, I. J., Rudd, A. G., Wang, Y., Douiri, A., Wolfe, C. D., & Bray, B. (2020). A systematic review of machine learning models for predicting outcomes of stroke with structured data. *PLOS ONE*, *15*(6), e0234722. https://doi.org/10.1371/journal.pone.0234722

Watkinson, P., Wood, A. M., Lloyd, D. M., & Brown, G. D. A. (2013). Pain ratings reflect cognitive context: A range frequency model of pain perception. *Pain*, *154*(5), 743–749. https://doi.org/10.1016/j.pain.2013.01.016

Wei, M., Liao, Y., Liu, J., Li, L., Huang, G., Huang, J., Li, D., Xiao, L., & Zhang, Z. (2020). EEG Beta-Band Spectral Entropy Can Predict the Effect of Drug Treatment on Pain in Patients With Herpes Zoster. *Journal of Clinical Neurophysiology*, *Publish Ah*. https://doi.org/10.1097/WNP.0000000000000758

Whitmarsh, S., Nieuwenhuis, I. L. C., Barendregt, H. P., & Jensen, O. (2011). Sensorimotor Alpha Activity is Modulated in Response to the Observation of Pain in Others. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00091

Whittingstall, K., Stroink, G., Gates, L., Connolly, J., & Finley, A. (2003). Effects of dipole position, orientation and noise on the accuracy of EEG source localization. *BioMedical Engineering OnLine*, *2*(1), 14. https://doi.org/10.1186/1475-925X-2-14

Whittington, J. C. R., & Bogacz, R. (2019). Theories of Error Back-Propagation in the Brain. *Trends in Cognitive Sciences*, *23*(3), 235–250. https://doi.org/10.1016/j.tics.2018.12.005

Wiech, K., Lin, C. -s., Brodersen, K. H., Bingel, U., Ploner, M., & Tracey, I. (2010). Anterior Insula Integrates Information about Salience into Perceptual Decisions about Pain. *Journal of Neuroscience*, *30*(48), 16324–16331. https://doi.org/10.1523/JNEUROSCI.2087-10.2010

Wiech, K., Ploner, M., & Tracey, I. (2008). Neurocognitive aspects of pain perception. *Trends in Cognitive Sciences*, *12*(8), 306–313. https://doi.org/10.1016/j.tics.2008.05.005

Williamson, A., & Hoggart, B. (2005). Pain: a review of three commonly used pain rating scales. *Journal of Clinical Nursing*, *14*(7), 798–804. https://doi.org/10.1111/j.1365-2702.2005.01121.x

Willmott, C., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, 79–82. https://doi.org/10.3354/cr030079

Wilson, J. E., & Pendleton, J. M. (1989). Oligoanalgesia in the emergency department. *The American Journal of Emergency Medicine*, *7*(6), 620–623. https://doi.org/10.1016/0735-6757(89)90286-6

Winkler, J. K., Fink, C., Toberer, F., Enk, A., Deinlein, T., Hofmann-Wellenhof, R., Thomas, L., Lallas, A., Blum, A., Stolz, W., & Haenssle, H. A. (2019). Association Between Surgical Skin Markings in Dermoscopic Images and Diagnostic Performance of a Deep Learning Convolutional Neural Network for Melanoma Recognition. *JAMA Dermatology*, *155*(10), 1135. https://doi.org/10.1001/jamadermatol.2019.1735

Witt, N., Coynor, S., Edwards, C., & Bradshaw, H. (2016). A Guide to Pain Assessment and Management in the Neonate. *Current Emergency and Hospital Medicine Reports*, *4*(1), 1–10. https://doi.org/10.1007/s40138-016-0089-y

Wolfe, F. (1997). The relation between tender points and fibromyalgia symptom variables: evidence that fibromyalgia is not a discrete disorder in the clinic. *Annals of the Rheumatic Diseases*, *56*(4), 268–271. https://doi.org/10.1136/ard.56.4.268

