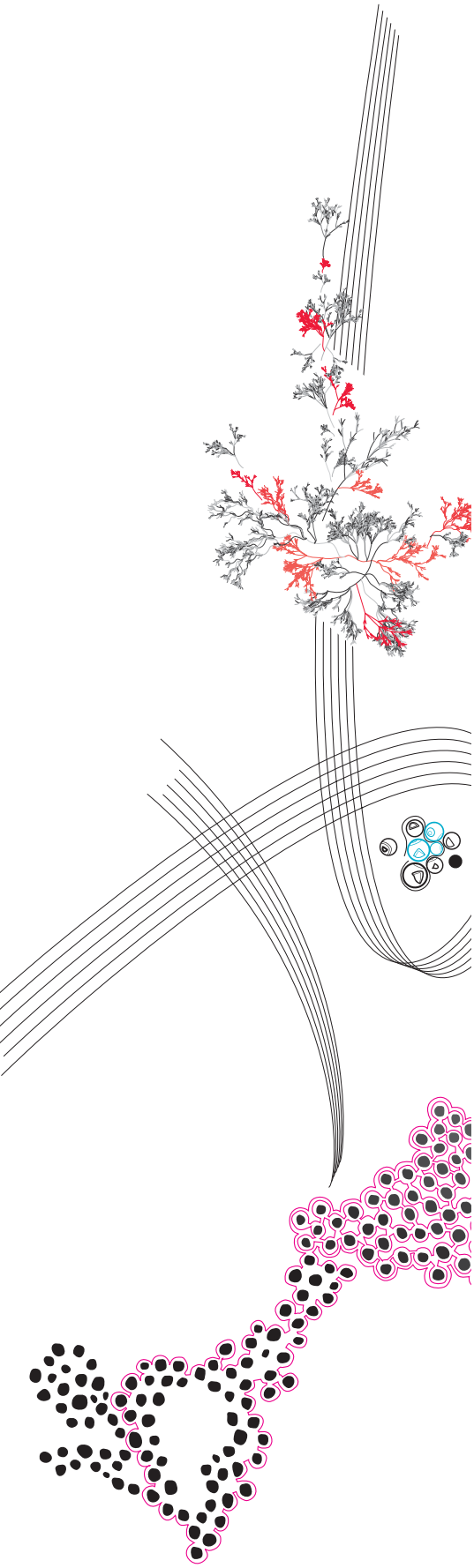


MSc. Thesis for Biomedical Engineering

Feasibility of channel reduction in real-time estimation of perceptual nociceptive threshold based on EEG measurements.

Marcela Martínez Pinedo

Supervisors:
Jan Buitenweg
Frodo Muijzer
Mark Vlutters

August, 2024

Biological Systems and Signals group
Faculty of Science and Technology,
Biomedical Engineering

UNIVERSITY OF TWENTE.

Abstract

Currently, chronic pain treatment is hindered by a lack of understanding and subjective measurement tools. To address this, a method at the University of Twente seeks to evaluate perceptual thresholds through intra-epidermal electrical stimulation, selectively targeting nociceptive fibers. The experimental procedures are based on psychophysical concepts, specifically the Go/No-Go (GN) and two-interval forced choice (2IFC) tasks. In the GN procedure, participants receive stimuli of varying amplitudes and press a button to record their perception. This feedback helps control the amplitude of subsequent stimuli, ensuring stimulation around threshold levels. Conversely, the 2IFC task involves presenting two intervals, with stimulation occurring in only one. The subject must identify the interval containing the stimulation.

The GN procedure is typically used but poses challenges as participants may lose focus, fail to report perceived stimuli, or be unable to communicate their perception. To overcome these issues, automatic classification of perceptual thresholds is proposed. A previous study successfully achieved reliable threshold estimates by using a convolutional neural network to perform the 2IFC task by classifing EEG activity recorded from a 32-electrode setup. However, this setup is time-consuming and uncomfortable for patients, limiting its clinical application. This study aims to explore the feasibility of real-time nociceptive threshold tracking using eight EEG-electrodes located at C3, Cz, C4, T7, T8, F3, Fz, and F4; which corresponds to the area where the largest changes in electrical activity are seen.

First, improving the performance of the previously used network was explored by comparing different architectures and performance-enhancing strategies. Various architectures from the literature were tested using data obtained from subsampling eight electrodes out of a 32-electrode dataset. Three strategies were employed to obtain the best-performing classifier: hyperparameter optimization algorithms (Bayesian optimization - Hyperband and Bayesian optimization), transfer learning, and an ensemble model. The best model resulted from the use of the EEG-Inception architecture with hyperparameter tuning, the evaluation of its performance in a separate testing set resulted in a 6% increase in accuracy.

Simulations were conducted to optimize parameters related to the GN and 2IFC tasks and to determine the minimum neural network accuracy required for reliable threshold estimates. These simulations explored the parameters that resulted in the greatest agreement between GN and 2IFC tasks and examined the effects of varying accuracies. It was determined that a neural network accuracy of 70% is the theoretical minimum for reliable threshold estimates. Experimental validation involved 15 healthy participants, each performing two tasks in random order: Task 1 involved tracking the perceptual threshold using the GN procedure while the neural network independently performed the 2IFC task; Task 2 involved tracking two independent thresholds using the neural network alone, without concurrent physical tasks.

In Task 1, despite no significant difference in the calculated thresholds, there was high variability (as shown by limits of agreement in a Bland-Altman plot) and low correlation (ICC=0.06) between the GN and 2IFC tasks. Additionally, amplitudes from the evoked potentials (EPs) were lower than expected, with no significant difference between average responses marked as correct by the network and those marked as incorrect in the 2IFC task. This suggests the presence of many false positives and poor classification ability. In Task 2, even greater variability and worse correlation (ICC=-0.009) between estimated thresholds were observed. The presence of clear EPs in the average responses when the network classified the EPs as incorrect suggests that the classifier may depend on the concurrent button press.

Although reliable real-time nociceptive threshold tracking was not achieved, this study provides valuable insights for future research. Key areas for improvement include electrode selection, the influence of button presses on EPs during the GN procedure, the potential impact of additional electrodes, and the use of different EEG recording references.

Contents

List of Acronyms

1 Introduction

Pain is a vital but subjective experience that depends on the dynamic integration of sensory and contextual processes, including biological, psychological, and social factors. It is defined by the International Association for the Study of Pain (IASP) as "an unpleasant sensory and emotional experience associated with, or resembling that associated with, actual or potential tissue damage" [1, 2, 3]. Pain is one of the leading causes of disability and, if left untreated, can lead to undesirable personal and social outcomes, such as depression, work absence, and increased costs for families and caregivers. Inaccuracies in pain assessment can result in suboptimal or inadequate treatment, potentially causing further clinical complications [4]. Chronic pain often arises from progressively disturbed processes in the central nervous system, which are poorly recognized in patients and insufficiently targeted by conventional pain management protocols.

Pain assessment is crucial for chronic pain classification and treatment guidance. However, current measurement options are limited, often relying on self-reports such as the Visual Analog Scale (VAS) and Numeric Rating Scales. These assessments may not be feasible for many vulnerable populations, including non-communicative patients, those with speech disorders, and individuals with impaired consciousness. Even with patients with none of the mentioned limitations, self-report techniques can lead to miscommunication or misjudgment [4].

Due to these limitations, alternative approaches, such as the evaluation of perceptual thresholds, have been explored. Perceptual threshold evaluation plays a crucial role in assessing various sensory systems, including vision, hearing, and touch, and can also be applied to nociception. In pain research, evaluating perceptual thresholds can help to assess altered nociceptive processing and sensitivity of ascending pathways in the nociceptive system [5]. These measured thresholds are of particular interest since it has been shown that events such as disease, clinical interventions, and experimental conditioning stimuli can activate endogenous mechanisms that lead to an increase in nociceptive thresholds [6].

Intra-epidermal electrical stimulation has been used to selectively stimulate nociceptive fibers with currents lower than twice the detection thresholds [7]. Recent experiments at the University of Twente utilized a needle electrode on the right hand to administer electrical stimuli, generated by a custom-built stimulator (NociTRACK AmbuStim) with a randomized interstimulus interval of 3.5-4.5 seconds during this threshold tracking experiments. In addition to selective stimulation, participants must remain focused and consistently respond to the stimulation to accurately assess perceptual thresholds. However, perception reports are not always accurate. Participants may lose focus during long or boring procedures, fail to report perceived stimuli (lapsing), or report stimulus perception without actual perception (guessing). Additionally, some subjects might have disabilities that hinder perception report [5].

Measurement tools used to obtain more objectives markers of pain include Electroenephalography (EEG), which due to its high temporal resolution, clinical convenience, non-invasiveness, and low cost, has been used to study pain-related cortical activity [8, 3]. Nociceptive activity can be quantified by recording and analyzing EEG signals time-locked to the stimuli, this averaged time-locked signal, known as the Evoked potential (EP), describes transient synchronized activity of large neural networks within the cortex.

In [5], researchers explored a fully automated method that used nociceptive-induced EPs to produce precise and objective assessment of perceptual thresholds across various patient populations. A deep learning classifier built upon a 32-channel EEG set-up was used to reliably analyze brain activity in response to nociceptive stimuli. The neural network successfully provided accurate estimates of the perceptual threshold, highlighting the potential of deep learning classification to control the adaptive stimulus sequence used in threshold tracking procedures. This approach could enhance accuracy and objectivity by eliminating the need for subject responses and minimizing the effects of poor task performance.

The next step in developing this method is reducing the number of EEG electrodes required. Fewer electrodes

would make the setup process quicker, making studies and clinical applications more time-efficient and scalable. Therefore, this study aims to explore the feasibility of real-time nociceptive threshold tracking using eight electrodes (C3, Cz, C4, T6, T8, F3, Fz, and F4) chosen based on EPs observed with a 32-electrode setup using a 10-20 montage.

To conduct this study, the background information is first presentedN. The methods section addresses the application of tools to enhance machine learning performance and details the experiments conducted to test the obtained classifier. Subsequently, the results and their analysis are presented, followed by a discussion on the study's limitations and potential future work.

2 Background

2.1 Nociceptive system

The perception of pain, called nociception, depends on specifically dedicated receptors and pathways. Since alerting the brain to the dangers implied by noxious stimuli differs substantially from informing it about innocuous sensory stimuli, the existence of a specialized subsystem for the perception of potentially threatening circumstances is necessary. The importance of pain in clinical practice, as well as the many aspects of it that remain poorly understood, continue to make nociception an extremely active area of research. [9]

2.1.1 Pain

Pain is a complex and subjective experience that signals impending or actual tissue damage, allowing individuals to avoid further injury. Similarly, it can also prevent harmful movement, as seen in injuries, where reduced mobility contributes to healing [10]. However, pain can also be pathological, for instance, in the case of nerve misfiring or damage, and its prolongation can lead to adverse outcomes [11]. The challenges of treating pain often stem from a limited understanding of the nervous system mechanisms underlying the pain experience [12].

Until the 1960s, pain was considered an inevitable sensory response to tissue damage. Relevant factors such as genetic differences, past experience, anxiety, or expectations were largely disregarded [13]. In 1965, Melzack and Wall proposed the gate control theory, suggesting the existence of endogenous modulatory mechanisms. This theory contrasted with the previous belief that pain was merely transmitted by the Peripheral Nervous System (PNS) to the Central Nervous System (CNS). According to the theory, pain transmission is modulated by a gating mechanism in the dorsal horn fibers can close (inhibit) or open (facilitate) the pain gate [14, 15].

Later, in 1999, the neuromatrix theory expanded on the understanding of pain by proposing that pain modulation occurs not only at the spinal level but also involves cerebral mechanisms of pain processing and transmission [15, 16]. Studies have corroborated the role of cognitive and limbic systems in pain processing, including patients' beliefs, understanding of their pain syndrome, pain sensitivity, and catastrophic thinking. Today, the IASP defines pain as "an unpleasant sensory and emotional experience associated with actual or potential tissue damage, or described in terms of such damage" [4, 2]

Pain perception is often termed nociception, which specifically refers to the detection of noxious stimuli, such as tissue damage or extreme temperatures. These stimuli activate nociceptors and their pathways through CNS and PNS processing. This perception depends on factors such as action potential frequency and the time interval between each action potential [11]. However, pain does not necessarily imply nociception. While pain is frequently triggered by noxious stimuli, it can also result from lesions in the nervous system, as seen pathologies such as diabetes mellitus, multiple sclerosis, stroke or post-amputation. In these cases, the intensity of pain often has little or no relation to the degree of tissue injury or other pathology [13].

2.1.2 Nociceptors

Nociceptors are relatively unspecialized nerve cell endings that typically initiate the sensation of pain. These receptors originate from cell bodies in the dorsal root ganglia or the trigeminal ganglion, sending one axonal process to the periphery and another into the spinal cord or brainstem [11]. They are located mainly in the epidermis, in contrast to non-nociceptive fibers that terminate more deeply, in the dermis [7]. Peripheral nociceptive axons terminate in unspecialized "free endings," which convey pain information. In the skin, nociceptors are classified into three major types: Aδ mechanosensitive nociceptors, Aδ mechanothermal nociceptors, and polymodal nociceptors associated with C fibers. Aδ fibers are myelinated, have small receptive fields, and conduct signals at about 20 m/s, while C fibers are small-diameter, unmyelinated axons

with large receptive fields, conducting at velocities less than 2 m/s. These differences result in both fast and slow pain pathways, contrasting with the faster conduction of somatic sensory receptors associated with mechanical stimuli. The receptive fields of pain-sensitive neurons are relatively large, especially at the thalamus and cortex levels, emphasizing the importance of detecting pain over a precise localization[9, 11].

Pain perception is generally categorized into two types: a sharp first pain and a longer-lasting, diffuse sensation called second pain. Activation of Aδ fibers is responsible for the initial tingling sensation or sharp pain, alerting the body to the presence of a noxious stimulus. In contrast, C fibers mediate the "second pain," which occurs at higher stimulus intensities and results in a duller, longer-lasting pain sensation associated with the intensity of the pain [9, 11].

Nociceptors relay information about noxious stimuli from various body parts, including the skin, joints, viscera, and muscles. They are activated by a wide variety of chemical substances such as globulin, protein kinases, arachidonic acid, histamine, nerve growth factor, substance P, calcitonin gene-related peptide, potassium, serotonin, acetylcholine, low-pH solutions, adenosine triphosphate, and lactic acid. Additionally, nociceptors respond to temperature extremes, high pressures, and tissue damage causing inflammation. Based on the type of stimuli they detect, nociceptors can be further subdivided into high-threshold mechanoreceptors for intense mechanical stimulation, thermal receptors for thermal and mechanical stimulation, chemical receptors, and polymodal receptors for high-intensity mechanical, thermal, and chemical stimulation [11].

2.1.3 Nociceptive pathway

Once a nociceptor is activated, the action potential is transmitted through primary afferent nociceptors, which terminate near second-order nerve cells in the dorsal horn of the gray matter. Synaptic connections are made in laminae I and/or II. The second-order cells then cross the midline and project to the brainstem and thalamus via the anterolateral (ventrolateral) quadrant of the contralateral side of the spinal cord. These fibers form the spinothalamic tract, the major ascending pathway for pain and temperature information [9, 17]

The spinothalamic tract, located in the white matter of the spinal cord, consists of two parts: the lateral spinothalamic tract and the anterior spinothalamic tract. The lateral spinothalamic tract primarily transmits pain and temperature sensations, while the anterior spinothalamic tract carries information related to touch and firm pressure sensations towards the thalamus [18].

The principal target of these pathways is the ventral posterior nucleus of the thalamus, which sends axons to the primary and secondary somatosensory cortices. The nociceptive information transmitted to these cortical areas is responsible for the discriminative component of pain, such as identifying the location, intensity, and quality of the stimulus. Electrophysiological recordings from nociceptive neurons in the primary somatosensory cortex show that these neurons have small receptive fields that relate to pain localization. Projections from the anterolateral system to the medial thalamic nuclei provide nociceptive signals to areas in the frontal lobe, the insula, and the cingulate cortex [9].The full experience of pain engages a complex network of interacting brain regions, a simplified version of which is illustrated in figure 1.

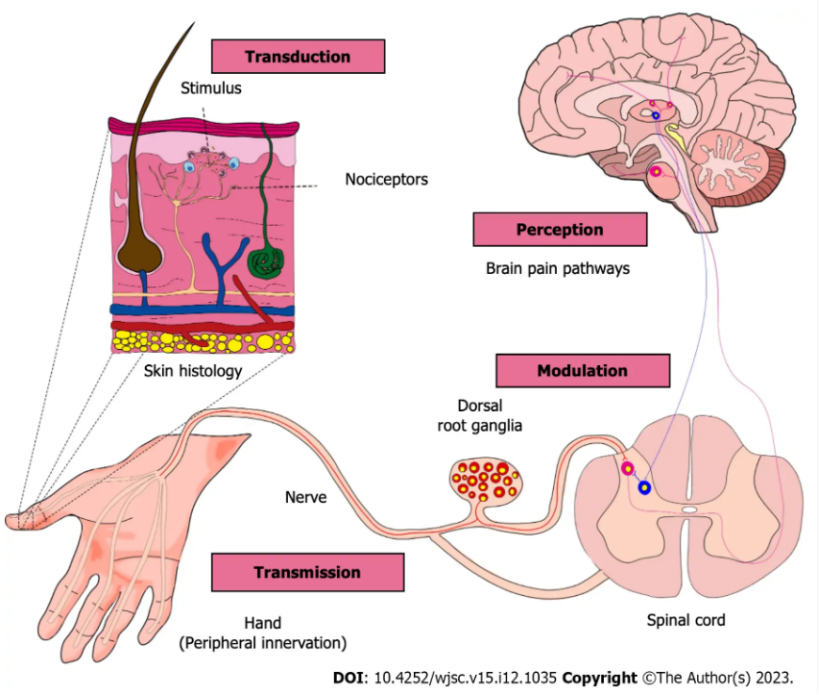


FIGURE 1: Diagram with the anatomical levels involved in pain processing from detection to modulation. The process begins with the detection of a noxious stimulus on the skin by nociceptors. These nociceptors with cell bodies in the dorsal root ganglia, relay the stimulus to the spinal cord via their axons. The signal then ascends to the brain (pink), where it is interpreted as pain. Descending pathways (blue) subsequently activate to modulate and limit nociceptive input at the spinal cord level. Figure adapted from [19]

2.2 Assessment of Pain

The foundation of an effective pain treatment is reliable patient assessment [20]. However, measuring pain is a complex task given its subjective and multidimensional nature. Pain can be modulated at peripheral, spinal, and supra-spinal levels, and the nature and number of these processes can change over time [21, 22]. The current gold standard pain assessment tool relies on self-reporting [4, 23]. While pain scores are considered an accurate and reliable measure for assessing a patient's pain and response to pain treatment, they have significant limitations. Primarily, they measure only intensity, ignoring factors such as time dependency and pain type [22]. Additionally, self-reports are not feasible for many vulnerable populations or non-communicative patients, such as those with speech disorders or consciousness impairments [4]. These limitations underscore the need for mechanism-based pain assessments that provide more comprehensive information and that are able to give insight into the mechanisms underlying an individual's pain experience [24].