Wolfe, F., Smythe, H. A., Yunus, M. B., Bennett, R. M., Bombardier, C., Goldenberg, D. L., Tugwell, P., Campbell, S. M., Abeles, M., Clark, P., Fam, A. G., Farber, S. J., Fiechtner, J. J., Michael Franklin, C., Gatter, R. A., Hamaty, D., Lessard, J., Lichtbroun, A. S., Masi, A. T., … Sheon, R. P. (1990). The american college of rheumatology 1990 criteria for the classification of fibromyalgia. *Arthritis & Rheumatism*, *33*(2), 160–172. https://doi.org/10.1002/art.1780330203

Wolff, R. F., Moons, K. G. M., Riley, R. D., Whiting, P. F., Westwood, M., Collins, G. S., Reitsma, J. B., Kleijnen, J., & Mallett, S. (2019). PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Annals of Internal Medicine*, *170*(1), 51. https://doi.org/10.7326/M18-1376

Wong, T.-T. (2015). Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, *48*(9), 2839–2846. https://doi.org/10.1016/j.patcog.2015.03.009

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, *20*(3), 365–377. https://doi.org/10.1038/nn.4478

Woodman, G. F. (2010). A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception & Psychophysics*, *72*(8), 2031–2046. https://doi.org/10.3758/APP.72.8.2031

Woolf, C. J. (2010). What is this thing called pain? *Journal of Clinical Investigation*, *120*(11), 3742–3744. https://doi.org/10.1172/JCI45178

Woolf, C. J. (2011). Central sensitization: Implications for the diagnosis and treatment of pain. *Pain*, *152*(Supplement), S2–S15. https://doi.org/10.1016/j.pain.2010.09.030

Wooten, M., Weng, H.-J., Hartke, T. V., Borzan, J., Klein, A. H., Turnquist, B., Dong, X., Meyer, R. A., & Ringkamp, M. (2014). Three functionally distinct classes of C-fibre nociceptors in primates. *Nature Communications*, *5*(1), 4122. https://doi.org/10.1038/ncomms5122

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The Increasing Dominance of Teams in Production of Knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Wydenkeller, S., Maurizio, S., Dietz, V., & Halder, P. (2009). Neuropathic pain in spinal cord injury: significance of clinical and electrophysiological measures. *European Journal of Neuroscience*, *30*(1), 91–99. https://doi.org/10.1111/j.1460-9568.2009.06801.x

Xu, A., Larsen, B., Baller, E. B., Scott, J. C., Sharma, V., Adebimpe, A., Basbaum, A. I., Dworkin, R. H., Edwards, R. R., Woolf, C. J., Eickhoff, S. B., Eickhoff, C. R., & Satterthwaite, T. D. (2020). Convergent neural representations of experimentally-induced acute pain in healthy volunteers: A large-scale fMRI meta-analysis. *Neuroscience & Biobehavioral Reviews*, *112*, 300–323. https://doi.org/10.1016/j.neubiorev.2020.01.004

Xu, X., & Huang, Y. (2020). Objective Pain Assessment: a Key for the Management of Chronic Pain. *F1000Research*, *9*, 35. https://doi.org/10.12688/f1000research.20441.1

Yam, M., Loh, Y., Tan, C., Khadijah Adam, S., Abdul Manan, N., & Basir, R. (2018). General

Pathways of Pain Sensation and the Major Neurotransmitters Involved in Pain Regulation. *International Journal of Molecular Sciences*, *19*(8), 2164. https://doi.org/10.3390/ijms19082164

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2018). Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, *9*(4), 611–629. https://doi.org/10.1007/s13244-018-0639-9

Yang, F., Banerjee, T., Narine, K., & Shah, N. (2018). Improving pain management in patients with sickle cell disease from physiological measures using machine learning techniques. *Smart Health*, *7–8*, 48–59. https://doi.org/10.1016/j.smhl.2018.01.002

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yao, D. (2017). Is the Surface Potential Integral of a Dipole in a Volume Conductor Always Zero? A Cloud Over the Average Reference of EEG and ERP. *Brain Topography*, *30*(2), 161–171. https://doi.org/10.1007/s10548-016-0543-x

Yao, D., Qin, Y., Hu, S., Dong, L., Bringas Vega, M. L., & Valdés Sosa, P. A. (2019). Which Reference Should We Use for EEG and ERP practice? *Brain Topography*, *32*(4), 530–549. https://doi.org/10.1007/s10548-019-00707-x

Yasoda, K., Ponmagal, R. S., Bhuvaneshwari, K. S., & Venkatachalam, K. (2020). Automatic detection and classification of EEG artifacts using fuzzy kernel SVM and wavelet ICA (WICA). *Soft Computing*, *24*(21), 16011–16019. https://doi.org/10.1007/s00500-020-04920-w

Yavandhasani, M., & Ghaderi, F. (2022). Visual Object Recognition From Single-Trial EEG Signals Using Machine Learning Wrapper Techniques. *IEEE Transactions on Biomedical*

Engineering, *69*(7), 2176–2183. https://doi.org/10.1109/TBME.2021.3138157

Yen, C.-T., & Lu, P.-L. (2013). Thalamus and pain. *Acta Anaesthesiologica Taiwanica*, *51*(2), 73–80. https://doi.org/10.1016/j.aat.2013.06.011

Yoshinaga, H., Nakahori, T., Ohtsuka, Y., Oka, E., Kitamura, Y., Kiriyama, H., Kinugasa, K., Miyamoto, K., & Hoshida, T. (2002). Benefit of Simultaneous Recording of EEG and MEG in Dipole Localization. *Epilepsia*, *43*(8), 924–928. https://doi.org/10.1046/j.1528-1157.2002.42901.x

Younger, J., McCue, R., & Mackey, S. (2009). Pain outcomes: A brief review of instruments and techniques. *Current Pain and Headache Reports*, *13*(1), 39–43. https://doi.org/10.1007/s11916-009-0009-x

Yu, F. H., & Catterall, W. A. (2003). Overview of the voltage-gated sodium channel family. *Genome Biology*, *4*(207), 1–7. https://doi.org/10.1186/gb-2003-4-3-207

Yu, M., Sun, Y., Zhu, B., Zhu, L., Lin, Y., Tang, X., Guo, Y., Sun, G., & Dong, M. (2020). Diverse frequency band-based convolutional neural networks for tonic cold pain assessment using EEG. *Neurocomputing*, *378*, 270–282. https://doi.org/10.1016/j.neucom.2019.10.023

Yu, M., Yan, H., Han, J., Lin, Y., Zhu, L., Tang, X., Sun, G., He, Y., & Guo, Y. (2020). EEG-based tonic cold pain assessment using extreme learning machine. *Intelligent Data Analysis*, *24*(1), 163–182. https://doi.org/10.3233/IDA-184388

Yusuf, M., Atal, I., Li, J., Smith, P., Ravaud, P., Fergie, M., Callaghan, M., & Selfe, J. (2020). Reporting quality of studies using machine learning models for medical diagnosis: a systematic review. *BMJ Open*, *10*(3), e034568. https://doi.org/10.1136/bmjopen-2019-034568

Yuvaraj, R., Murugappan, M., Mohamed Ibrahim, N., Iqbal, M., Sundaraj, K., Mohamad, K., Palaniappan, R., Mesquita, E., & Satiyan, M. (2014). On the analysis of EEG power, frequency and asymmetry in Parkinson's disease during emotion processing. *Behavioral and Brain Functions*, *10*(1), 12. https://doi.org/10.1186/1744-9081-10-12

Zander, T. O., Lehne, M., Ihme, K., Jatzev, S., Correia, J., Kothe, C., Picht, B., & Nijboer, F. (2011). A Dry EEG-System for Scientific Research and Brain–Computer Interfaces. *Frontiers in Neuroscience*, *5*. https://doi.org/10.3389/fnins.2011.00053

Zanocchi, M., Maero, B., Nicola, E., Martinelli, E., Luppino, A., Gonella, M., Gariglio, F., Fissore, L., Bardelli, B., Obialero, R., & Molaschi, M. (2008). Chronic pain in a sample of nursing home residents: Prevalence, characteristics, influence on quality of life (QoL). *Archives of Gerontology and Geriatrics*, *47*(1), 121–128. https://doi.org/10.1016/j.archger.2007.07.003

Zhang, Z. G., Hu, L., Hung, Y. S., Mouraux, A., & Iannetti, G. D. (2012). Gamma-Band Oscillations in the Primary Somatosensory Cortex--A Direct and Obligatory Correlate of Subjective Pain Intensity. *Journal of Neuroscience*, *32*(22), 7429–7438. https://doi.org/10.1523/JNEUROSCI.5877-11.2012