Several approaches and stimuli are used to assess pain, including thermal (cold and heat), tactile, ischemic, reactions to pain mediators (e.g., capsaicin or hypertonic saline injection), electrical, among others. The choice of stimulus depends on factors such as the target sensory pathway, desired specificity, equipment availability, and ease of use. Monofilaments (von Frey hairs) or small brushes are commonly used to assess tactile sensation threshold due to their simplicity and cost-effectiveness. These tools are suitable for diagnosing condition like allodynia (pain from typically non-painful stimuli) or hyperalgesia (increased response to painful stimuli) [21, 25]. However, inconsistencies in using monofilaments for sensory detection thresholds have been described, including the subjective judgment required by the examiner to determine if there was pain or not and the subjective pain scores reported by subjects [26, 27]. Other tactile stimulation method

is pressure threshold, which can be assessed using an algometer probe. This probe is pressed against a predefined skin area with ascending stimulus intensities until pain is reported. While reliable, this method's results can be affected by protocol variations such as instrument size and material, application rate, angle, etc [28, 29].

Thermal testing can range from simple tools such as cold water buckets to specialized thermodes or laser stimulators These stimuli provide information about factors like temporal summation of pain or conditioned pain modulation due to their broad range of pain responses. However, specialized equipment is often needed for accurate stimulation and there is a higher risk of skin overheating or tissue damage compared to other methods. Accuracy may also be limited by the skin's biothermal properties [7, 25, 30]. Electrical stimuli are valued for their precision and control over stimulus intensity, directly targeting nerve fibers specific to pain reception by selectively activating Aδ axons. These stimuli are useful for studying nerve conduction, temporal summation of pain, and conditioned pain modulation, although responses may not always correlate with natural stimuli like heat or cold [7, 21, 30].

The described methods can be employed in Quantitative Sensory Testing (QST) to get more comprehensive results. This method determines thresholds or stimulus response curves for sensory processing under normal and pathological conditions. In this type of test, the stimulus is quantified and used to measure perception. The protocol and interpretation can focus on different characteristics such as the minimum threshold perceived, localization, threshold perceived as painful, tolerance, among others [24, 30].

2.3 Psychophysics

Psychophysics is the scientific study of the relation between a physical stimuli to the sensation or perception they produce. This field aims to quantitatively investigate how physical events relate to sensory experiences. The utility of its application lies on its ability to provide a framework to quantify and describe sensory systems. Psychophysics can be applied to any sensory system such as hearing, touch taste and even pain. Following this framework, potential methods to characterize human nociceptive processing involve measuring psychophysical responses to well defined stimuli [31, 32, 33].

2.3.1 Experimental set-ups

For describing the pain system, one of the main variables is threshold. Several psychophysical paradigms are available to estimate this measure. A common method to classify the different paradigms is based on the number of stimulus presented on each trial. In case of one stimulus per trial, the term "yes/no" or Go/No-Go (GN) is commonly used, while if two or more stimulus alternatives are presented (one of which is the "target") it is usually talked about a "forced choice" task [32].

In the case of GN procedure, the taget is normally presented on half of the trials, and the observer responds yes or no on each trial depending on whether or not they perceive the target. These experiments are prominent in signal detection literature [32].

Forced-choice tasks can be sub-classified into Interval Forced Choice (IFC) to specify procedures in which the stimulus alternatives are presented in temporal order. The number of response choices, m is an important parameter as it determines a guessing rate ($1/m$). Thus the guessing rate of both a yes/no procedure and a 2IFC tends to be considered 0.5 [32].

Another classification is based on the subject's internal criteria before they respond yes. This phenomena is mostly seen in GN procedures where based on only one stimulus the subject has to determine whether there was a perception or not, while 2IFC procedures are considered criterion free. However, 2IFC tasks can also be prone to a different bias, for instance towards selecting the first interval or the second. Nonetheless, this sort of biases are observed less frequently [32].

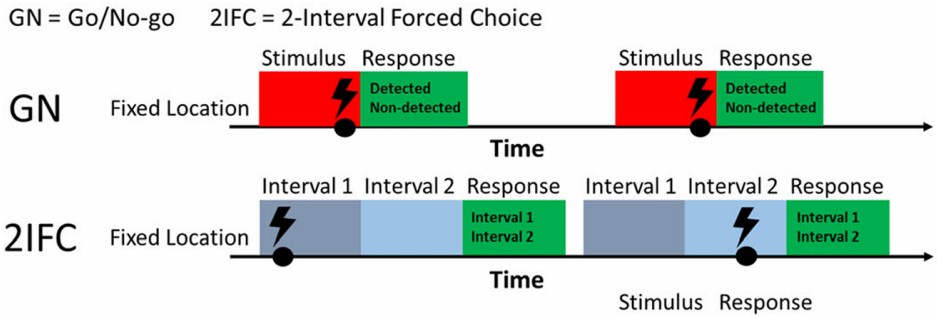


FIGURE 2: Representation of the two experimental procedures described. Adapted from [34].

2.3.2 Psychometric function

Psychometric Function (PF) relate the probability of detecting a certain stimulus to its magnitude and give a description of the subject's behaviour during a psychophysical task. The shape of the PF is similar across a wide variety of tasks and is typically well described by a sigmoidal function. Its estimation allows to determine parameters that summarize behavior, such as the threshold. The sensory or perceptual threshold corresponds to the point in which the subject perceives the stimulus 50% [32, 35].

In addition to the threshold and slope, other parameters necessary to describe the PF are γ (guessing rate) and λ (lapsing rate). The guessing rate is typically determined by the psychometric procedure, in a forced choice task it is usually assumed to be the reciprocal of the number of alternatives. The lapse rate corresponds to the probability of responding incorrectly when a stimulus level is high enough to be felt, it relates to alertness or motivation of the subject. As a result of these lapses the PF will have a asymptote to a value slightly less than one $(1 - \lambda)$ [32]. Taking a guessing rate of 0.5 in the case of no actual perception and an unknown lapsing rate, a expected confusion matrix for a 2IFC would look like the one presented in Figure 3.

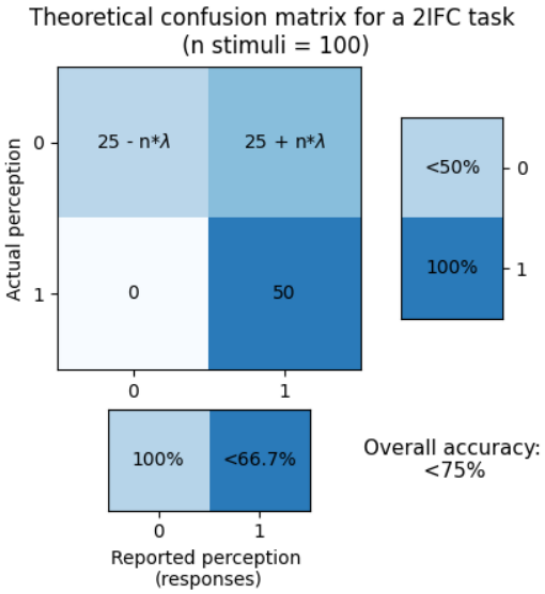


FIGURE 3: Values expected for a confusion matrix in a 2IFC task, assuming a guessing rate of 0.5 and a lapsing rate of λ.

Different theories provide a framework to interpret and describe the results from a psychophysical task. Under

the High Theshold Theory (HTT), the detection of a stimulus depends on the amount of sensory evidence accumulated by the system, which can fluctuate randomly due to external and internal noise. Detection occurs when this evidence exceeds a fixed internal threshold. This threshold is set high enough that when a stimulus of amplitude zero is presented, the probability of exceeding it is also zero. The decision process relies on binary information: whether the evidence surpasses the threshold or not. The PF under HTT represents the cumulative normal distribution, indicating the probability that the stimulus intensity exceeds the threshold [32, 35]. This theory is illustrated in Figure 4.

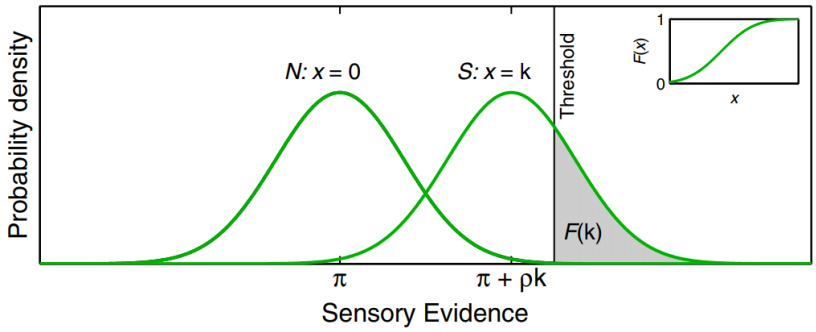


FIGURE 4: Probability density with which a stimulus at a given intensity, k, generates different levels of sensory evidence, following the HTT. In this representation it is assumed that the amount of sensory evidence is a linear $(\pi + \rho x)$ dependent on the stimulus intensity (x) . Figure adapted from [32].

Under the Signal Detection Theory (SDT), the critical difference is that there is not a fixed threshold. Instead, it proposes that sensory mechanisms generate a graded signal for all stimulus intensities, available to the decision process. This theory uses probability density function for both stimulus and noise and there is an overlap between them. The subject's decision is based on the relative amplitude of sensory evidence from two intervals, where the interval with greater evidence is chosen, and the PF is built as the upper half of the cumulative normal density function [32, 35].

In line with the assumption of HTT the expression for the PF, $\Psi(x)$, is presented in equation 2.3.2. This expression differs from the detection probability $F(x)$ as it offers insight into the underlying mechanisms of the sensory system. The PF thus accounts for task specific factors and performance-related variables, which may influence the probability function derived directly from the response pairs in psychophysical procedures [32]. In this equation, the threshold is denoted by α and the slope by β. Parameters related to the tasks such as the lapse rate (λ) and guessing rate (γ) are also taken into account. From the application of the formula, a figure similar to 5 can be obtained.

$$\Psi(x; \alpha, \beta, \gamma, \lambda) = \gamma + (1 - \gamma - \lambda)F(x; \alpha, \beta) \tag{1}$$

Despite HTT's critical assumptions being largely discredited in favor of SDT, the expression derived for HTT is preferred as it distinguishes the term that describe sensory mechanisms (threshold and slope) from the terms that are determined by the task design and non-sensory characteristics. In HTT, the threshold is the sensory evidence needed for detection, while in SDT, even though there is no fixed evidence amount for detection, the term is still used refer to the PF's location parameter [32, 35].

Additionally, a key requirement for fitting an adequate PF is selecting the appropriate range of stimulus levels. Ideally, these levels should result in performance ranging from just above chance to nearly 100% correct [32]. However, in certain sensory systems, such as nociception, a temporal drift can occur, where the

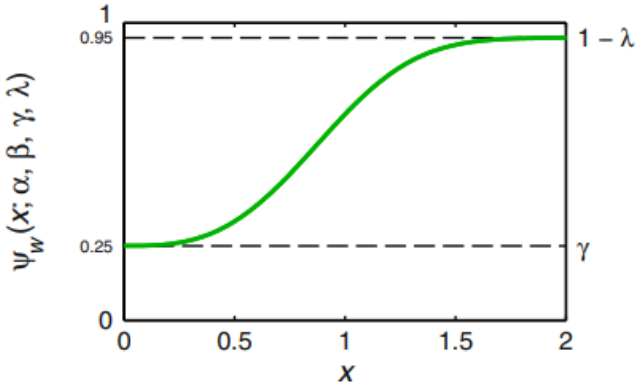


FIGURE 5: Example of a PF with a threshold of 1, a guessing rate of 0.25 and a lapse rate of 0.05. Figure adapted from [32].

threshold appears to increase over time. This drift may be due to factors such as loss of attention, fatigue, changes in decision criteria, or underlying psychophysical functions. Such changes can affect the estimation of threshold and slope, making the stimulus selection procedure particularly important [36].

Stimulus selection procedures can be classified into adaptive and non-adaptive methods. Non-adaptive procedures use a pre-defined set of amplitudes for all stimuli, while adaptive procedures adjust the new stimulation amplitudes based on the preceding stimulus. Adaptive procedures, such as a simple up-down staircase method, where the amplitude of the next stimulus is increased if the subject responds incorrectly or decreased if the subject responds correctly, have demonstrated greater efficiency than non-adaptive procedures [32, 36].

2.3.3 Model fitting

To choose a function to estimate the PF, the decision can be based on an a priori theory of the internal shape of the PF or a posteriori considerations, using the function that most accurately fits the data. The most common method for fitting the data is Maximum Likelihood Estimation (MLE), which defines the best fitting PF as the one most likely to replicate the experiment as performed by the subject [32].

MLE is a standard approach to parameter estimation and inference in statistics. It estimates the parameters by defining the likelihood function, which represents the probability densities most likely to have produced the observed data:

$$L(w|y) = f(y|w) \tag{2}$$

Where $f(y|w)$ denotes the Probability Density Function (PDF) specifying the probability of observing the data vector y given the parameter w and $L(w|y)$ represents the likelihood of the parameter w given the observed data y, and is therefore a funcition of w. The value of interest is the parameter vector that maximizes the likelihood function [37].

In practice, the application of the MLE criterion can be done through fitting a linear model or one of its variants, such as a Generalized Linear Model (GLM). Linear models provide a way of describing a response variable in terms of a linear combination of predictor variables. However, GLMs extend linear modeling to accommodate different types of response variables, such as binary responses, which are particularly suitable for GN experiments where the answers are yes/no [38, 39]. In these models, the effect of the predictors on the response is expressed through a linear predictor:

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q \tag{3}$$

To adapt to different response types, a link function must be defined [38, 39]. The link function g describes how the mean response, μ is linked to the covariates through the linear predictor:

$$\eta = g(\mu) \tag{4}$$

In the case of binary responses, a logit function is often used. This specific type of GLM is termed logistic regression [38, 39]. In logistic regression, the model is presented as follows:

$$\text{logit}(\pi) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q \tag{5}$$

A logit is the mathematically inverse of the standard logistic function, which in the case of a PF can be written as:

$$F_{T,s}(x) = \frac{1}{1 + e^{-s(x-T)}} \tag{6}$$

Where $F_{T,s}$ is the detection probability, x represents the stimulus intensity, T is the threshold and s is the slope. Even though theoretically F is described by a cumulative normal distribution, a logistic regression is a close approximation commonly used due to its easy implementation [32].

2.4 Cortical activity measurements

EEG is a valuable noninvasive tool for assessing pain responses due to its high temporal resolution, clinical convenience, low cost, potential portability, and ease of maintenance, especially when compared to other imaging tools such as Functional Magnetic Resonance Imaging (fMRI) [8]. A key characteristic derived from EEG recordings are the EPs, often referred to as an Event Related Potential (ERP). These are structured, time-locked responses induced by sensory, motor, or cognitive events. One of the most studied ERPs in the field of Brain Computer Interface (BCI) is the P300, which presents an observable peak roughly 300 ms after stimulus onset. The P300 is an induced potential, related not to the physical characteristics of the stimulus, but to the task of recognizing or perceiving it [40].

EEG has been used in various ways to assess and study pain. In a review conducted my Mussigmann (2022) different study types are identified, such as those using provoked pain protocols in healthy controls and those related to resting-state EEG biomarkers associated with chronic pain [41]. In provoked pain tasks, EPs are commonly described by peaks based on their polarity and are termed N1, N2, and P2 [42]. The temporal onset of these peaks varies depending on the study, and their shape has proven to be dependent on the choice of filtering. The first two negative peaks are mainly present in contralateral channels and are usually reported at 150 to 230 ms after stimulation. P2 is typically a larger central peak occurring between 290 to 415 ms. An example of these peaks and their variation due to filtering is shown in Figure 6 [43].

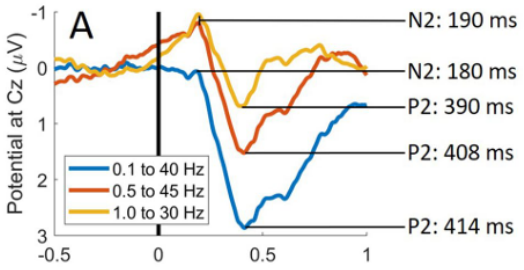


FIGURE 6: Example of nociceptive EP at EEG channel Cz with common average reference with varying filtering properties. Figure adapted from [43].

Brain structures involved in the processing and perception of pain are termed the pain matrix and include areas such as the anterior cingulate cortex, primary and secondary somatosensory cortices, insular cortex, thalamus, and prefrontal cortex [4, 44]. The described peaks mainly originate from the somatosensory cortices, and their amplitudes are sensitive to damage to nociceptive pathways, making them clinically useful measures of the integrity of these pathways in the brain. Additionally, they are proportional with stimulus intensity and can be modulated by factors such as attention or placebo effects [42].

2.5 Analysis of EEG data

EEG data is inherently noisy given that it can be easily contaminated with different artifacts. These artifacts can be broadly classified into two categories: (1) physiological artifacts, which relate to sources in the human body such as eye blinks, eye movement, muscle artifacts, electrocardiogram, among others; and (2) non-physiological artifacts, which relate to external factors such as power line noise, electrode malfunction, variation in impedances, etc [45]. Consequently, preprocessing is traditionally the first step in EEG analysis, focusing primarily on artifact identification and removal. After preprocessing, feature extraction is performed to select the most discriminative features from the data. This process can be highly complex and time-consuming, often leading to a loss of important information. To address this, approaches using deep learning have gained popularity to automatically extract discriminative features [46, 47]. The following subsections will discuss topics related to the analysis and processing of EEG in more detail.

2.5.1 Preprocessing

Different approaches can be followed depending on the study. In the case of ERP analysis, one common strategy to remove noise is the repeated presentation of stimuli and to average out the noise. A complementary strategy is to use digital signal processing to remove artifacts. Among these strategies filtering, baseline correction and artifact identification and removal are common. Additionally, visual inspection of the raw EEG data by expert EEG researchers continues to be widely used and recommended, however is both time-consuming and imprecise as it relies on the observers subjective interpretation of artifact, specially consider a large inter-subject variability inherent of EEG recordings [47, 48].

In ERP analysis, noise is commonly reduced by repeatedly presenting stimuli and averaging the potentials. Digital signal processing techniques such as filtering, baseline correction, and artifact removal are also used. Expert visual inspection of raw EEG data is common and usually considered the gold standard, however, it is highly subjective due to inter-subject variability and time-consuming [47, 48].

In a study conducted by Delorme (2023), performance of different preprocessing techniques were compared against the gold standard of visual inspection by several raters. It was found that high-pass filtering, particularly above 0.1Hz, was the method with the best performance [48]. On the other hand, baseline

correction tends to be considered standard in ERP research, however it has been found that subtracting the pre-stimulus baseline from the entire epoch can introduce distortions and confounds, ultimately, impacting negatively the quality of the signal. Nonetheless, filtering and baseline correction may aid in visualizing the pre-stimulus interval and removing drifts that may obscure peaks of interest [48, 49].