Zheng, X., Chen, W., You, Y., Jiang, Y., Li, M., & Zhang, T. (2020). Ensemble deep learning for automated visual classification using EEG signals. *Pattern Recognition*, *102*, 107147. https://doi.org/10.1016/j.patcog.2019.107147

Zhou, F., Li, J., Zhao, W., Xu, L., Zheng, X., Fu, M., Yao, S., Kendrick, K. M., Wager, T. D., & Becker, B. (2020). Empathic pain evoked by sensory and emotional-communicative cues share common and process-specific neural representations. *ELife*, *9*. https://doi.org/10.7554/eLife.56929

Zis, P., Liampas, A., Artemiadis, A., Tsalamandris, G., Neophytou, P., Unwin, Z., Kimiskidis, V. K., Hadjigeorgiou, G. M., Varrassi, G., Zhao, Y., & Sarrigiannis, P. G. (2022). EEG

Recordings as Biomarkers of Pain Perception: Where Do We Stand and Where to Go?

Pain and Therapy, *11*(2), 369–380. https://doi.org/10.1007/s40122-022-00372-2

Appendices

Supplementary Material 1

Search Strategy

MEDLINE

1. exp Electroencephalography/
2. "EEG*".af.
3. electroencephalo*.af.
4. or/1-3
5. exp Pain/
6. exp Pain Perception/
7. pain*.af.
8. exp Nociception/
9. nocicept*.af.
10. or/5-9
11. exp Machine Learning/
12. machine learning.af.
13. supervised.af.
14. exp Multivariate Analysis/
15. exp Support Vector Machine/
16. multivariate*.af.
17. support vector machine*.af.
18. SVM.af.
19. exp Decision Trees/
20. decision tree*.af.
21. random Forest.af.
22. nearest neighbor.af.
23. nearest neighbour.af.
24. Naive bayes.af.
25. exp Bayes Theorem/
26. exp Regression Analysis/
27. regression.af.
28. exp Neural Networks, Computer/
29. Models, Neurological/
30. Neural Net*.af.
31. exp DISCRIMINANT ANALYSIS/
32. linear discriminant analysis.af.
33. exp ARTIFICIAL INTELLIGENCE/
34. "artificial intelligence".af.
35. or/11-34
36. 4 and 10 and 35
37. limit 36 to english language

Cochrane

#1 MeSH descriptor: [Electroencephalography] explode all trees
#2 ("EEG*"):ti,ab,kw
#3 (electroencephalo*):ti,ab,kw
#4 {OR #1-#3}
#5 MeSH descriptor: [Pain] explode all trees
#6 MeSH descriptor: [Pain Perception] explode all trees
#7 (pain*):ti,ab,kw
#8 MeSH descriptor: [Nociception] explode all trees
#9 (nocicept*):ti,ab,kw
#10 {OR #5-#9}
#11 #4 AND #10
#12 MeSH descriptor: [Machine Learning] explode all trees
#13 ("machine learning"):ti,ab,kw
#14 (supervised):ti,ab,kw
#15 MeSH descriptor: [Multivariate Analysis] explode all trees
#16 MeSH descriptor: [Support Vector Machine] explode all trees
#17 (multivariate*):ti,ab,kw
#18 (support vector machine*):ti,ab,kw
#19 (SVM):ti,ab,kw
#20 MeSH descriptor: [Decision Trees] explode all trees
#21 (decision NEXT tree*):ti,ab,kw
#22 ("random forest"):ti,ab,kw
#23 ("nearest neighbor"):ti,ab,kw
#24 ("nearest neighbour"):ti,ab,kw
#25 ("Naive bayes"):ti,ab,kw
#26 MeSH descriptor: [Regression Analysis] explode all trees
#27 (regression):ti,ab,kw
#28 MeSH descriptor: [Neural Networks, Computer] explode all trees
#29 MeSH descriptor: [Models, Neurological] explode all trees
#30 (Neural NEXT Net*):ti,ab,kw
#31 MeSH descriptor: [Discriminant Analysis] explode all trees
#32 ("linear discriminant analysis"):ti,ab,kw
#33 MeSH descriptor: [Artificial Intelligence] explode all trees
#34 "artificial intelligence":ti,ab,kw
#35 {OR #12-#34}
#36 #11 AND #35

EMBASE

1. exp electroencephalography/
2. "EEG*".af.
3. electroencephalo*.af.
4. 1 or 2 or 3
5. exp pain/

6. exp nociception/
7. pain.af.
8. nocicept*.af.
9. 5 or 6 or 7 or 8
10. 4 and 9
11. exp machine learning/
12. machine learning.af.
13. supervised.af.
14. exp multivariate analysis/
15. exp support vector machine/
16. multivariate*.af.
17. support vector machine*.af.
18. SVM.af.
19. exp "decision tree"/
20. decision tree*.af.
21. random Forest.af.
22. nearest neighbor.af.
23. nearest neighbour.af.
24. exp random forest/
25. exp Bayesian learning/
26. Naive bayes.af.
27. exp regression analysis/
28. regression.af.
29. exp artificial neural network/
30. Neural Net*.af.
31. exp discriminant analysis/
32. linear discriminant analysis.af.
33. exp artificial intelligence/
34. "artificial intelligence".af.
35. or/11-34
36. 10 and 35
37. limit 36 to english language

PsycINFO

S1 DE "Electroencephalography" OR DE "Alpha Rhythm" OR DE "Beta Rhythm" OR DE "Delta Rhythm" OR DE "Gamma Rhythm" OR DE "Theta Rhythm"
S2 "EEG*"
S3 electroencephalo*
S4 S1 OR S2 OR S3
S5 DE "Pain" OR DE "Aphagia" OR DE "Back Pain" OR DE "Chronic Pain" OR DE "Headache" OR DE "Myofascial Pain" OR DE "Neuralgia" OR DE "Neuropathic Pain" OR DE "Somatoform Pain Disorder"
S6 pain*
S7 DE "Pain Perception" OR DE "Analgesia" OR DE "Pain Thresholds"
S8 nocicept*

S9 S5 OR S6 OR S7 OR S8
S10 S4 AND S9
S11 DE "Machine Learning" OR DE "Computational Reinforcement Learning" OR DE "Inductive Logic Programming" OR DE "Machine Learning Algorithms" OR DE "Pattern Recognition (Computer Science)"
S12 "machine learning"
S13 supervised
S14 DE "Multivariate Analysis" OR DE "Factor Analysis" OR DE "Mixture Modeling" OR DE "Multiple Regression" OR DE "Path Analysis" OR DE "Principal Component Analysis"
S15 multivariate*
S16 "support vector machine*"
S17 SVM
S18 "decision tree*"
S19 "random Forest"
S20 "nearest neighbor"
S21 "nearest neighbour"
S22 DE "Bayesian Analysis"
S23 "Naive bayes"
S24 DE "Statistical Regression" OR DE "Linear Regression" OR DE "Logistic Regression" OR DE "Multiple Regression" OR DE "Nonlinear Regression"
S25 regression
S26 DE "Neural Networks" OR DE "Artificial Neural Networks" OR DE "Biological Neural Networks"
S27 "linear discriminant analysis" OR “artificial intelligence”
S28 S11 OR S12 OR S13 OR S14 OR S15 OR S16 OR S17 OR S18 OR S19 OR S20 OR S21 OR S22 OR S23 OR S24 OR S25 OR S26 OR S27
S29 S10 AND S28 Limit to English

Web of Science

TS=(electroencephalo* OR eeg) AND TS=(pain* OR nocicept*) AND TS=("machine learning" OR supervised OR multivariate* OR "support vector machine*" OR SVM OR "decision tree*" OR "random Forest" OR "nearest neighbor" OR "nearest neighbour" OR "Naive bayes" OR regression OR "linear discriminant analysis" OR "artificial intelligence")
Refined by: LANGUAGES: (ENGLISH)
Indexes=SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, BKCI-S, BKCI-SSH, ESCI, CCR-EXPANDED, IC Timespan=All years