Choosing an appropriate reference site is crucial since EEG signals represent the potential difference between scalp locations and a reference. The Common Average Reference (CAR), which subtracts the average of all electrode locations from each electrode, is widely used. Other strategies include using mastoids, earlobes, nose, Cz, or rereferencing using a reference electrode standardization technique (REST), their selection will depend on the location of the potential being studied [50].

For specific artifacts like eye blinks, regression can be used to remove them using Electrooculography (EOG) channels. Independent Component Analysis (ICA) is more commonly used, assuming sources are independent and non-Gaussian, but it requires a sufficient number of electrodes and manual intervention [51]. Automatic pipelines are prefered, such as Autoreject, which estimates peak-to-peak thresholds to detect bad trials and interpolate sensors if necessary. Comparisons with different datasets have shown that Autoreject effectively detects and interpolates artifacts while maintaining data integrity [52]. However, some argue that rejecting bad segments may reduce statistical power [48] and that minimal to no preprocessing could be beneficial, specially in machine learning applications [47].

2.5.2 Machine learning

Machine learning algorithms, particularly artificial neural networks, have proven highly effective in the study and classification of biomedical signals due to their ability to automatically learn complex features from large datasets [4]. This is especially useful for large and structured data like EEG [47], where their application consistently outperforms conventional pattern recognition methods [53].

Traditional preprocessing and feature extraction methods are often complex and time-consuming, leading to potential loss of critical information. Automated feature extraction algorithms address this issue by allowing deep learning models to analyze large datasets without the need for extensive preprocessing or detailed feature extraction, thus preserving essential information. [54]. Consequently, various deep learning strategies have been adopted in many studies. In EEG research, deep learning is particularly valuable for single-trial classification tasks, such as P300 detection [53]. Detection is commonly performed by using a Neural Network (NN), specifically, a Convolutional Neural Network (CNN) is one of the most popular types of NN applied in this field [47, 55].

CNNs are powerful tools for feature extraction from data matrices [54]. They extract features by convolving the input data with filters, passing these features through multiple layers for further convolution until reaching the output layer, which provides the probability of belonging to each class [56]. CNNs can be designed to operate convolutions along temporal and spatial dimensions separately to extract various types of features [54]. Among these networks, the model proposed by Lawhern et al. (2018), called EEGNet, is one of the most successful for single trial ERP detection [57]. Further improvements to EEGNet include integrating a Recurrent Neural Network (RNN), as seen in [58]. Other research often uses EEGNet as a benchmark, such as Santamaría-Vázquez et al. (2020), which applied inception modules to EEG analysis [59], and Zhang et al. (2021), which utilized permutation layers to extract spatiotemporal features [54].

2.6 Improving machine learning models

A significant drawback of using deep learning methods, such as CNNs, is the large number of hyperparameters that define their structure and learning process [56]. These hyperparameters are not learnable and must be selected manually, often based on experience. However, systematic approaches for hyperparameter optimization are available [60]. In addition to hyperparameter optimization, transfer learning, which leverages

large available datasets similar to the target, and ensemble models, which combine different architectures, are common strategies to improve model accuracy. These methods are discussed further below.

2.6.1 Hyper-parameter optimization

Hyperparameters define a model's overall structure and learning process. Unlike model parameters, hyperparameters must be selected beforehand and cannot be learned during training[60]. For CNNs, these include parameters related to network architecture (e.g., number of layers, kernel size, pooling type), training (e.g., epochs, batch size, learning rate), regularization (e.g., dropout probability, weight decay, number of training epochs), and others such as the loss function and activation function. Input parameters like sampling frequency and post-stimulus interval length also influence CNN performance [60, 61, 62].

Selecting hyperparameters typically involves empirical evaluation of a few configurations, often resulting in suboptimal models. Manual exploration of hyperparameter configurations requires significant expertise and is time-consuming. Automated hyperparameter optimization algorithms offer a more systematic approach but often have their own hyperparameters, which tend to be easier to choose given that an acceptable performance can be usually obtained with different configurations. However, CNN training is resource-intensive, and the large number of hyperparameters makes exhaustive search impractical. Current methods optimize only a portion of the hyperparameters leading to shorter tuning time but still allowing for the quality of the results an still be improved [60, 62].

Grid search and random search are common methods for hyperparameter optimization. Grid search evaluates all possible combinations of predefined hyperparameter values but is impractical for more than three hyperparameters and can miss optimal values due to discretization. Random search, while faster and more flexible, does not require discretization and can explore a larger set of hyperparameters. However, it can be inefficient as it does not leverage information from previous evaluations and may require extensive computational time [60, 62].

More advanced methods like Bayesian Optimization (BO) use conditional probability to learn from past evaluations and focus on promising hyperparameter regions, enhancing efficiency. BO is more effective because it concentrates on regions of the hyperparameter space that are likely to yield better performance. However, BO can struggle with the high computational demands of CNN training and lacks effective parallelization [60].

Metaheuristic algorithms, inspired by biological theories, are also used for hyperparameter optimization. Genetic algorithms and Particle Swarm Optimization (PSO) are notable examples. Genetic algorithms apply evolutionary concepts to iteratively improve hyperparameters by simulating natural selection processes, including mutation and crossover. PSO uses a swarm of particles that traverse the search space in a semi-random manner, where the movement of the next iteration is guided by the best-found solutions of each particle and the overall swarm. PSO is advantageous for parallel computing but is limited to continuous hyperparameters and can be computationally expensive [60].

For limited time and resources, multifidelity optimization algorithms like successive halving and Hyperband are preferable. Successive halving, an improvement over random search, starts with a large set of hyperparameter configurations evaluated with small budgets. Poor-performing configurations are progressively discarded, and resources are reallocated to better-performing ones. However, successive halving struggles with balancing the number of configurations and budget allocations. Hyperband addresses this by dividing the total budget into smaller pieces allocated to configurations in rounds of increasing fidelity. This method uses successive halving as a subroutine to efficiently allocate computational resources. Despite its efficiency, Hyperband's random initial configurations may limit effectiveness, as it does not consider parameter correlations and can prematurely discard promising solutions if the initial budget is too small [60, 63].

Combining Bayesian Optimization with Hyperband (BOHB) addresses these limitations by guiding the

selection of configurations in Hyperband with BO. This hybrid approach leverages BO's efficiency in selecting promising configurations and Hyperband's ability to parallelize computations, making it particularly suitable for CNNs. BOHB's systematic configuration selection enhances its effectiveness, and its ability to scale well to handle various hyperparameter types makes it an excellent choice for complex optimization tasks [60, 63].

2.6.2 Transfer learning

In deep neural networks, an interesting behavior is observed when training on images: the first-layer features often resemble Gabor filters or color blobs. These features are consistent across various tasks, regardless of the cost function and dataset [64]. Conversely, the features computed by the last layer of a trained network are highly dependent on the specific dataset and task. Transfer learning leverages this by training a base network on a base dataset and task, then repurposing the learned features for a second target network on a target dataset and task. This approach is effective if the features are general enough to be applicable to both the base and target tasks. Transfer learning is particularly useful when the target dataset is significantly smaller than the base dataset, as it enables training a large target network without overfitting [65].

The standard transfer learning method involves training a base network and copying its first n layers to the target network's first n layers. The remaining layers of the target network are randomly initialized and trained on the target task. The next step often involves fine-tuning the network for the new task. Depending on the target dataset's size, a smaller learning rate and fewer epochs can be used to avoid overfitting. The decision to fine-tune the first n layers depends on the target dataset's size and the number of parameters in those layers. If the target dataset is small and the first n layers have many parameters, fine-tuning may cause overfitting, so the features are typically left frozen. Conversely, if the target dataset is large or the first n layers have few parameters, fine-tuning can enhance performance [64, 65].

Transferability is negatively affected by two main issues: where the networks are split and the specialization of higher-layer features to the original task, which can reduce performance on the target dataset. Additionally, the effectiveness of feature transfer diminishes as the similarity between target tasks decreases [64].

In a P300 detection study, transfer learning showed slight overall performance improvement in the CNN used, but significant benefits for disabled subjects. It is hypothesized that subjects with pathological conditions may struggle with protocol compliance, making pre-training the network on a larger, distinct dataset beneficial [66]. Transfer learning features are especially useful for small target datasets. Furthermore, initializing with transferred features can improve generalization performance even after extensive fine-tuning on a new task, suggesting this technique could generally enhance deep neural network performance [64].

2.6.3 Ensembles

Ensemble methods in machine learning involve combining multiple models to improve overall performance. By aggregating predictions from several individual models, the ensemble can often outperform the best single model within the group. This technique is based on the idea that while individual models might have different types of errors, their combined decision-making can lead to more accurate predictions. A classic example is Random Forest, which combines multiple Decision Trees to achieve better performance than any individual tree [61].

There are several types of ensemble methods, such as voting classifiers, bagging, boosting, and stacking. Voting classifiers aggregate predictions from different models, predicting the class with the most votes in the case of hard voting. A "soft" voting can also be implemented by averaging the outputs of each model. Bagging involves training the same algorithm on different subsets of the training data, either with replacement (bagging) or without replacement (pasting). Boosting trains models sequentially, each one attempting to correct the errors of its predecessor, effectively creating a strong learner from multiple weak learners.

Stacking, or stacked generalization, involves training model to aggregate the predictions of base learners by learning the best way to combine these predictions [61].

Ensemble methods are particularly useful when individual models are diverse and have different strengths and weaknesses. They are effective in improving model accuracy, reducing variance, and handling overfitting issues, especially when the individual models are weak learners or when the training dataset is limited. By using techniques like bagging, boosting, or stacking, ensembles can enhance predictive performance and robustness, making them suitable for a wide range of applications where model reliability and accuracy are critical [61].

2.7 NDT-EP set up

Research at the University relies on intra-epidermal electric stimulation to selectively target nociceptive fibers [7]. This is accomplished using a custom-made electrode consisting of five 0.2 mm microneedles embedded in a flexible silicone layer (Figure 7), placed on the hand for stimulation. The electrical stimuli are delivered as square wave pulses via a one-channel constant current cathode stimulator (NociTRACK AmbuStim) with a uniformly randomized interstimulus interval of 3.5 to 4.5 seconds. The stimulus amplitudes are selected using an adaptive randomized procedure developed by Doll et al. (2015) [36].

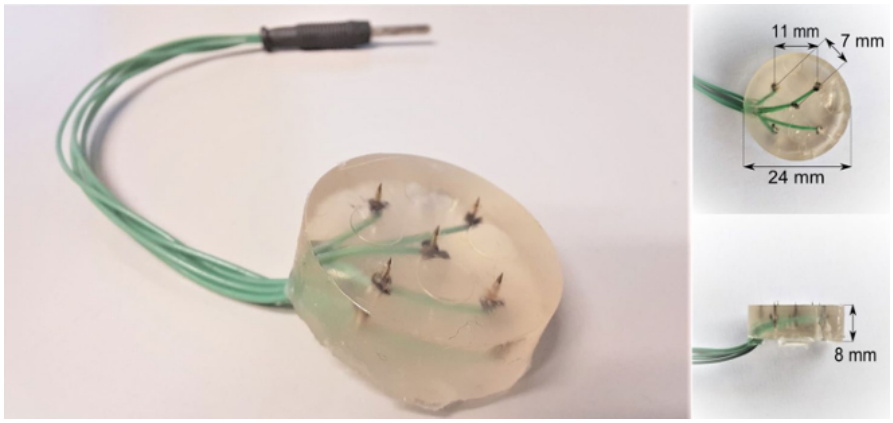


FIGURE 7: Custom made electrode used for intra-epidermal electric stimulation. Figure adapted from [43]

The adaptive procedure for selecting stimulus amplitudes improves upon the classic staircase method. In a traditional staircase procedure, the subject is presented with a series of stimuli of increasing or decreasing magnitude, uniformly spaced out (temporally or spatially). However, this method can lead to subjects becoming accustomed to reporting that they perceived or did not perceived the stimuli or developing anticipation or expectation biases due to the predictability of the stimulation. Despite these drawbacks, the staircase procedure is still useful in the familiarization phase and for obtaining an initial estimate of the stimulation range [36].

The adaptive randomized procedure by Doll et al. begins by defining a vector of k amplitudes, each separated by a distance s (in mA). For each stimulation, one amplitude is randomly chosen from the vector. Based on the subject's response, the next vector of amplitudes shifts towards smaller values if the stimulus is perceived (by a $d_{correct}$ amount) or towards larger values if it is not perceived (by a $d_{incorrect}$ amount). The values for $d_{correct}$ and $d_{incorrect}$ are set such that $d_{incorrect}/(d_{correct} + d_{incorrect}) = p_{threshold}$. In a GN procedure,

where the perceptual threshold is defined as the value at which 50% of stimuli are correctly classified, $d_{correct}$ and $d_{incorrect}$ are equal. In contrast, in a 2IFC task, where the threshold is set at 75% correct classification, $d_{incorrect} = 3 * d_{correct}$ [5, 36, 43].

This approach allows for the estimation of the psychometric curve and the associated parameters such as the Nociceptive Detection Threshold (NDT) by analyzing stimulus-response pairs, minimizing biases due to expectation, and ensuring stimulation around the perceptual threshold. The method is illustrated in figure 8.

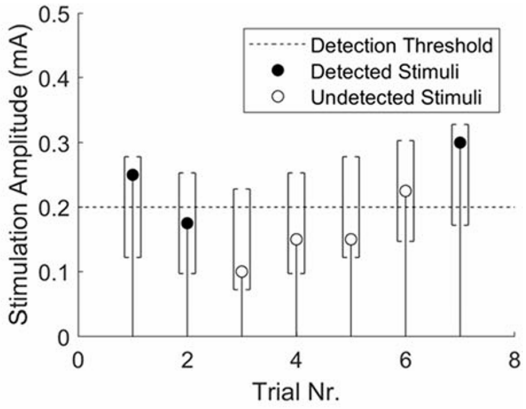


FIGURE 8: Adaptive staircase procedure proposed by [36] used for selecting stimuli during the threshold tracking experiments.

The simulator is controlled via Bluetooth using a LabView application on a computer, which sends the stimulus sequences to the stimulator. This program also supports Multiple Threshold Tracking (MTT), which allows for the simultaneous tracking of perceptual thresholds using different stimulation characteristics. Multiple stimulus settings are defined, and the stimulus type is randomly alternated to reduce predictability. The use of MTT aims to avoid differences in values due to non-stationarities of the nociceptive system [36].

Currently, perception reports rely on a button press, requiring patients to remain focused, to keep the provided button pressed during the experiment and to release when they perceive the stimulation. However, this approach is dependent on the subject's internal criterion and attention. A fully automated approach was created to facilitate an objective and accurate assessment of thresholds, particularly in patient populations where perception report may be challenging. A deep neural network was implemented to detect stimulus perception based on EEG, replacing manual perception responses. The network performs a 2IFC task using pre- and post-stimulus intervals, instead of the GN procedure conducted by the subject. This difference addresses non-stationarities in the EEG and reduces calibration bias introduced by cross-subject application of the neural network (since the classification score is not calibrated per subject). The method was tested using 32 electrodes in a proof-of-concept study and it was demonstrated that neural networks can accurately estimate perceptual thresholds and control adaptive stimulus sequences [5].

2.8 Implications and Objectives

The automatic classification of nociceptive detection thresholds, while promising, has so far been tested only in a small patient group (n=8) using 32 EEG-electrodes. However, fitting this many EEG-electrodes and ensuring good impedance values for each is both time-consuming and uncomfortable for subjects, limiting its practicality for clinical applications [5]. To enhance feasibility and patient comfort, future research should

focus on adapting this automated method so that it relies on fewer electrodes while maintaining a sufficient accuracy. Studies in P300 detection for various BCI applications have successfully employed as few as 8 electrodes, achieving strong classification metrics [59, 58]. It is hypothesized that reducing the number of electrodes further could still maintain signal detection accuracy under certain conditions [66].

Considering this, the feasibility of tracking the nociceptive threshold in real-time with the information from only 8 channels of interest is explored. The channels were selected based on the known distribution of the nociceptive EPs, channels F3, Fz, F4, T7, C3, Cz, C4, and T8 were used given that the potentials are more pronounced on these areas. The objectives treated through this report are:

1. **To improve the performance of the previously used CNN by exploring different architectures and performance-increasing strategies.** This is done in order to obtain an optimized classifier which will be used in further testing.
2. **To explore the accuracy of the neural network model in estimating nociceptive thresholds by comparing the thresholds reported by subjects to those predicted by the model using new experimental data from 8 EEG electrodes.** This objective is guided by the following questions:
    a Is the network able to reliably distinguish between perceived and non-perceived stimuli when compared against subject responses?
    b How do NDT estimates from subjects' GN task and network's 2IFC task compare in terms of reliability across subjects and general values?
    c As an additional measure of the network's classification ability, Is it possible to obtain the expected EP shape under both threshold tracking procedures and how do they compare?
3. **To explore the performance of the neural network model in real-time nociceptive threshold tracking without the performance of concurrent physical tasks.** This objective is set given that a question that may arise is whether the network's ability to perform is primarily reliant on the behavior related to the GN task. Questions related to this objective are the following:
    a When the two networks are used concurrently, is it possible to obtain coherent NDT values?
    b How does the overall performance of the two networks relate to each other in terms of agreement and differences in the estimated parameters?
    c How do the EPs compare between the two networks' estimations?

Finally, a secondary objective is **to explore additional relevant descriptors, such as the psychophysical curve and its related slope, to gain insights into the neural network's capability to estimate these parameters accurately**. By addressing the previously stated objectives and questions, the overarching question this study aims to answer is:

Is it possible to automatically and in real-time track the nociceptive detection probability using only the eight selected EEG electrodes?

3 Methods

To conduct this study, three main stages were performed: (1) exploration of neural networks and methods to improve their performance, (2) selection of parameters through simulations, and (3) application of the network in real-time during experiments conducted at the university. This chapter details the specific methodologies followed during each stage.

3.1 Selection of a neural network

3.1.1 Datasets

During the initial stage of selecting an optimal NN, two datasets were available:

- **DS1**(training dataset): This training dataset comprised EEG measurements from 32 electrodes obtained from 64 healthy participants at St. Antonius Hospital in Nieuwegein. Each participant performed two NDT-EP measurements, one on each hand. Each experiment consisted in approximately 460 stimuli. MTT was performed using three configurations: two with double pulses at a 10 ms interpulse interval and one with a 40 ms interval. The mean age of the subjects was 45.9 years (SD = 16.9), with 35 females. The EEG recordings followed the standard 10-20 montage.
- **DS2**(testing dataset): This dataset was exclusively used for testing, it included 32 EEG-electrode measurements from 10 healthy participants recorded at the University of Twente. Each participant completed a NDT-EP measurement consisting of a total of 150 stimuli, with double-pulse stimuli at a 10ms interpulse interval. The mean age of subjects was 24.4 years (SD = 1.85), with 2 females.

For both testing and most of the training phases, the eight channels of interest were subsampled from the 32 available electrodes.