Risk of Bias

Table S1. ROB assessment for all studies

	Participants	Predictors	Outcomes	Analysis	Overall
Intensity					
Alazrai, AL-Rawi, et al. (2019)	Low	Low	Low	High	High
Alazrai, Momani, et al. (2019)	Low	Low	Low	High	High
Bai et al. (2016)	Low	Low	Low	High	High
Cao et al. (2020)	High	Low	Low	High	High
Elsayed et al. (2020)	Low	Low	Low	High	High
Furman et al. (2018)	Low	Low	Low	High	High
Hadjileontiadis (2015)	High	Low	Unclear	High	High
Kaur et al. (2019)	Unclear	Low	Low	High	High
Kimura et al. (2021)	Low	Low	Low	High	High
Li et al. (2018)	Low	Low	Low	High	High
Misra et al. (2017)	Low	Low	Low	High	High
Nezam et al. (2021)	Low	Low	Low	High	High
Okolo & Omurtag (2018)	High	Low	Low	High	High
Prichep et al. (2018)	Low	Low	Low	High	High
Sai et al. (2019)	High	Low	High	High	High
Schulz et al. (2012)	Unclear	Low	Low	High	High
Tripanpitak et al. (2020)	High	Low	Low	High	High
Tu et al. (2016)	Unclear	Low	Low	High	High
Vatankhah et al. (2013)	High	Low	Low	High	High
Vijayakumar et al. (2017)	Low	Low	Low	Low	Low
Yu, Sun, et al. (2020)	Low	Low	Low	Low	Low
Yu, Yan, et al. (2020)	High	Low	Low	High	High
Phenotyping					
Akben et al. (2012)	Low	Low	Low	High	High
Akben et al. (2016)	Low	Low	Low	High	High
Cao et al. (2018)	Low	Low	Low	High	High
De Tommaso et al. (1999)	High	Low	Low	High	High
Frid et al. (2020)	High	Low	Low	High	High
Graversen et al. (2011)	High	Low	Low	High	High
Levitt et al. (2020)	Low	Low	Low	High	High
Paul et al. (2019)	Low	Low	Low	High	High
Saif et al. (2021)	Unclear	Low	Low	High	High
Sarnthein et al. (2006)	Low	Low	Low	High	High
Subasi et al. (2019)	Low	Low	Low	High	High
Ta Dinh et al. (2019)	Low	Low	Low	High	High
Vanneste et al. (2018)	Low	Low	Low	High	High

Vuckovic et al. (2018)	Low	Low	Low	High	High
Wydenkeller et al. (2009)	Unclear	Low	High	High	High
Treatment					
Gram et al. (2015)	Low	Low	High	High	High
Gram et al. (2017)	Low	Low	Low	High	High
Graversen et al. (2012)	Low	Low	Low	High	High
Graversen et al. (2015)	High	Low	Low	High	High
Grosen et al. (2017)	Low	Low	Low	High	High
Hunter et al. (2009)	High	Low	Low	High	High
Wei et al. (2020)	Low	Low	Low	High	High

Applicability Assessment

Table S2. Applicability assessment for all studies

	Participants	Predictors	Outcomes	Overall
Intensity				
Alazrai, AL-Rawi, et al. (2019)	Low	Low	Low	Low
Alazrai, Momani, et al. (2019)	Low	Low	Low	Low
Bai et al. (2016)	Low	Low	Low	Low
Cao et al. (2020)	Low	Low	Low	Low
Elsayed et al. (2020)	Low	Low	Low	Low
Furman et al. (2018)	Low	Low	Low	Low
Hadjileontiadis (2015)	Low	Low	Low	Low
Kaur et al. (2019)	Low	Low	Low	Low
Kimura et al. (2021)	Low	Low	Low	Low
Li et al. (2018)	Low	Low	Low	Low
Misra et al. (2017)	Low	Low	Low	Low
Nezam et al. (2021)	Low	Low	Low	Low
Okolo & Omurtag (2018)	Low	Low	High	High
Prichep et al. (2018)	Low	Low	Low	Low
Sai et al. (2019)	Low	Low	Low	Low
Schulz et al. (2012)	Low	Low	Low	Low
Tripanpitak et al. (2020)	Low	Low	Low	Low
Tu et al. (2016)	Low	Low	Low	Low
Vatankhah et al. (2013)	Low	Low	Low	Low
Vijayakumar et al. (2017)	Low	Low	Low	Low
Yu, Sun, et al. (2020)	Low	Low	Low	Low
Yu, Yan, et al. (2020)	Low	Low	Low	Low
Phenotyping				
Akben et al. (2012)	Low	Low	Low	Low
Akben et al. (2016)	Low	Low	Low	Low
Cao et al. (2018)	Low	Low	Low	Low
De Tommaso et al. (1999)	Low	Low	Low	Low
Frid et al. (2020)	Low	Low	Low	Low
Graversen et al. (2011)	Low	Low	Low	Low
Levitt et al. (2020)	Low	Low	Low	Low
Paul et al. (2019)	Low	Low	Low	Low
Saif et al. (2021)	Low	Low	Low	Low
Sarnthein et al. (2006)	Low	Low	Low	Low
Subasi et al. (2019)	Low	Low	Low	Low
Ta Dinh et al. (2019)	Low	Low	Low	Low
Vanneste et al. (2018)	Low	Low	Low	Low

Vuckovic et al. (2018)	Low	Low	Low	Low
Wydenkeller et al. (2009)	Low	Low	Low	Low
Treatment				
Gram et al. (2015)	Low	Low	Low	Low
Gram et al. (2017)	Low	Low	Low	Low
Graversen et al. (2012)	Low	Low	Low	Low
Graversen et al. (2015)	Low	Low	Low	Low
Grosen et al. (2017)	Low	Low	Low	Low
Hunter et al. (2009)	Low	Low	Low	Low
Wei et al. (2020)	Low	Low	Low	Low

Supplementary Material 2

Methods

Model Evaluation

The primary measures of discrimination in the current study were the AUC and accuracy. The AUC measures the model's overall performance, which is the ability of the algorithm to correctly discriminate between low and high pain across different classification thresholds. An AUC of 0.5 represents chance discrimination, whilst an AUC of 1 represents perfect discrimination. Moreover, accuracy assesses the overall effectiveness of the algorithm and represents the number of correctly classified events over the total number of events. Precision measures the ratio of correctly labelled positive events across all positive predictions. In contrast, recall assesses the ratio of true positive cases correctly identified. F1 represents the harmonic mean of recall and precision. For accuracy, precision, recall and F1, outputs of 1 demonstrate perfect predictions, whilst 0.5 represents chance performance for binary classification (note, classification metrics, excluding AUC, are reported as a percentage in-text for improved readability). Finally, the Brier score measures the mean squared error of the probability prediction. Here, 0 represents perfect performance and 1 reflects the worst theoretical performance. The Brier score was assessed as it is affected by discrimination and calibration, which is advantageous over other metrics. Equations (1) to (5) provide mathematical descriptions of the metrics.

$$Accuracy = \frac{tp + tn}{tp + fp + tn + fn} \tag{1}$$

$$Precision = \frac{tp}{tp + fp} \tag{2}$$

$$Recall = \frac{tp}{tp + fn} \tag{3}$$

$$F1 = \frac{2tp}{2tp + fp + fn} \tag{4}$$

Where tp, tn, fp, fn represent the number of true positives, true negatives, false positives, and false negatives, respectively.

$$Brier\ Score = \frac{1}{n}\sum_{i=1}^{n}(p_i - \ o_i)^2 \tag{5}$$

Where n is the number of samples, p_i is the probability prediction and o_i is the outcome event.

Confusion Matrices

The confusion matrices for all models and external validation assessments are presented in Table S3.

Table S3. *Confusion matrices for all models for both external validations.*

Model	**External Validation One**		**External Validation Two**	
	Predicted Low Pain	Predicted High Pain	Predicted Low Pain	Predicted High Pain
AdaBoost				
Low Pain	311	192	272	232
High Pain	172	332	212	292
LDA				
Low Pain	346	157	335	169
High Pain	245	259	272	232
LR				
Low Pain	324	179	320	184
High Pain	234	270	272	232
NB				
Low Pain	355	148	325	179
High Pain	215	289	223	281
RF				
Low Pain	426	77	367	137
High Pain	242	262	262	242
SVM				
Low Pain	301	202	281	223
High Pain	182	322	216	288
XGBoost				
Low Pain	320	183	292	212
High Pain	195	309	216	288