3.1.2 Architectures

In order to select an architecture fitted for the classification of the noceceptive EPs, various neural network architectures from literature were selected and tested, focusing on those used for P300 response detection and compared against EEGNet. These architectures were tested using 10-fold cross-validation with 9 epochs on DS1. All hyperparameters were set according to van den Berg (2022)[5], except for the frequency rate after downsampling and dropout rate, which used the values specified by each architecture. The following architectures were tested:

EEGNet [57]

EEGNet is a compact, robust CNN designed for BCI classification tasks. In the study were it was propossed, the input data is first preprocessed with a 1-40 Hz band-pass filter, downsampled to 128 Hz, and the signals are extracted from 0 to 1 second post-stimulus. The architecture includes two main blocks: sequential 2D and depthwise convolutions for spatial filtering, followed by separable convolution. It uses the Adam optimizer, categorical cross-entropy loss function, 500 epochs of training, and a 0.25 dropout rate to prevent overfitting in cross-subject applications. This architecture has successfully been applied to nociceptive EPs in [5] with minor changes, such as a 0.1-40 Hz band-pass filter, 512 Hz downsampling, a depthwise convolution depth of 4, 9 training epochs, a batch size of 128, and a learning rate of 0.1 reduced by a factor of 0.2 every 3 epochs.

EEG-Inception [59]

EEG-Inception is a novel CNN tailored for EEG signal and ERP processing, incorporating inception modules inspired by image classification networks to capture dependencies between features at different scales. In the original study, the raw EEG data is first downsampled to 128 Hz, band-pass filtered between 0.5 and 45 Hz, rereferenced to CAR and processed with a spatial filter to improve the Signal-to-Noise Ratio (SNR) of

ERPs. Epochs are extracted from 0 to 1000 ms after stimulus onset, using eight electrodes. The architecture of EEG-Inception leverages inception modules to adapt to the EEG data context. It employs concurrent convolutions with different kernel sizes, which are then concatenated and pooled. EEG-Inception is trained using the Adam optimizer, categorical cross-entropy loss function, with a batch size of 1024 and a learning rate of 0.001. To prevent overfitting the netwrok is trained over a maximum of 500 epochs with early stopping using a patience of 10 and a dropout rate of 0.25. EEG-Inception showed an improvement of about 5.1% in performance over EEGNet, demonstrating its superior ability to handle complex EEG signal patterns.

P3Cnet [66]

P3CNet focuses on optimizing ERP classification using a dataset of signals preprocessed with a 50 Hz notch-filter and a band-pass filter between 2-30 Hz, downsampled to 140 Hz. The architecture comprises three convolutional layers with different kernel sizes, designed to capture a broad range of features. The model uses the Adam optimizer with a learning rate of 0.0005, and early stopping with a patience of 45, training for up to 1000 epochs with a batch size of 128 and dropout probabilities ranging from 0.2 to 0.4 in different layers. P3CNet outperformed the BCIAUT CNN model (an adaptation of EEGNet) on both the dataset it was optimized for and on their newly recorded dataset.

P3Net [67]

P3Net adopts a systematic approach to model selection, examining a wide range of deep learning architectures to identify the best-performing model for P300 ERP classification. This comprehensive method evaluated 232 CNNs, 36 Long short-term memory (LSTM)s, and 320 hybrid CNN-LSTM models across four datasets, ultimately selecting P3Net. The chosen architecture features two convolutional layers with 32 and 16 filters of 7x7 kernel size, followed by two fully connected layers. The EEG data is downsampled to 128 Hz, with training set to 100 epochs, a batch size of 64, and a dropout rate of 0.1. P3Net demonstrated significant performance improvements over EEGNet for most subjects in the dataset, with statistically significant accuracy gains according to the Wilcoxon signed-rank test.

RNN-EEGNet [58]

RNN-EEGNet integrates a RNN with CNNs to leverage RNNs' ability to incorporate temporal information, enhancing the detection of ERP features with time-related sequences of peaks and patterns. The dataset is preprocessed with a band-pass filter of 2-30 Hz and downsampled to 250 Hz, focusing on the 0-600 ms post-stimulus period. The model uses the Adam optimizer with a learning rate of 0.0005, categorical cross-entropy loss, and a dropout rate of 0.25. Training is set for 1000 epochs with early stopping and a patience of 100. RNN-EEGNet showed a slight overall improvement in accuracy compared to EEGNet, indicating some benefits of incorporating temporal dynamics into EEG signal processing.

PLNet [54]

PLNet aims to enhance accuracy by considering the phase-locked characteristics of ERPs, often overlooked by conventional CNNs. EEG data is sampled at 1000 Hz, band-pass filtered between 0.3-28 Hz, and down-sampled to 128 Hz. The architecture consists of three main modules: two for temporal feature extraction, one for spatial features, and a classification module. The use of permutation layers and vector kernels allows PLNet to focus on domain-specific characteristics. PLNet achieved superior performance compared to existing deep learning models like EEGNet, demonstrating its effectiveness in capturing phase-locked features for improved ERP classification.

A summary of the information of the selected architectures is presented in table 1. Complementary information about the architectures is found in Appendix A.

Network	Preprocessing	Architecture	Training Details	Remarks
EEGnet [57]	- 1-40 Hz band-pass filter - Downsampled to 128 Hz - Signals from 0 to 1s post-stimulus. - 32 EEG electrodes	Two main blocks: sequential 2D and depthwise convolutions for spatial filtering, followed by separable convolution.	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Epochs: 500 - Early stopping - Batch Size: 128 - Dropout Rate: 0.25	Wide applicability in different BCI tasks.
EEGnet for nociceptive EPs[5]	- 0.1-40 Hz band-pass filter - Downsampled to 512 Hz - Rereferenced to CAR - Epochs: 0.05 to 1.05s post-stimulus. - 32 EEG electrodes	EEGnet architecture with a depthwise convolution of depth 4	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Epochs: 9 - Batch Size: 128 - Learning Rate: 0.1, reduced by 0.2 every 3 epochs	Accurately identified nociceptive EPs
EEG-Inception [59]	- 0.5-45Hz band-pass filter - Downsampled to 128 Hz - Rereferenced to CAR - Spatial filtering for improved SNR - Epochs from 0 to 1s post-stimulus - 8 EEG electrodes (F7,Cz,Pz,P3,P4, PO7, PO8, Oz, ground at Fp)	Application of inception modules: concurrent convolutions with different kernel sizes, which are then concatenated and pooled	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Epochs: 500 - Early stopping (patience = 10) - Batch Size: 1024 - Learning Rate: 0.001 - Dropout Rate: 0.25	5.1% improvement over EEGNet
P3CNet [66]	- 2-30 Hz band-pass filter - 50 Hz notch-filter - Downsampled to 140 Hz - Epochs: -200 to 1200ms-post stimulus - 8 EEG electrodes (C3, Cz,C4, CPz, P3, Pz, P4, Poz, groud at Afz)	Three convolutional layers with different kernel sizes for feature extraction	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Epochs:1000 - Early stopping (patience = 45) - Batch Size: 128 - Learning Rate: 0.0005 - Dropout: 0.2 to 0.4	Outperformed BCIAUT CNN (an adaptation of EEGNet) on original and new datasets.
P3Net [67]	- 0.1-15Hz band-pass filter - Downsampled to 128 Hz - Epochs from 0-600 ms post-stimulus - 16 EEG electrodes	Two convolutional layers with 32 and 16 filters (7x7 kernel size), followed by two fully connected layers.	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Epochs:1000 - Early stopping (patience = 100) - Batch size: 64 - Learning Rate: 0.0005 - Dropout Rate: 0.25	Significant performance improvement over EEGNet
RNN-EEGNet [58]	- 2-30 Hz band-pass filter - Downsampled to 250 Hz- - Epochs from 0-600 ms post-stimulus - 8 EEG electrodes (C3, Cz, C4, CPz, P3, Pz, P4, POz)	Combination of RNN and CNN layers to leverage temporal and spatial feature extraction.	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Learning Rate: 0.0005 - Early stopping (patience=100) - Dropout Rate: 0.25	Benefits from temporal dynamics in EEG signal processing
PLNet [54]	- 0.3-28 Hz band-pass filter - Downsampled to 128 Hz - Signals from 0 to 1s post-stimulus. - 60 EEG electrodes	Permutation layers and vector kernels for domain-specific feature capture.	- Optimizer: Adam - Loss: Categorical Cross-Entropy - Early stopping (patience=20) - Dropout Rate: 0.25 to 0.5	Effective in capturing phase-locked features for ERP classification.

TABLE 1: Summary of the used architectures with the hyperparameters used in each study.

3.1.3 Performance tuning strategy

Hyperparameter tuning focused on the parameters that significantly impact performance according to literature and exhibit the most variation across different architectures. Parameters such as the optimizer, pooling type, and activation function, which appeared to be consistent across prior studies, were not adjusted. Instead, tuning was focused on parameters like learning rate and dropout probability, which varied significantly among the architectures listed in Table 1. The number of training epochs was not directly tuned; instead, early stopping was implemented to prevent overfitting. Early stopping interrupts training when validation loss stops improving, with the patience value determining the number of epochs to wait for before stopping. This approach balances optimizing training time and preventing overtraining. The search space for the selected parameters is detailed in Table 2.

Hyperparameter type	Hyperparameter	Search space
Training	Batch size	[32, 64, 128, 1024]
	Learning rate	[1e-5: 1e-1] (logarithmic scale)
Input size	Sampling frequency	[128, 140, 250, 500, 512]
	Window start	[0.05:0.35]
	Window end	[0.6:1.5]
Regularization	Dropout probability	[0.1:0.6]
	Dropout type	["SpatialDropout2D", "Dropout", "AlphaDropout"]
	Patience	[2:20]

TABLE 2: Hyperparameters selected for hyperparameter tuning along with their search spaces.

Initially, the hyperparameters related to training and input size are explored using BOHB as implemented by the python library BOHB-HPO. In this library a budget is defined as a integer which in the case of NN is usually defined as the number of epochs. In this study, the computational budget was related to the number of folds in cross-validation. Additionally, the maximum budget determines the number of combinations being tested by the algorithm, based on this, a maximum budget of 100 is selected as it allows for testing a total of 143 hyperparameter combinations. To ensure a maximum of 10 folds, a function ($n_{folds} = Budget^{\log(10)/\log(MaxBudget)}$) was used to map the assigned budget to a range of 1 to 10 folds. With this budget, the algorithm initially selects 81 random combinations to explore the search space. Successive halving is applied, with two-thirds of the non-promising configurations being discarded and assigning more resources to the best ones. After this process, additional rounds are conducted where the selected combinations are based on the regions of interest identified by the BO algorithm.

Once the optimal values for the selected training and input size hyperparameters are determined, these values are used in the second round of hyperparameter tuning related to regularization parameters. A maximum budget of 30 is chosen, allowing for three different groups of hyperparameter combinations and a total of 23 combinations. Similarly, these budget values are adjusted to a maximum of a 10-fold cross-validation. To implement early stopping in hyperparameter tuning, an additional term related to the time needed to complete the run is used to prevent the methods from being biased towards higher values of patience that contribute little to performance.

After this initial search, each hyperparameter is changed from its optimal value while keeping the rest fixed to determine which hyperparameter potentially has the most significant influence on overall performance. Once identified, hyperparameter tuning using only BO (implemented by the hyperopt library) is employed to circumvent the problem of prematurely discarding promising results. To enhance time efficiency, parallel computing is implemented, with each fold running in a different thread.

Once the optimal network is chosen in terms of architecture and hyperparameters, other strategies such as transfer learning and ensembles are tested. Transfer learning is performed using the dataset presented by Santamaría-Vázquez et al., which includes data from 73 subjects (42 healthy, 31 with a disability) using an ERP-based speller recorded at 256 Hz, using 8 channels ('FZ', 'CZ', 'PZ', 'P3', 'P4', 'PO7', 'PO8', 'OZ'), amounting to 701,615 observations [59]. The optimal network previously selected is first trained with this dataset and then fine-tuned with the dataset already used for training. In the case of ensembles, the networks with the highest metrics are combined by averaging the outputs to build a new classifier. These networks are tested with DS2. The selected strategy is then further improved by implementing longer training.

3.2 Simulations

To explore the effects of various parameters associated with the GN and 2IFC procedures, as well as the impact of varying accuracy values, a Python class is implemented to simulate a threshold tracking procedure.

In these simulations, given a threshold and slope, the method constructs the PF curve. Using the stimulation amplitude, the PF maps stimulus intensity to the probability of detecting the stimulus, simulating in this way subjects responses. Other input parameters are related to the size of the vector from which the stimulation amplitude are chosen (k) the step ($d_{correct}$ and $d_{incorrect}$) used to shift the amplitudes of the vector depending on the response and the distance between each of the vector elements (s). In the case of a 2IFC task $d_{incorrect}$ is set to be three times $3 * d_{correct}$, to account for the high guessing rate in the case of non-perceived stimuli. An additional parameter, termed "agreement probability" is introduced to assess the effects of accuracy. To implement the agreement probability, first a decision (1 or 0) is made according to the stimulus intensity and the probability value given by the PF function, after this, the "agreement probability" is used to determine if the made decision will be altered or not. From this process a set of stimulus response pairs is created. A GLM is then fitted in the simulated data, considering responses, stimuli amplitues and trial number, in order to obtain the simulated PF, from which a threshold and a slope can be estimated.

First, the effects of varying k, s, and $d_{correct}$ on how closely the threshold estimates of the 2IFC match the real threshold obtained from a GN procedure are studied. Two vector sizes (5 and 7) are tested, with s values of 0.008 (the minimum amplitude supported by the stimulator), 0.01, 0.015, 0.02, 0.025, and 0.03 explored. Similarly, $d_{correct}$ values are varied within the range of 0.008 to 0.035. All possible combinations of these values for each vector size are tested with agreement probabilities ranging from 0.5 to 1 in steps of 0.05. The sum of the absolute differences in thresholds across tested agreement probabilities is used as a metric for comparison, with the lowest value being ideal.

To assess the effects of accuracy and determine a theoretical minimum accuracy required for experiments, both slope and threshold values are obtained after simulating the threshold tracking procedure data with agreement probabilities ranging from 0.5 to 1. From this stage, the optimal parameters for the 2IFC task are chosen, and the minimum expected accuracy needed for reliable threshold estimates is determined as the value at which the estimates start to converge towards the real threshold.

3.3 Experiments

The following section outlines the experimental setup and methodology employed in this study.

3.3.1 Participants

The study is conducted only on healthy participants recruited at the University of Twente. Before participation, subjects receive all relevant information through an information brochure (Appendix B). The exclusion criteria include skin abnormalities, abnormal blood pressure, heart problems, diabetes, chronic pain, implanted stimulation or electronic devices such as pacemakers, pregnancy, and the use of stimulants, narcotics, or analgesic drugs within 24 hours before the experiment, as well as any pain complaints at the time of the experiments. In total, 15 subjects with an average age of 24.06 years (std = 1.83, 6 females) participate in the study. Ethical approval for conducting this study is provided by the Natural Sciences & Engineering Science committee of the University of Twente.

3.3.2 Experimental procedure

The NDT-EP method is followed, using a TMSi 32-channel EEG low-noise cap with only the electrodes F4, Fz, F3, C3, Cz, C4, T7, and T8 enabled and maintaining impedances below $5k\Omega$. Intra-epidermal electric stimulation is applied with double pulse stimulation at a 10ms interpulse interval. To evaluate the performance of the developed network against the usual button press method and when no concurrent motor task is performed, two tasks are conducted by each subject. The definition of these tasks and their parameters are based on the experiments conducted by van den Berg, et al. [5]. Each subject completes both tasks in a randomized order, with 4-block randomization ensuring a balanced distribution. Detailed descriptions of each task are provided below:

Task 1

The perceptual threshold based on reported stimulus perception is compared to the perceptual threshold estimated by the neural network. Participants perform a GN task by pressing a button and briefly releasing it when a stimulus is perceived. Simultaneously, the selected neural network performs a 2IFC classification task where pre- and post-stimulus intervals are given. Multiple threshold tracking is used to independently track the perceptual threshold via the GN task and the threshold estimated by the neural network. Stimulus amplitudes are selected by two independent adaptive methods of limits in randomized order to center amplitudes around the perceptual threshold, with one using neural network classification as feedback and the other using the button release. A total of 200 stimuli are presented

Task 2

Instead of using a button to report the perceived stimuli, participants count each stimulus when perceived. To simplify the task, participants report the number of stimuli they have felt at various intervals during the procedure. The task is divided, in most cases, into two segments, but depending on the participant's request, the report of the count can be done on three occasions during the procedure. Multiple threshold tracking is used to independently track two pain thresholds using the same neural network. A total of 200 stimuli are presented.

A more detailed description of the followed protocol is found in appendix C.

3.3.3 Analysis

The analysis of the obtained data, including figures, is done entirely using python libraries. The code is available in the BSS Git repository. The neural network's performance is initially assessed using the obtained experimental data, which was recorded using 8 EEG electrodes. Specifically, the network's ability to distinguish between pre- and post-stimulus epochs, in cases where subjects reported perception, is evaluated. This method aligns with the original training and evaluation of the networks. A confusion matrix is constructed to visualize this performance, and its significance is tested using a chi^2 test from the sci-py library. In this study, a 5% significance level is applied to all statistical tests.

For each individual experiment, the initial analysis involves visualizing the tracking procedure (see Appendix D), fitting the PF (which allows to obtain the NDT, slope and drift), and extracting and plotting the EPs. To estimate the PF, a GLM with a logit link function is used as implemented by the statsmodel library. In this model, the dependent variable is the participant's response, while the explanatory variables are the stimulus amplitudes and trial numbers. An additional variance weight term is included to prevent underestimation of the threshold in the 2IFC procedure. This term assumes that a larger weight (bounded between 0 and 1) corresponds to a more credible observation. In this context, network responses indicating a perceived stimulus are given a weight of 1 (as they are considered more credible), while non-perceived stimuli are weighted at 0.5 to account for the 50% guessing rate inherent in the 2IFC task. For plotting the EPs, the signal is cleaned using the autoreject algorithm from the mne-python library.

For the group analysis, both independent threshold tracking procedures obtained in each task are initially visualized using boxplots to display the NDT, slope, and drift. These values are further analyzed by testing for significant differences in the means using a chi^2 test and assessed for agreement by calculating the Intraclass Correlation Coefficient (ICC) according to equation 3.3.3. This formula refers to the Two-way mixed effects, consistency, single rater/measurement or ICC(3, 1), which was determined to be the most suitable for this case [68]. In equation 3.3.3, MSB denotes the mean square between subjects, MSR represents the mean square of the residual, and k is the number of classes (in this case, 2) . To further visualize agreement, Bland-Altman plots are constructed. Additionally, the obtained EPs are examined using butterfly plots and Global

Field Power (GFP), which allows to identify time points with the highest overall amplitude [69]. Individual channels of interest are also plotted and contrasts between the perceived and non-percieved EPs for both GN and 2IFC procedures are tested using cluster-based non-parametric statistical testing. Topographical maps of the EPs are created at time steps identified through the GFP analysis to understand the spatial localization of responses.

$$ICC = \frac{MSB - MSR}{MSB + (k-1) * MSR} \tag{7}$$

For Task 1, an additional analysis is conducted to assess the network's performance during real-time application. A confusion matrix is constructed using the button responses as true labels and the outputs from the NN application as the predictions. In Task 2, an additional analysis involves visualizing the participants' reported counts and calculating the ICC between reported counts in each half of the experiment to assess the reliability of the NN throughout the task.

4 Results

The results obtained from this work are divided into three stages: (1) exploring neural networks and methods to improve their performance, (2) conducting simulations to select parameters for the experiments, and (3) applying the network in real-time during the experiments.

4.1 Selection of an improved neural network

The metrics from the initial exploration of the previously used network with all 32 EEG electrodes is shown in table 3. This table also presents the performance metrics when the network was applied to the eight (subpsampled) electrodes of interest. A noticeable drop in average of approximately 5% is seen due to the subsampling. These initial performance metrics served as comparison points to assess changes and improvements to the networks.

	Sampling frequency	Hyperparameters	Av. Loss	Av. Accuracy	Best loss	Best accuracy
32 electrodes	512	van den Berg, 2022 [5]	0.52 (0.056)	0.73 (0.039)	0.44	0.796
	128	Lawhern, 2018 [57]	0.52 (0.055)	0.74 (0.038)	0.43	0.798
8 electrodes	512	van den Berg, 2022 [5]	0.60 (0.046)	0.69 (0.029)	0.50	0.76
	128	Lawhern, 2018 [57]	0.59 (0.043)	0.69 (0.038)	0.52	0.76

Table 3: Results from the initial exploration of the neural networks with all 32 electrodes and a subsample of eight electrodes of interest.

Several other architectures were implemented and evaluated using DS1, with the performance metrics from a 10-fold cross-validation summarized in Table 4. The EEG-Inception network showed a 2% increase in accuracy compared to the obtained baseline. While the accuracy of PLnet remained similar to that of EEGnet, PLnet exhibited slightly better performance, as indicated by a marginally lower loss. The remaining networks tested showed accuracy values slightly below those of EEGnet. Due to its superior improvement in accuracy and loss the Inception network was selected for further exploration. EEGnet was also retained for comparison purposes after performing hyperparameter tuning.

Network	Average Loss	Average Accuracy	Best Loss	Best Accuracy
Inception [59]	**0.562 (0.052)**	**0.714 (0.032)**	**0.478**	**0.771**
P3Cnet [66]	0.604 (0.038)	0.681 (0.027)	0.538	0.726
P3net [67]	0.602 (0.053)	0.686 (0.035)	0.496	0.768
PLnet [54]	0.577 (0.037)	0.697 (0.027)	0.495	0.758
RNN - EEGnet [58]	0.667 (0.157)	0.658 (0.057)	0.504	0.751

Table 4: Results from testing different architectures obtained from literature.

In the first round of hyperparameter tuning, which focused on training parameters and input size, the values that resulted in the best metrics are described in Table 5. The evaluation of these parameters using 10-fold cross-validation with DS1 is shown in Table 6. Little to no improvement during this round of hyperparameter tuning is observed. EEGnet's average accuracy decreased slightly by around 1% while EEG-Inception showed a minimal increase of 0.3%.

Network	Batch Size	Learning Rate	Sampling Freq [Hz]	Window Start [s]	Window end [s]
EEGnet	64	0.003	250	0.25	1.49
Inception	64	0.05	140	0.35	1.3

Table 5: Values for the best performing parameters after the first round of hyperparameter tuning.

Network	Average Loss	Average Accuracy	Best Loss	Best Accuracy
EEGnet	0.589 (0.0447)	0.682 (0.0428)	0.524	0.729
Inception	0.570 (0.065)	0.717 (0.0406)	0.472	0.779

TABLE 6: Results from testing the best found hyperparameters described in Table 5 with 10-fold cross validation.

Using the hyperparameters listed in Table 5, a second round of hyperparameter tuning focused on regularization parameters was conducted. For EEGnet, the optimal hyperparameters identified were a patience value of 8 for early stopping, a dropout rate of 0.18, and the use of "SpatialDropout2D" as the dropout method. For Inception, the selected hyperparameters were a patience value of 2, a dropout rate of 0.21, and also "SpatialDropout2D". The performance results from the 10-fold validation are shown in 7. Here, a consisted performance of EEGnet compared to the metrics obtained before tuning is observed. However, for EEG-Inception, a slight decrease in accuracy of about 1%was observed after applying the tuned hyperparameters, similar to the results seen in the previous round of tuning.

Network	Average Loss	Average Accuracy	Best Loss	Best Accuracy
EEGnet	0.566 (0.0418)	0.699 (0.0399)	0.517	0.744
Inception	0.577 (0.0463)	0.699 (0.0329)	0.498	0.765

TABLE 7: Results from testing the best-found hyperparameters after improving the regularization parameters with 10-fold cross-validation.

Window size appeared to have the most significant impact on network performance. To test this, each hyperparameter was varied from the optimal value while keeping the others fixed. Since Inception outperformed EEGnet, further tuning was done only with this network. Another round of hyperparameter tuning using BO resulted in selecting a window size between 0.05 and 1.4 seconds post-stimulus. The 10-fold cross validation of this last hyperparameter tuning resulted in an average loss of 0.536 (std=0.05), an average accuracy of 0.728 (std=0.028), which constitutes a modest increase in accuracy of about 1%. The performance of this tuned network was compared with a network trained with transfer learning (with and without fine-tuning with a window size of 1 second), an ensemble network built by averaging the outputs of EEGnet, PLnet, and Inception (the best-performing networks according to Table 4), and the networks with the original parameters. Results of this comparison on dataset DS2 are shown in Table 8.

Network	Loss	Accuracy
Inception (tuned window)	0.526	0.735
Transfer learning + fine tuning	0.571	0.731
Transfer learning	0.606	0.718
Ensemble (EEGnet, Inception and PLnet)	0.548	0.710
Inception	0.547	0.709
EEGnet	0.591	0.695

TABLE 8: Comparison of the best-tuned network with other strategies to improve network performance.

Based on the results shown, the Inception network with a tuned window was chosen as the best model, since it represents a 4% increase in accuracy when compared to the application of the same dataset on the original EEGnet. This model was then trained for a longer period on DS1, with a patience of 25 and a duration

of 200 epochs. After this, the trained network was tested again on DS2, resulting in a loss of 0.516 and an accuracy of 0.759. This trained network was used for further experiments.

4.2 Simulations

Simulations of the GN and 2IFC threshold estimation procedures were carried out to explore the impact of various parameters. The simulations evaluated differences in estimated thresholds by varying factors such as the size of the current amplitude vector, the distance between vector elements, and the step size used to adjust the vector in subsequent trials. For each combination of step and distance between elements, the difference between the threshold obtained from the simulation and the real value were evaluated. These difference was calculated for varying simulated accuracies and their sum is shown in Figure 9.

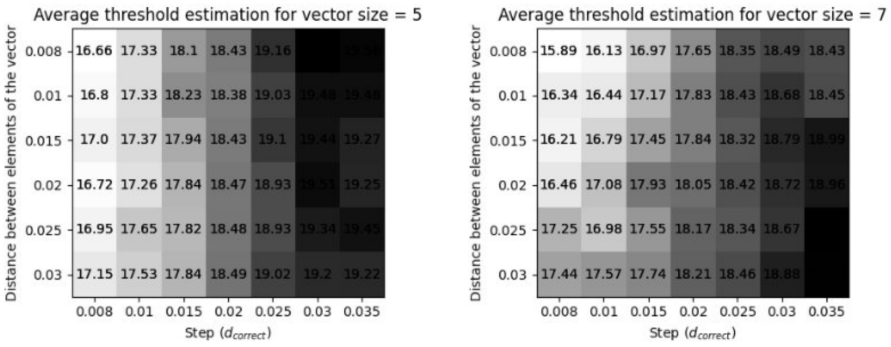


FIGURE 9: Effects of vector size, distance between elements, and step size on the accuracy of threshold estimation. Lower values indicate a more accurate estimated threshold.

In addition, the potential effects of different accuracy values for a machine learning model were explored, as shown in Figure 10. This analysis helped establish a possible minimum accuracy standard for evaluating the networks. For threshold estimates, a minimum accuracy around 70% appeared to be sufficient to yield reliable threshold estimates. Conversely, achieving an accuracy higher than 90% seemed to be necessary for the estimated slope to approximate real values.

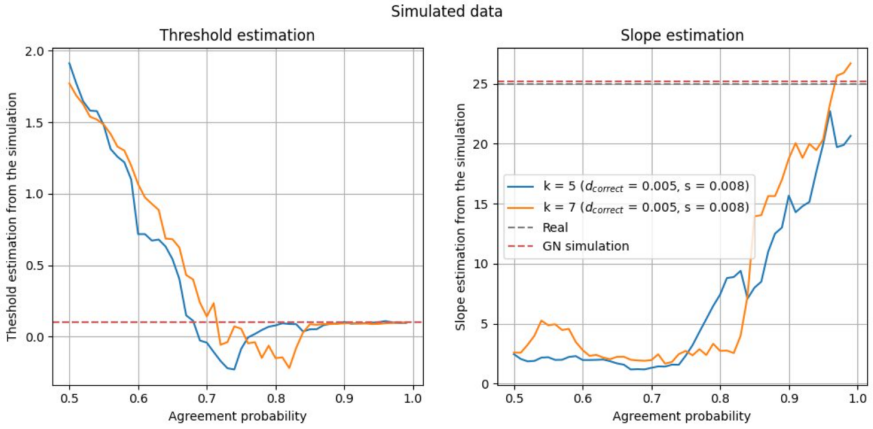


FIGURE 10: Expected impact of neural network accuracy on threshold estimation. The gray dotted line represents the threshold and slope obtained from the real experiment. The red dotted line corresponds to the threshold and slope from simulated GN data using experimental parameters. The blue and orange lines show the estimated threshold and slope over different agreement probabilities for the simulation of a 2IFC procedure.

Based on the results obtained here, the parameters listed in table 9 were selected for the experiments, which align with those proposed in [5].

Procedure	Vector Size	Distance Between Elements	Step
GN	5	0.025	0.025
2IFC (NN)	7	0.008	0.008 (0.024 incorrect)

TABLE 9: Selected parameters for the real-time experiments for each procedure. The GN procedure corresponds to the subject's response, and the 2IFC procedure relates to the neural network responses.

4.3 Experimental results

All participants successfully completed the two proposed tasks. However, there were specific issues with data collection for some subjects. For Subjects 3 and 5, the T6 electrode was broken, therefore 7 electrodes were active instead of the intended 8. Additionally, Subject 1 was excluded from the analysis of Task 2 due to a damaged cable, which lead to unreliable results from this test. Furthermore, the data for Subject 4 during Task 1 was not recorded correctly because of an error during task setup, which resulted in the settings from task 2 being mistakenly used.

The performance of the network was evaluated using the recorded dataset, which used only the eight electrodes of interest. Figure 11 shows the metrics related to the classification of pre- and post-stimulus activity in the subset of stimuli marked as felt by the subject. An overall accuracy of 71% is seen with a significant association between the predicted and actual classes ($p<0.001$). Compared to the accuracy obtained in the test set, a drop of 4% is seen. The network appears to be slightly better at classifying post-stimuli compared to pre-stimuli intervals.

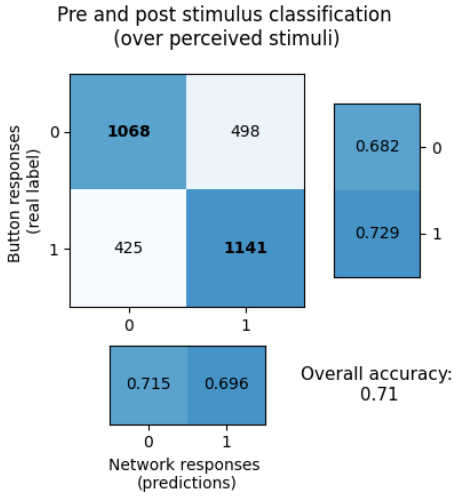


FIGURE 11: Neural network performance in classifying pre- and post-stimulus activity for stimuli marked as felt in the recorded dataset.

Further analysis of the data obtained during the experiments is divided into the following sections:

4.3.1 Task 1

To further assess the network's performance, Figure 12 shows a confusion matrix built with data from the real-time experiment. Here, instead of post- and pre-stimulus, the predicted labels correspond to those

assigned by the network while the participant was performing the GN procedure and are compared to the subject's responses. Despite a low overall accuracy of 60%, the results are not due to chance (as seen by a p-value 0.034 in the conducted chi^2 test) and follow the expected pattern present in Figure 3.

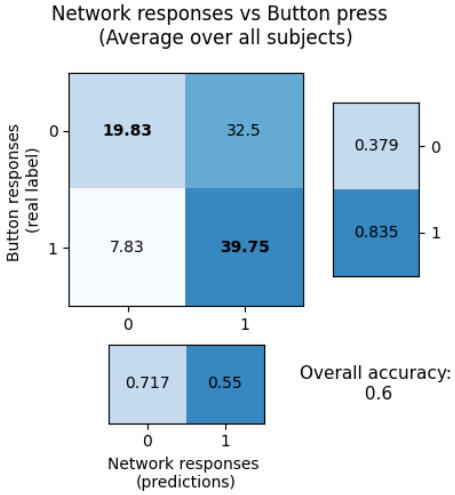


FIGURE 12: Neural network's real-time performance in classifying stimuli averaged over all subjects. The real labels refer to the participants' responses during the GN procedure, while the predicted labels are the results of the 2IFC task performed concurrently by the neural network.

Figure 13 presents the estimated perceptual thresholds, slopes, and drifts for each participant, with measurements from the same individual connected by a dotted line to visually compare the two methods. The p-values from the Wilcoxon rank test and the corresponding ICC values for each metric are also provided. Although the 2IFC method generally yields lower NDT estimates compared to those obtained using the GN procedure, this difference is not statistically significant ($p > 0.05$). The substantial crossing of lines between the same subjects indicates a high variability between methods, as evidenced by the low ICC, which suggests poor reliability. Regarding the other metrics, the average drift obtained from the two methods was not significantly different, the ICC of 0.67 suggests moderate reliability. In terms of slope, the 2IFC method resulted in significantly lower slopes and low reliability between the methods with an ICC of 0.19.

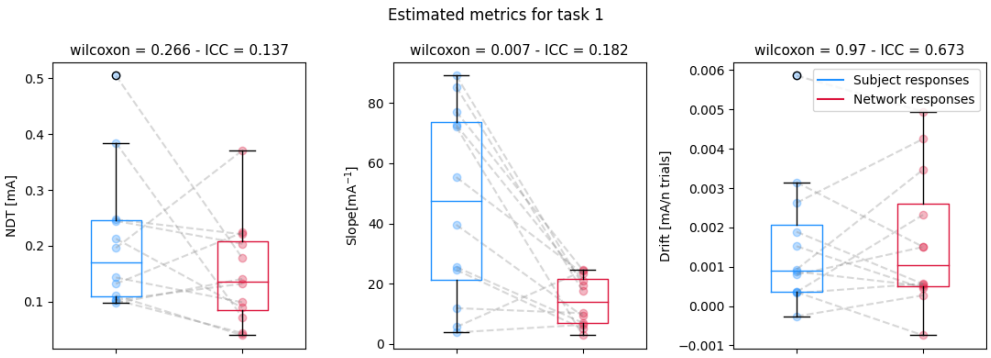


FIGURE 13: Boxplots of the estimated threshold, slope, and drift after fitting a GLM in the data obtained for Task 1. The dotted lines join data points from the same subject.

The estimated metrics are also analyzed using a Bland-Altman plot, as shown in Figure 14. In general, for all metrics a very big variability, with large confidence intervals. In the figure related to the threshold,

most differences (9/12) seem to lay close to zero. Also, a tendency of the button press values to be higher can be observed considering a positive mean difference. The small value of this mean differences and the described observation aling with the non-significant difference observed in Figure 13. In the case of the slope, a seemingly positive linear tendency is seen, while for the drift, most values seem to lay to the left of the graph, towards small means between methods.

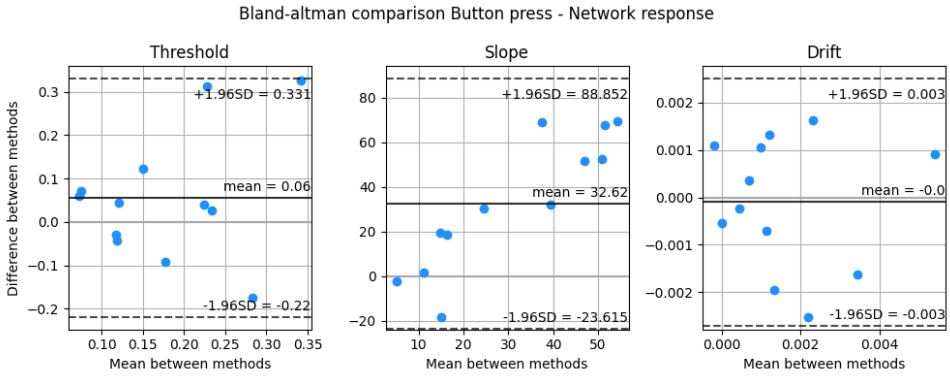


FIGURE 14: Bland-Altman plot for comparing agreement between estimated thresholds, slopes and drifts obtained by subject responses during a GN procedure in contrasts to those obtained through the implementation of a neural network performing a GN procedure.

The EPs generated by the stimulation are shown in a butterfly plot along with their GFP in Figure 15. EPs obtained from the GN procedure and the 2IFC are contrasted for both detected and non-detected stimuli. EPs are also shown in Figure 16, where only responses from channel Cz, which has the largest potential according to the butterfly plot, are depicted. In both figures the potentials derived from the GN procedure appear to be about two times larger than those derived from the 2IFC task. Additionally, only in the EPs for the subject responses, a significant contrast between stimuli classified as perceived and non-perceived was identified.

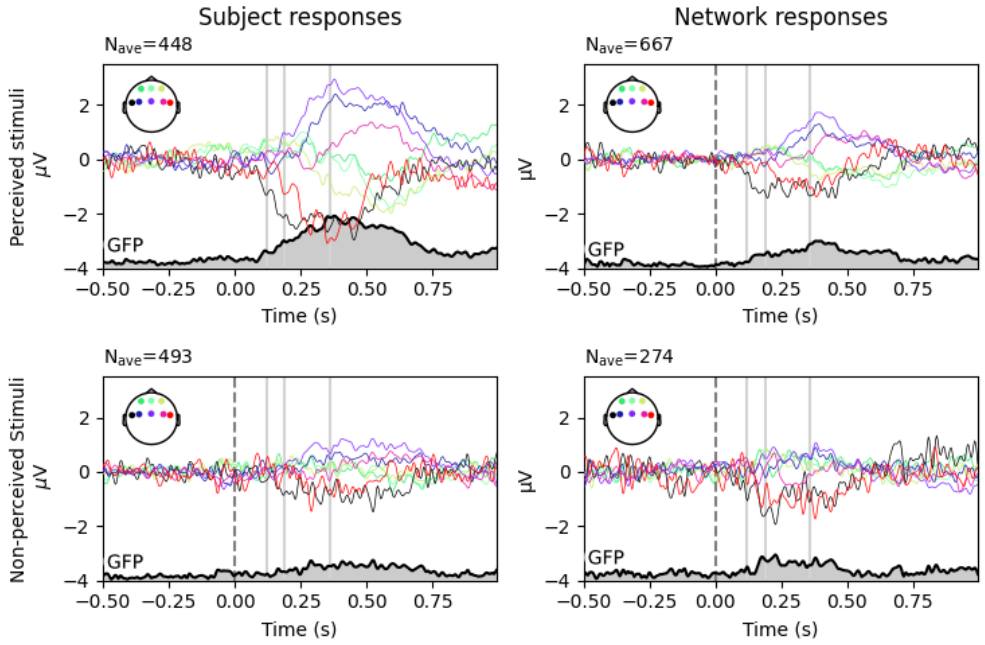


FIGURE 15: Butterfly plot of the grand average potentials and GFP of the EPs resulting from intra-epidermal stimulation during task 1. The vertical dashed line represents the time point when the stimulus is given while the other vertical lines represent the possible time points in which the nociceptive peaks are observed.

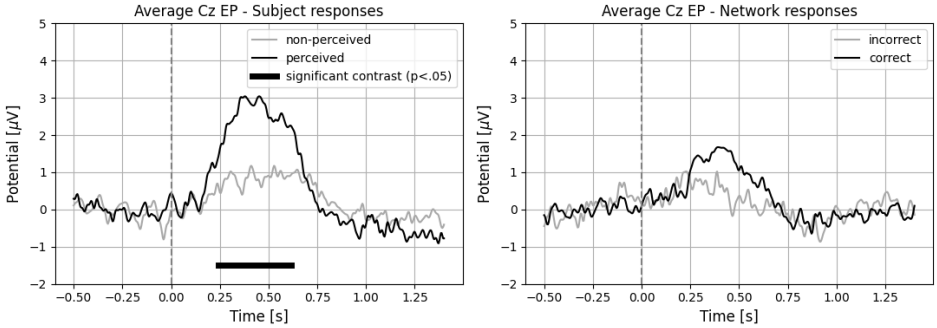


FIGURE 16: Grand average over subjects of the evoked potential at Cz during task 1. The vertical line corresponds to the time point when the stimulus is applied.

Peaks of interest where determined from the GFP of the subject responses for the felt stimuli in Figure 15. The first two early peaks where defined at 12ms and 19ms. The larger, later peak was defined at the maximum point of the GFP, at around 36ms. These time points of interest are additionally plotted as topographies in figure 17, where the expected ERP distribution is observed.

Based on the results obtained in this section, it was of interest to assess which factors might contribute to a better performance of the neural network. An ordinary linear model was built in order to identify possible explanatory variables, however, from the analysis, no significant relations were observed. More details are found on appendix E.

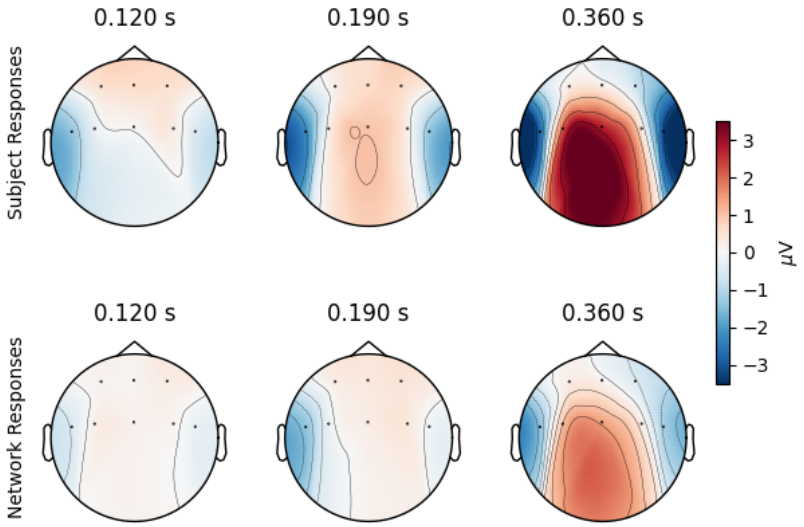


FIGURE 17: Topographical maps of average evoked potentials at three time points of interest.

4.3.2 Task 2

Figure 18 shows the estimated perceptual thresholds, slopes, and drifts for each participant. Measurements corresponding to the same subject are connected with a dotted line to visually explore differences between the two methods. A non-significant difference between the average NDT of the two networks is seen despite the larger mean NDT observed for the second network. Additionally, negative and close to zero ICC values indicate very poor agreement, due to a greater variability within subjects' measurements compared to the variability between different subjects in the same group.

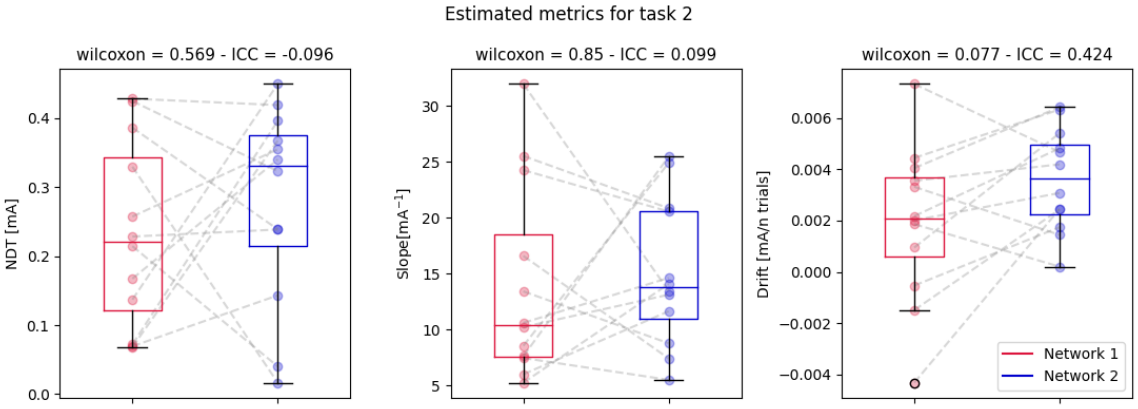


FIGURE 18: Boxplots of the estimated threshold, slope, and drift after fitting a GLM to the data obtained from Task 2. Dotted lines join data points from the same subject.

The differences between the two networks for the estimated metrics are also studied by means of a Bland-Altman plot as shown in Figure 14. Large intervals of confidence are seen in the case of the Threshold and slope. No clear distribution of the difference between methods is seen for the NDT and drift estimates, while

in the case of the slope, higher agreement between the methods seems to be present for mean slopes below 13mA−1

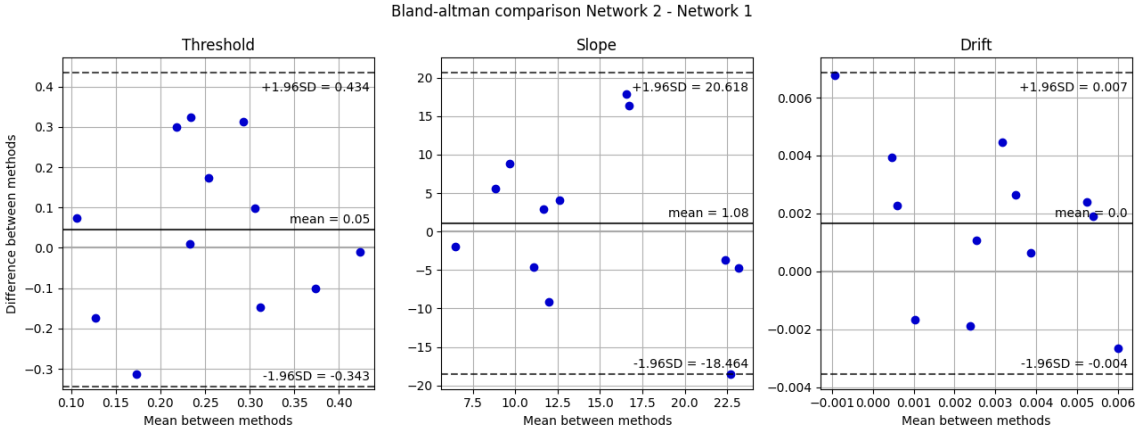


FIGURE 19: Bland-Altman plot comparing the estimated thresholds, slopes and drifts for Task 2.

The reported number of felt stimuli is shown in Figure 20. Two plots represent each half of the experiment, each containing approximately 100 stimuli. Dotted lines connect data points from the same subject. The obtained ICC is also presented on the figure. The average reported count on each half of the experiment was around 65, with some subjects reporting a number of perceived stimuli as high as 99 out of 100.

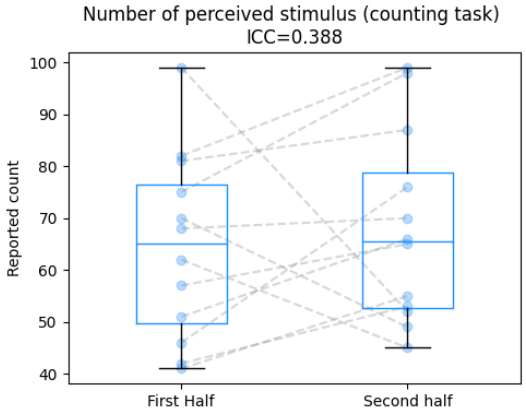


FIGURE 20: Reported number of felt stimuli for each half of the experiment, with dotted lines connecting data points from the same subject.

The EPs generated by the stimulation where depicted in a butterfly plot along with their GFP. Figure 21 contrasts the EPs obtained from the two concurrent. EPs are also presented in figure 22, where only responses of channel Cz are presented. No significant differences were found between "correct" and "incorrect" EPs from the same network. In both correct and incorrect classified stimuli, a clear peak is seen for both networks with the average peaks of network 2 being slightly larger than network 1.

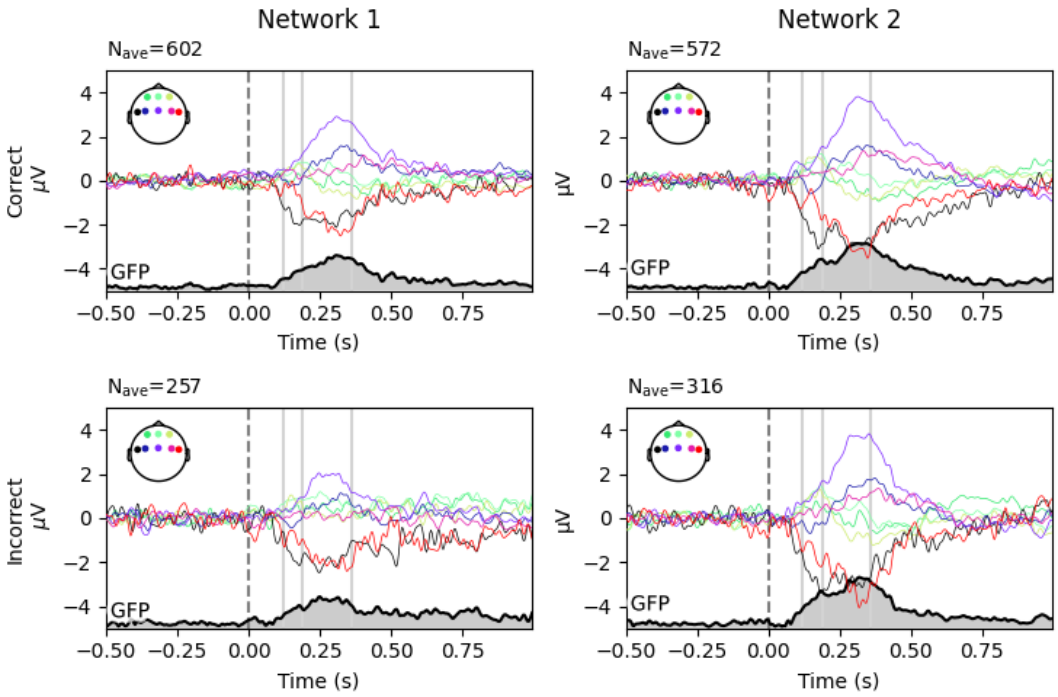


FIGURE 21: Butterfly plot of the grand average potentials and GFP of the EPs resulting from the intra-epidermal stimulation performed in task 2. The vertical dashed line represents the time point when the stimulus is given while the other vertical lines represent the possible time points in which the nociceptive peaks are observed.

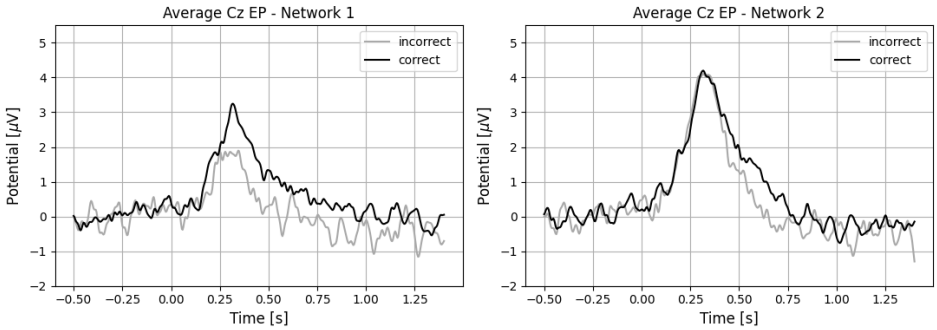


FIGURE 22: Grand average over subjects of the evoked potential at Cz during task 2. The vertical line corresponds to the time point when the stimulus is applied.

The same peaks of interest identified in Figure 15 were chosen to represent the evoked potentials in Figure 23. Here, the expected EP distribution is also seen.

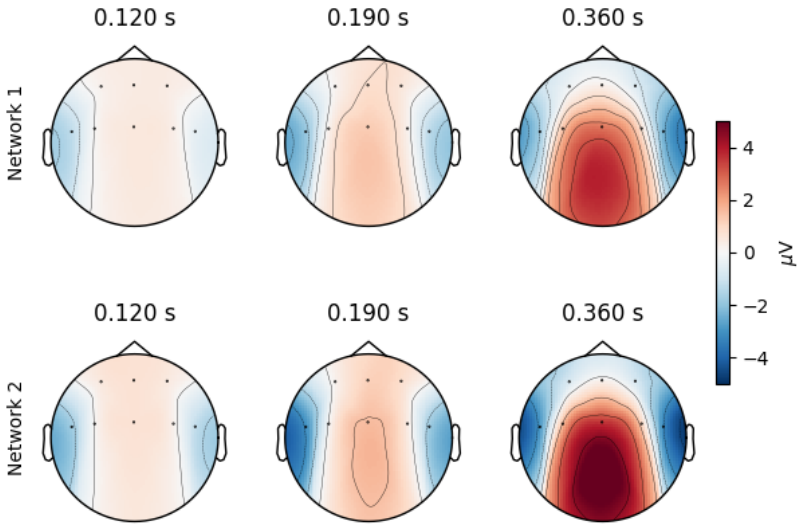


FIGURE 23: Average topographical maps of evoked potentials at three time points of interest.

5 Discussion

5.1 Network selection

The application of the EEG-net architecture to the DS1 dataset using 32 electrodes results in metrics that differ minimally from those reported in [5], with Table 3 showing a 1-2% lower accuracy. These differences could be attributed to randomization effects, such as the use of a different random seed for network initialization and dataset shuffling. Additionally, there are differences in the used datasets; the original study involved 38 participants, while currently, the dataset includes 64 participants. Considering these factors and the minor variation in average accuracy, it is confirmed that the network has been applied correctly.

The minimal performance improvements observed after hyperparameter tuning (as shown in Tables 6 and 7) suggest that the search space for the tuning process may have been too large relative to the computational budget. This could have led to the selection of a suboptimal combination of parameters. However, due to computational constraints, increasing the budget was not feasible. When considering a reduced search space, such as tuning only the window size, a modest increase in average accuracy (about 1%) was observed relative to applying the Inception architecture with its original parameters. Considering also the small differences in accuracy when comparing the original EEGnet parameters proposed by [57] and those proposed in [5], suggests that the impact of hyperparameter tuning on performance was minimal.

A possible limitation to the effect of this strategy is the training duration, which was restricted to 9 epochs. This is considerably fewer than the epochs typically used in training similar networks [57, 59, 54, 66]. This shorter training duration may have prevented the model from fully optimizing, thereby limiting the discovery of the best possible parameter combinations. This effect is particularly evident in the final selected network, where a 2% improvement in accuracy is observed only by increasing training time. However, using a longer training time during a performance tuning strategy can significantly increase the time needed to evaluate one combination of hyperparameters, leading to less combinations being tested.

In the case of transfer learning, a limiting factor was the duration of the epochs in the dataset (one-second), as opposed to the 1.4 seconds deemed optimal by hyperparameter tuning. Another possibility is the similarity between the training and testing datasets. Dag et al. [66] found that transfer learning had minimal effect on accuracy for datasets involving healthy subjects but significantly impacted datasets with disabled subjects. Since this study only used datasets from healthy subjects, transfer learning's impact was likely reduced. Additionally, the training dataset's size was adequate for training a neural network without overfitting; larger accuracy improvements might be observed with smaller training datasets [64].

Regarding ensembles, only one type was tested, more complex strategies such as bagging or stacking might help increasing the accuracy. Using different types of models such as decision trees or support vector machines can more variety of the patterns of the data. Additionally, for an ensemble to perform optimally, independence of the classifier should exist, this however cannot be obtained when the training is done in the same data, as the classifiers are more likely to make similar mistakes [61, 70]. All of the classifiers used in the ensemble presented here were CNNs, limiting the variability of the patterns that are studied, also, since only one training dataset was used, it is possible that some errors related to the differences in testing and training data set are present in all classifiers.

The results indicate that all the attempted strategies, despite the identified limitations, positively impacted performance. A thorough exploration using various approaches was conducted. This provides valuable information about the suitability of each paradigm for the type of data being studied, allowing future research to be better guided. Notably, hyperparameter tuning, while resulting in small changes on its own, led to more substantial performance improvements when combined with a smaller search space, the adoption of a different architecture, and extended training duration. Together, these adjustments led to an overall accuracy increase of 6% in the final selected architecture when tested on DS2.

5.2 Simulations

The conducted simulation provided a basis for justifying the values used for vector length, step, and distance between vectors, ensuring a more confident selection process. It should be noted, however, that there are already expected differences due to the use of a GN or a 2IFC. In [71], comparing subject performance between the two tasks, no significant differences were seen in threshold or slope estimates, but mean slope between methods had a difference of about 7mA-1, with the slope of the GN being higher, this is consistent with the lower slope estimates seen in Figure 10 even with relatively higher agreement probabilities of around 80%.

A main limitation in directly applying these results to neural network accuracies is the constant agreement probability in these estimates. In [5], it was observed that trained networks might perform better in classifying pre-stimulus compared to post-stimulus. Additionally, as shown in Figure 3, the classification in the case of non-perceived stimuli during a 2IFC task, will be mostly influenced by the guessing rate instead of the PF. Therefore, varying agreement probabilities dependent on the initial response derived from the PF and considering the effects of the guessing rate could better model neural network performance. Despite these simplifications, this simulation allows to obtain useful information, aligns with results obtained from literature and provides a base to interpret the obtained accuracy levels and how these might relate to reliable estimates.

5.3 Experimental results

The drop in accuracy seen when testing the selected neural network in the recorded dataset with 8 electrodes should still allow for accurate threshold estimations, according to the conducted simulations. This reduction in accuracy suggests that subsampling the electrodes provides different signal quality compared to recording from only eight. This discrepancy can be explained by the use of CAR during signal recording, which could result in potentially useful information from the other channels being present when subsampling eight electrodes from 32 and in cleaner signals, which might enhance the network's ability to classify the stimuli accurately.

The following sections provide a specific discussion related to the two tasks performed by the subject:

5.3.1 Task 1

The additional test of the neural network's performance, illustrated in Figure 12, provides a better understanding of its behavior during actual experiments. Considering the theory related to the 2IFC task (see section 2.3.1), it is anticipated that when a stimulus is not perceived in a GN (button response 0), the 2IFC classifier should have an equal probability of classifying the stimulus as perceived or not. However, the network shows a bias towards indicating that the stimulus was perceived. Theoretically, assuming perfect network performance, these discrepancies could be attributed to the lapse rate or by a subject's internal criterion, which may result in stimuli being perceived but deemed too weak to report. These phenomena could partially explain the values observed in the confusion matrix, particularly since they suggest an average lapse rate of only seven trials, a plausible number out of 100 trials.

In cases where subjects reported perception of the stimuli (button responses = 1), misclassified trials could be partially attributed to a guessing rate, where a subject accidentally releases the button without actual perception. However, given the setup of the GN procedure, this guessing rate should be extremely low, and the discrepancies are more likely due to errors in the neural network's classification. Therefore, the proportion of correctly classified responses in reported perception cases (84%) might better reflect the network's behavior than the analysis of responses where no perception was reported.

Differences in thresholds are expected as the overall accuracy of the neural network was lower than the standard set by the simulations. Interestingly, based on simulations and results from [71], it would be

expected for the 2IFC threshold estimates to be higher, rather than lower as seen in Figure 13. However, considering the network's bias towards predicting that a stimulus is perceived, as observed in Figure 12, it is logical that the threshold decreases slightly, though not significantly. The low ICC seen between NDT estimates indicates a poor agreement between the methods. A lower ICC could be attributed to random errors in the network's performance, which would be amplified due to subject-specific variations in EP shape and noisy recordings not adequately averaged using CAR. However, a value as low as 0.137 is most likely hinting at a suboptimal performance of the network for the given task.

Regarding the other obtained estimates, a better reliability of the drift suggests a stable behaviour of the network throughout the experiment. The difference in slope between the methods aligns with the conducted simulations, which anticipated lower slopes in case of low accuracy values, additionally due to the inherent differences between GN and 2IFC methods, a lower slope would also be logical as seen in [71].

The Bland-Altman plot (Figure 14)confirmed significant variability by showing large confidence intervals across all metrics when evaluating the agreement between methods. For the threshold metric, larger differences between methods are seen for the data points where the mean between the methods is larger. The plot indicated a tendency for button press thresholds to be higher, as evidenced by a positive mean difference, although this difference was not substantial. In contrast, the distribution of the data points related to the slope metric indicates less agreement at higher mean values. This pattern can be attributed to generally higher slope estimates from subjects compared to the 2IFC task, with only 2 subjects presenting higher slopes with the 2IFC.

One of the most noticeable aspects in the butterfly plot of the collected EPs from both procedures (see figure 15) is the difference in amplitudes of the responses of the perceived stimuli of the GN compared to the 2IFC method. Van den Berg et al. (2022) reported that the EPs from 2IFC trials were about half the amplitude of those from GN procedures when analyzed at Cz-M1M2 [71]. In the mentioned study, the average P2 peak amplitude for GN reached approximately 16µV, whereas for 2IFC, it was around 8µV. Comparing these results to another study by the same author [43], where an average reference was used instead of mastoids, the P2 peak amplitudes were similar to those obtained in this study, around 3µV for the GN task. These findings underscore the importance of the reference electrode, as M1 and M2 potentially reveal critical information. Furthermore, the reduction in amplitude observed from GN to 2IFC was consistent with the twofold reduction expected from the literature. However, due to the difference in reference and probably due to the inclusion of many false positives, the contrast between epochs classified as correct and incorrect did not reach statistical significance (Figure 16).

For non-perceived stimuli, a similar overall behavior is seen in the butterfly plot, except for the lateral electrodes (T6 and T7), which exhibit slightly larger post-stimulus activity in the 2IFC task compared to the GN. When isolating the Cz electrode (Figure 16), a small peak at the P2 location is evident even in the non-perceived responses of the GN procedure. This could indicate that the small amplitude stimuli still elicited a response, or it might suggest subjects missed reporting some stimuli. In contrast, for non-perceived stimuli in the 2IFC, this peak is absent, and the potential for perceived stimuli is significantly smaller, similar in amplitude to the non-perceived GN stimuli. This suggests that, while some differences could be attributed to factors other than network performance, the low EP for perceived stimuli in 2IFC indicates suboptimal network classification.

Topographical maps of the EPs show expected patterns when comparing with a similar study [43], with central activity for P2 and bilateral differences between electrodes such as F4 and T7. However, the bilateral behavior appears more pronounced in the 2IFC procedure compared to GN.

In summary, the comparison between the network's performance and subject responses reveals that, although the network behaves as expected, it exhibits a bias towards classifying stimuli as perceived, which compromises its reliability. This bias affects the accuracy of threshold estimates, which, despite not being

significantly different on average, show high variability and poor reliability, as indicated by large confidence intervals and low ICC. The EP analysis supports these findings, with observed peak amplitudes showing no significant difference between epochs classified as correct and incorrect. However, it is important to note that expected EP shapes are generally observed, demonstrating that the network is capable of capturing relevant neural activity. Despite the challenges, the analysis underscores the potential for improvement, particularly in addressing network bias to achieve a more reliable classification.

5.3.2 Task 2

A previous study [71] and the results from the current task indicate that mean threshold estimates for the studied age group should be around 0.2 mA or lower. In Task 2, the NDT estimates are slightly higher but remain within a reasonable range (Figure 18). However, poor reliability across all metrics, as evidenced by low ICC values, suggests suboptimal network performance. Despite a higher average NDT and drift for Network 2, the high variability between subjects and groups makes it difficult to conclusively determine if this is due to a network bias or chance. The Bland-Altman plot (Figure 19) further highlights this variability, showing large confidence intervals and a small positive mean difference between methods. While this might suggest a slight bias towards the second network, the magnitude and variability make it uncertain whether this bias would persist with a larger sample size.

Additionally, the high count of reported stimuli suggests inaccuracies in the network's classification, resulting in elevated threshold estimates. The inconsistency in behavior throughout the task is further reflected in the low reliability between the two halves of the experiment.

Regarding the EPs, the peaks are higher than in the previous 2IFC task, confirming that subjects were stimulated above their perception threshold. The presence of clear peaks in incorrectly classified epochs and the non-significant differences between EPs further emphasize the network's poor classification performance during the task. Despite these issues, the topographical maps of the EPs show the expected patterns.

5.4 Comparison of Neural Network performance across tasks

Noticeable differences in neural network behavior are observed between tasks. In Task 1, the neural network was closer to the threshold, potentially underestimating it. In contrast, in Task 2, the thresholds were overestimated, as evidenced by a count of reported stimuli higher than expected from the type of task and generally higher threshold values than anticipated for the studied population.

These findings are further corroborated by the analysis of the EPs. Although the EPs from the 2IFC task in Task 1 were lower than expected, the presence of a peak for detected stimuli and the absence of such a peak for undetected stimuli indicate that the network was performing the classification with a high rate of false positives. Conversely, in Task 2, clear and high peaks are observed for both perceived and non-perceived stimuli, confirming poor classification performance that resulted in above-threshold stimulation.

These differences indicate that the network's ability to detect stimuli is highly influenced by the physical task of pressing a button. While this outcome is understandable given that the networks were trained exclusively on responses where subjects indicated perception with a button press, it was not anticipated. In a similar setup used by van den Berg (2022), involving 32 electrodes, the network had a satisfactory performance without a concurrent physical task, despite being trained on a dataset with similar characteristics. These results underscore the importance of a more thorough selection of EEG electrodes.

5.5 Implications

The observed classification performance issues may stem from various factors, which cannot be precisely identified with the available data. Factors such as the presence of a physical task component in the EPs from

the training set, the reference used for recording, the choice of electrodes, and the hyperparameter tuning strategy all call for more research to enhance the reliability of automatic classification estimates.

The network's worse performance without a concurrent physical task suggests that the training set or selected electrodes are not optimal. Future work should focus on the EEG responses from the button press and how that reflects on the EPs to better select the EEG channels and address the training issue. Regarding network improvement, all chosen strategies did result in an improvement, but limited computational power prevented a thorough exploration of the hyperparameter space. Further study into the most influential hyperparameters would better guide decisions on which to tune.

Signal quality is another limiting factor when using only eight electrodes. Common post-hoc analysis techniques, such as ICA and regression, were not feasible due to the limited number of signals and the absence of an EOG recording. Exploring the impact of alternative references on signal quality could be valuable. The average reference used in this study may not be optimal, as it assumes that the surface potential integral of a dipole in a volume conductor is zero, which does not hold with only eight midline electrodes. Possible alternatives used in literature include placing the reference at FPz [59] and at Afz [66].

Additionally, explaining the network scores, which was beyond the scope of this study, provides a useful point for further research. Algorithms such as occlusion sensitivity, which tests the effect of occluding input features, and Layerwise Relevance Propagation, which traces activations through the network layers to the output and back, could reveal what the network focuses on during classification [72]. This could inform decisions regarding channel selection and potentially reveal whether the network uses information other than nociceptive evoked EPs. Exploring these methods could enhance our understanding and improve the network's accuracy in detecting nociceptive signals.

6 Conclusion

In this thesis, significant steps were made towards the clinical application of automatic nociceptive threshold classification. Building on the encouraging results previously presented by van den Berg (2022) [5], this study explored the feasibility of reducing the number of EEG channels needed.

While is was possible to observe an increased accuracy of the previously used CNN by 6%, several shortcomings were evident during the experimental procedures. The neural network's accuracy, compared to a concurrent GN procedure, showed only a small bias towards classifying stimuli as perceived. Although the average thresholds obtained from the GN and 2IFC tasks were not significantly different, the poor correlation between methods demonstrated that the network's performance during this task was not optimal. Additionally, a smaller than expected peak amplitude in the obtained EP along with non-significant contrasts between EPs classified as correct and incorrect confirmed the presence of many false positives. When analyzing the neural network without a concurrent button press, a very low correlation between the two estimated thresholds demonstrated that the network could not reliably estimate the threshold. Furthermore, clear nociceptive EPs seen when the network classified a stimulus as non-perceived proved the network's low ability to distinguish between perceived and non-perceived stimuli.

Considering these findings, it was determined that with the current setup, automatic tracking of the nociceptive detection threshold is not possible. Despite this, several areas for improvement were identified, providing possible directions for future research. More attention should be given to electrode selection by examining the network's activities and the information it uses. Additionally, exploring the possibility of using different references during recording could improve the quality of the recorded signals.

Moreover, this thesis explored several methods for improving neural network performance. Future research should focus on optimizing these strategies, such as refining hyperparameter tuning to target a smaller, more influential set of parameters or applying transfer learning with a large dataset followed by fine-tuning with a smaller, more relevant dataset.

Despite the less-than-ideal results presented here, it provides confidence on the next steps to follow. The ongoing advancements in artificial intelligence and its applications in BCIs offer an expanding array of tools that might help to enhance the methods explored in this work. Achieving automatic tracking of nociceptive thresholds can significantly contribute to objective pain measurement, benefiting a broader range of patients with chronic pain. This advancement could lead to a more comprehensive understanding of pain, improved evaluation methods, and ultimately, better treatment options.

References

[1] Srinivasa N. Raja et al. "The Revised IASP definition of pain: concepts, challenges, and compromises". In: *Pain* 161.9 (Sept. 2020), p. 1976. ISSN: 18726623. DOI: 10.1097/J.PAIN.0000000000001939. URL: /pmc/articles/PMC7680716/%20/pmc/articles/PMC7680716/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7680716/.

[2] Markus Ploner, Christian Sorg, and Joachim Gross. "Brain Rhythms of Pain". In: *Trends in Cognitive Sciences* 21.2 (Feb. 2017), p. 100. ISSN: 1879307X. DOI: 10.1016/J.TICS.2016.12.001. URL: /pmc/articles/PMC5374269/%20/pmc/articles/PMC5374269/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5374269/.

[3] Duo Chen et al. "Scalp EEG-Based Pain Detection Using Convolutional Neural Network". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 30 (2022), pp. 274–285. ISSN: 15580210. DOI: 10.1109/TNSRE.2022.3147673.

[4] Mahmoud Elsayed, Kok Swee Sim, and Shing Chiang Tan. "A novel approach to objectively quantify the subjective perception of pain through electroencephalogram signal analysis". In: *IEEE Access* 8 (2020), pp. 199920–199930. ISSN: 21693536. DOI: 10.1109/ACCESS.2020.3032153.

[5] Boudewijn van den Berg et al. "Real-time estimation of perceptual thresholds based on the electroencephalogram using a deep neural network". In: *Journal of Neuroscience Methods* 374 (May 2022). ISSN: 1872678X. DOI: 10.1016/j.jneumeth.2022.109580.

[6] Robert J. Doll et al. "Tracking of nociceptive thresholds using adaptive psychophysical methods". In: *Behavior Research Methods* 46.1 (July 2014), pp. 55–66. ISSN: 15543528. DOI: 10.3758/S13428-013-0368-4/TABLES/5. URL: https://link.springer.com/article/10.3758/s13428-013-0368-4.

[7] A. Mouraux, G. D. Iannetti, and L. Plaghki. "Low intensity intra-epidermal electrical stimulation can activate Aδ-nociceptors selectively". In: *Pain* 150.1 (July 2010), pp. 199–207. ISSN: 03043959. DOI: 10.1016/J.PAIN.2010.04.026. URL: https://journals.lww.com/pain/fulltext/2010/07000/low_intensity_intra_epidermal_electrical.29.aspx.

[8] Vishal Vijayakumar et al. "Quantifying and Characterizing Tonic Thermal Pain across Subjects from EEG Data using Random Forest Models". In: *IEEE transactions on bio-medical engineering* 64.12 (Dec. 2017), p. 2988. ISSN: 15582531. DOI: 10.1109/TBME.2017.2756870. URL: /pmc/articles/PMC5718188/%20/pmc/articles/PMC5718188/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5718188/.

[9] D Purves. "Neuroscience, 3rd Edition". In: ().

[10] Biji Bahuleyan, Tatiana von Hertwig Fernandes de Oliveira, and Andre G. Machado. "Chronic Pain, Failed Back Surgery Syndrome, and Management". In: *Benzel's Spine Surgery: Techniques, Complication Avoidance and Management: Volume 1-2, Fourth Edition* 1-2 (Jan. 2017), pp. 1548–1559. DOI: 10.1016/B978-0-323-40030-5.00177-5.

[11] Sarah Kendroud et al. "Physiology, Nociceptive Pathways". In: *StatPearls* (Sept. 2022). URL: https://www.ncbi.nlm.nih.gov/books/NBK470255/.

[12] Robert C. Coghill. "The Distributed Nociceptive System: A Framework for Understanding Pain". In: *Trends in Neurosciences* 43.10 (Oct. 2020), pp. 780–794. ISSN: 0166-2236. DOI: 10.1016/J.TINS.2020.07.004.

[13] John D. Loeser and Ronald Melzack. "Pain: an overview". In: *The Lancet* 353.9164 (May 1999), pp. 1607–1609. ISSN: 0140-6736. DOI: 10.1016/S0140-6736(99)01311-2.

[14] Richard P. Szumita, Paul M. Szumita, and Nancy Just. "Understanding and Managing Patients with Chronic Pain". In: *Oral and Maxillofacial Surgery Clinics of North America* 22.4 (Nov. 2010), pp. 481–494. ISSN: 10423699. DOI: 10.1016/j.coms.2010.07.005.

[15] Ronald Melzack. "The future of pain". In: *Nature Reviews Drug Discovery 2008 7:8* 7.8 (2008), pp. 629–629. ISSN: 1474-1784. DOI: 10.1038/nrd2640. URL: https://www.nature.com/articles/nrd2640.

[16] Ronald Melzack. "From the gate to the neuromatrix". In: *Pain* Suppl 6.SUPPL.1 (1999). ISSN: 0304-3959. DOI: 10.1016/S0304-3959(99)00145-1. URL: https://pubmed.ncbi.nlm.nih.gov/10491980/.

[17] Disability Institute of Medicine (US) Committee on Pain et al. "The Anatomy and Physiology of Pain". In: (1987). URL: https://www.ncbi.nlm.nih.gov/books/NBK219252/.

[18] Mun Fei Yam et al. "General Pathways of Pain Sensation and the Major Neurotransmitters Involved in Pain Regulation". In: *International Journal of Molecular Sciences* 19.8 (Aug. 2018). ISSN: 14220067. DOI: 10.3390/IJMS19082164. URL: /pmc/articles/PMC6121522/%20/pmc/articles/PMC6121522/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6121522/.

[19] Matheus Deroco Veloso da Silva et al. "Stem cells and pain". In: *World journal of stem cells* 15.12 (2023), pp. 1035–1062. ISSN: 1948-0210. DOI: 10.4252/WJSC.V15.I12.1035. URL: https://pubmed.ncbi.nlm.nih.gov/38179216/.

[20] Regina Fink. "Pain assessment: the cornerstone to optimal pain management". In: *Proceedings (Baylor University. Medical Center)* 13.3 (July 2000), p. 236. ISSN: 0899-8280. DOI: 10.1080/08998280.2000.11927681. URL: /pmc/articles/PMC1317046/%20/pmc/articles/PMC1317046/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1317046/.

[21] Richard H. Gracely. "Pain measurement". In: *Acta Anaesthesiologica Scandinavica* 43.9 (1999), pp. 897–908. ISSN: 00015172. DOI: 10.1034/J.1399-6576.1999.430907.X.

[22] Ozgur Karcioglu et al. "A systematic review of the pain scales in adults: Which to use?" In: *The American Journal of Emergency Medicine* 36.4 (Apr. 2018), pp. 707–714. ISSN: 0735-6757. DOI: 10.1016/J.AJEM.2018.01.008.

[23] R. Cowen et al. "Assessing pain objectively: the use of physiological markers". In: *Anaesthesia* 70.7 (July 2015), pp. 828–847. ISSN: 1365-2044. DOI: 10.1111/ANAE.13018. URL: https://onlinelibrary.wiley.com/doi/full/10.1111/anae.13018%20https://onlinelibrary.wiley.com/doi/abs/10.1111/anae.13018%20https://associationofanaesthetists-publications.onlinelibrary.wiley.com/doi/10.1111/anae.13018.

[24] Zakir Uddin and Joy C. Macdermid. "Quantitative Sensory Testing in Chronic Musculoskeletal Pain". In: *Pain Medicine* 17.9 (Sept. 2016), pp. 1694–1703. ISSN: 1526-2375. DOI: 10.1093/PM/PNV105. URL: https://dx.doi.org/10.1093/pm/pnv105.

[25] Shengai Li, Danielle H. Melton, and Sheng Li. "Tactile, thermal, and electrical thresholds in patients with and without phantom limb pain after traumatic lower limb amputation". In: *Journal of Pain Research* 8 (Apr. 2015), pp. 169–174. ISSN: 11787090. DOI: 10.2147/JPR.S77412. URL: https://www.dovepress.com/tactile-thermal-and-electrical-thresholds-in-patients-with-and-without-peer-reviewed-fulltext-article-JPR.

[26] Doeke Keizer et al. "Quantitative sensory testing with von frey monofilaments in patients with allodynia what are we quantifying?" In: *Clinical Journal of Pain* 24.5 (June 2008), pp. 463–466. ISSN: 07498047. DOI: 10.1097/AJP.0B013E3181673B80. URL: https://journals.lww.com/clinicalpain/fulltext/2008/06000/quantitative_sensory_testing_with_von_frey.14.aspx.

[27] Geoffrey Bove. "Mechanical sensory threshold testing using nylon monofilaments: The pain field's "Tin Standard"". In: *Pain* 124.1-2 (Sept. 2006), pp. 13–17. ISSN: 0304-3959. DOI: 10.1016/J.PAIN.2006.06.020.

[28] Priyanka Iyer and Yvonne C. Lee. "Why It Hurts: The Mechanisms of Pain in Rheumatoid Arthritis". In: *Rheumatic Disease Clinics of North America* 47.2 (May 2021), pp. 229–244. ISSN: 0889-857X. DOI: 10.1016/J.RDC.2020.12.008.

[29] Anit Bhattacharyya et al. "The reliability of pressure pain threshold in individuals with low back or neck pain: a systematic review". In: *British Journal of Pain* 17.6 (Dec. 2023), pp. 579–591. ISSN: 20494645. DOI: 10.1177/20494637231196647/ASSET/IMAGES/LARGE/10.1177{_}20494637231196647-FIG1.JPEG. URL: https://journals.sagepub.com/doi/10.1177/20494637231196647.

[30] Monica Sean et al. "Comparison of Thermal and Electrical Modalities in the Assessment of Temporal Summation of Pain and Conditioned Pain Modulation". In: *Frontiers in Pain Research* 2 (Sept. 2021), p. 659563. ISSN: 2673561X. DOI: 10.3389/FPAIN.2021.659563/BIBTEX. URL: www.frontiersin.org.

[31] Boudewijn van den Berg et al. "Simultaneous tracking of psychophysical detection thresholds and evoked potentials to study nociceptive processing". In: *Behavior Research Methods* 52.4 (Aug. 2020), p. 1617. ISSN: 15543528. DOI: 10.3758/S13428-019-01338-7. URL: /pmc/articles/PMC7406487/%20/pmc/articles/PMC7406487/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7406487/.

[32] Frederick A.A. Kingdom and Nicolaas Prins. *Psychophysics: A Practical Introduction*. Elsevier, Jan. 2016, pp. 1–331. ISBN: 9780124071568. DOI: 10.1016/B978-0-12-407156-8.01001-X. URL: http://www.sciencedirect.com:5070/book/9780124071568/psychophysics.

[33] George A. Gescheider. "Psychophysics : the fundamentals". In: (1997), p. 435. URL: https://books.google.com/books/about/Psychophysics.html?hl=es&id=fLYWFcuamPwC.

[34] Boudewijn van den Berg and Jan R. Buitenweg. "Observation of Nociceptive Processing: Effect of Intra-Epidermal Electric Stimulus Properties on Detection Probability and Evoked Potentials". In: *Brain Topography* 34.2 (Mar. 2021), pp. 139–153. ISSN: 15736792. DOI: 10.1007/S10548-020-00816-Y/TABLES/5. URL: https://link.springer.com/article/10.1007/s10548-020-00816-y.

[35] Simon Grondin. *Psychology of Perception*. Université Laval , Québec , Canada: Springer International, 2016.

[36] Robert J. Doll, Peter H. Veltink, and Jan R. Buitenweg. "Observation of time-dependent psychophysical functions and accounting for threshold drifts". In: *Attention, Perception & Psychophysics* 77.4 (May 2015), p. 1440. ISSN: 1943393X. DOI: 10.3758/S13414-015-0865-X. URL: /pmc/articles/PMC4415976/%20/pmc/articles/PMC4415976/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4415976/.

[37] In Jae Myung. "Tutorial on maximum likelihood estimation". In: *Journal of Mathematical Psychology* 47.1 (Feb. 2003), pp. 90–100. ISSN: 0022-2496. DOI: 10.1016/S0022-2496(02)00028-7.

[38] Brian Everitt. "Generalized Linear Models (GLM)". In: *Encyclopedia of Statistics in Behavioral Science* (Apr. 2005). DOI: 10.1002/0470013192.BSA252. URL: https://onlinelibrary.wiley.com/doi/full/10.1002/0470013192.bsa252%20https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013192.bsa252%20https://onlinelibrary.wiley.com/doi/10.1002/0470013192.bsa252.

[39] J. J. Faraway. "Generalized Linear Models". In: *International Encyclopedia of Education, Third Edition* (Jan. 2010), pp. 178–183. DOI: 10.1016/B978-0-08-044894-7.01331-2.

[40] Vartika Gupta et al. "Comparative Performance Analysis of Scalp EEG and Ear EEG based P300 Ambulatory Brain-Computer Interfaces using Riemannian Geometry and Convolutional Neural Networks". In: *2022 National Conference on Communications, NCC 2022* (2022), pp. 314–319. DOI: 10.1109/NCC55593.2022.9806815.

[41] Thibaut Mussigmann, Benjamin Bardel, and Jean Pascal Lefaucheur. "Resting-state electroencephalography (EEG) biomarkers of chronic neuropathic pain. A systematic review". In: *NeuroImage* 258 (Sept. 2022), p. 119351. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2022.119351.

[42] Markus Ploner and Elisabeth S. May. "Electroencephalography and magnetoencephalography in pain research - Current state and future perspectives". In: *Pain* 159.2 (Feb. 2018), pp. 206–211. ISSN: 18726623. DOI: 10.1097/J.PAIN.0000000000001087. URL: https://journals.lww.com/pain/fulltext/2018/02000/electroencephalography_and_magnetoencephalography.4.aspx.

[43] Boudewijn van den Berg and Jan R. Buitenweg. "Observation of Nociceptive Processing: Effect of Intra-Epidermal Electric Stimulus Properties on Detection Probability and Evoked Potentials". In: *Brain Topography* 34.2 (Mar. 2021), p. 139. ISSN: 15736792. DOI: 10.1007/S10548-020-00816-Y. URL: /pmc/articles/PMC7892744/%20/pmc/articles/PMC7892744/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7892744/.

[44] Jair Stern, Daniel Jeanmonod, and Johannes Sarnthein. "Persistent EEG overactivation in the cortical pain matrix of neurogenic pain patients". In: *NeuroImage* 31.2 (June 2006), pp. 721–731. ISSN: 1053-8119. DOI: 10.1016/J.NEUROIMAGE.2005.12.042.

[45] Wajid Mumtaz, Suleman Rasheed, and Alina Irfan. "Review of challenges associated with the EEG artifact removal methods". In: *Biomedical Signal Processing and Control* 68 (July 2021), p. 102741. ISSN: 1746-8094. DOI: 10.1016/J.BSPC.2021.102741.

[46] Qi Li et al. "A P300-Detection Method Based on Logistic Regression and a Convolutional Neural Network". In: *Frontiers in Computational Neuroscience* 16 (June 2022), p. 909553. ISSN: 16625188. DOI: 10.3389/FNCOM.2022.909553/BIBTEX. URL: www.frontiersin.org.

[47] Alexander Craik, Yongtian He, and Jose L Contreras-Vidal. "Deep learning for electroencephalogram (EEG) classification tasks: a review". In: *J. Neural Eng* 16.3 (2019). DOI: 10.1088/1741-2552/ab0ab5. URL: https://doi.org/10.1088/1741-2552/ab0ab5.

[48] Arnaud Delorme. "EEG is better left alone". In: *Scientific reports* 13.1 (Dec. 2023). ISSN: 2045-2322. DOI: 10.1038/S41598-023-27528-0. URL: https://pubmed.ncbi.nlm.nih.gov/36759667/.

[49] Phillip M. Alday. "How much baseline correction do we need in ERP research? Extended GLM model can replace baseline correction while lifting its limits". In: *Psychophysiology* 56.12 (Dec. 2019). ISSN: 1469-8986. DOI: 10.1111/PSYP.13451. URL: https://pubmed.ncbi.nlm.nih.gov/31403187/.

[50] Federico Chella et al. "Impact of the reference choice on scalp EEG connectivity estimation". In: *Journal of Neural Engineering* 13.3 (May 2016), p. 036016. ISSN: 1741-2552. DOI: 10.1088/1741-2560/13/3/036016. URL: https://iopscience.iop.org/article/10.1088/1741-2560/13/3/036016%20https://iopscience.iop.org/article/10.1088/1741-2560/13/3/036016/meta.

[51] C. R. Rashmi and C. P. Shantala. "EEG artifacts detection and removal techniques for brain computer interface applications: a systematic review". In: *International Journal of Advanced Technology and Engineering Exploration* 9.88 (Mar. 2022), pp. 354–383. ISSN: 23947454. DOI: 10.19101/IJATEE.2021.874883. URL: https://www.researchgate.net/publication/360670106_EEG_artifacts_detection_and_removal_techniques_for_brain_computer_interface_applications_a_systematic_review.

[52] Mainak Jas et al. "Autoreject: Automated artifact rejection for MEG and EEG data". In: *NeuroImage* 159 (Oct. 2017), pp. 417–429. ISSN: 1095-9572. DOI: 10.1016/J.NEUROIMAGE.2017.06.030. URL: https://pubmed.ncbi.nlm.nih.gov/28645840/.

[53] Sinam Ajitkumar Singh et al. "A deep neural network approach for P300 detection-based BCI using single-channel EEG scalogram images". In: *Physical and engineering sciences in medicine* 44.4 (Dec. 2021), pp. 1221–1230. ISSN: 2662-4737. DOI: 10.1007/S13246-021-01057-4. URL: https://pubmed.ncbi.nlm.nih.gov/34550551/.

[54] Meng Xu et al. "A deep learning method for single-trial EEG classification in RSVP task based on spatiotemporal features of ERPs". In: *Journal of Neural Engineering* 18.4 (Aug. 2021), p. 0460c8. ISSN: 1741-2552. DOI: 10.1088/1741-2552/AC1610. URL: https://iopscience.iop.org/article/10.1088/1741-2552/ac1610%20https://iopscience.iop.org/article/10.1088/1741-2552/ac1610/meta.

[55] Seyed Vahab Shojaedini, Sajedeh Morabbi, and Mohammad Reza Keyvanpour. "A new method for detecting P300 signals by using deep learning: Hyperparameter tuning in high-dimensional space by minimizing nonconvex error function". In: *Journal of Medical Signals and Sensors* 8.4 (Oct. 2018), pp. 205–214. ISSN: 22287477. DOI: 10.4103/JMSS.JMSS{_}7{_}18. URL: https://journals.lww.com/jmss/fulltext/2018/08040/a_new_method_for_detecting_p300_signals_by_using.1.aspx.

[56] Endre Pap, ed. *Artificial Intelligence: Theory and Applications*. Vol. 973. Studies in Computational Intelligence. Cham: Springer International Publishing, 2021. ISBN: 978-3-030-72710-9. DOI: 10.1007/978-3-030-72711-6. URL: https://link.springer.com/10.1007/978-3-030-72711-6.

[57] Ziwei Wang et al. "EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces You may also like Unsupervised domain adaptation for cross-patient seizure classification-

Multi-source deep domain adaptation ensemble framework for cross-dataset motor imagery EEG transfer learning EEGNet: a compact convolutional neural network for EEG-based brain-computer interfaces". In: *J. Neural Eng* 15 (2018), pp. 56013–56030. DOI: 10.1088/1741-2552/aace8c. URL: https://doi.org/10.1088/1741-2552/aace8c.

[58] Haifeng Zhao et al. "Can recurrent neural network enhanced EEGNet improve the accuracy of ERP classification task? An exploration and a discussion". In: *Health and Technology* 10.4 (July 2020), pp. 979–995. ISSN: 21907196. DOI: 10.1007/S12553-020-00458-X/TABLES/4. URL: https://link.springer.com/article/10.1007/s12553-020-00458-x.

[59] Eduardo Santamaria-Vazquez et al. "EEG-Inception: A Novel Deep Convolutional Neural Network for Assistive ERP-Based Brain-Computer Interfaces". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 28.12 (Dec. 2020), pp. 2773–2782. ISSN: 15580210. DOI: 10.1109/TNSRE.2020.3048106.

[60] Louis Owen. *Hyperparameter Tuning with Python*. Ed. by David Sugarman and Devanshi Ayare. Birmingham: Packt Publishing Ltd., 2022.

[61] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*. Ed. by Nicole Tache. 2nd ed. O'Reilly Media, Inc, 2019.

[62] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[63] Stefan Falkner, Aaron Klein, and Frank Hutter. "BOHB: Robust and Efficient Hyperparameter Optimization at Scale". In: (2018).

[64] Jason Yosinski et al. "How transferable are features in deep neural networks?" In: *Advances in Neural Information Processing Systems* 27 (2014).

[65] Ahmad Waleed Salehi et al. "A Study of CNN and Transfer Learning in Medical Imaging: Advantages, Challenges, Future Scope". In: *Sustainability 2023, Vol. 15, Page 5930* 15.7 (Mar. 2023), p. 5930. ISSN: 2071-1050. DOI: 10.3390/SU15075930. URL: https://www.mdpi.com/2071-1050/15/7/5930/htm%20https://www.mdpi.com/2071-1050/15/7/5930.

[66] Ihsan Da et al. "Leveraging Deep Learning Techniques to Improve P300-Based Brain Computer Interfaces". In: *IEEE journal of biomedical and health informatics* 26.10 (Oct. 2022), pp. 4892–4902. ISSN: 2168-2208. DOI: 10.1109/JBHI.2022.3174771. URL: https://pubmed.ncbi.nlm.nih.gov/35552154/.

[67] Berdakh Abibullaev and Amin Zollanvari. "A Systematic Deep Learning Model Selection for P300-Based Brain-Computer Interfaces". In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 52.5 (May 2022), pp. 2744–2756. ISSN: 21682232. DOI: 10.1109/TSMC.2021.3051136.

[68] Terry K. Koo and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research". In: *Journal of Chiropractic Medicine* 15.2 (June 2016), p. 155. ISSN: 15563707. DOI: 10.1016/J.JCM.2016.02.012. URL: /pmc/articles/PMC4913118/%20/pmc/articles/PMC4913118/?report=abstract%20https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4913118/.

[69] Micah M. Murray, Denis Brunet, and Christoph M. Michel. "Topographic ERP Analyses: A Step-by-Step Tutorial Review". In: *Brain Topography* 20.4 (June 2008), pp. 249–264. ISSN: 0896-0267. DOI: 10.1007/s10548-008-0054-5.

[70] Terry Windeatt. "Ensemble MLP Classifier Design". In: *Studies in Computational Intelligence* 137 (2008), pp. 133–147. ISSN: 1860-9503. DOI: 10.1007/978-3-540-79474-5{_}6. URL: https://link.springer.com/chapter/10.1007/978-3-540-79474-5_6.

[71] Boudewijn van den Berg et al. "Observation of nociceptive detection thresholds and cortical evoked potentials: Go/no-go versus two-interval forced choice". In: *Attention, Perception, and Psychophysics* 84.4 (May 2022), pp. 1359–1369. ISSN: 1943393X. DOI: 10.3758/S13414-022-02484-5/FIGURES/8. URL: https://link.springer.com/article/10.3758/s13414-022-02484-5.

[72] Wojciech Samek et al. "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications". In: *Proceedings of the IEEE* 109.3 (Mar. 2021), pp. 247–278. ISSN: 15582256. DOI: 10.1109/JPROC.2021.3060483.

[73] Michael Olusegun Akinwande, Hussaini Garba Dikko, and Agboola Samson. "Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis". In: *Open Journal of Statistics* 05.07 (2015), pp. 754–767. ISSN: 2161-718X. DOI: 10.4236/OJS.2015.57075. URL: http://file.scirp.org/Html/%20http://www.scirp.org/journal/PaperInformation.aspx?PaperID=62189&#abstract.

A Architectures

EEGNet [57]

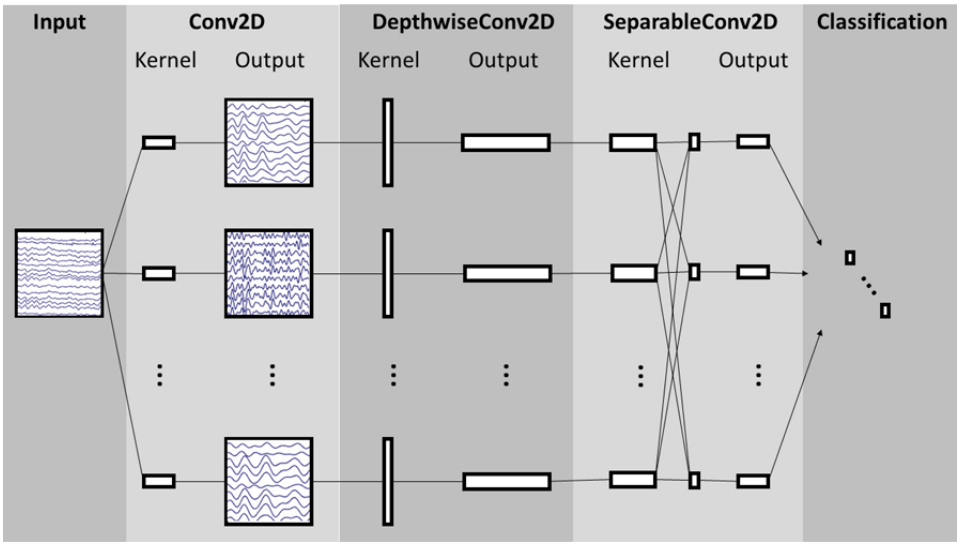


FIGURE 24: Architecture for EEGnet.

Block	Layer	# filters	Size	Output	Options
1	Input			(C, T)	
	Reshape			(1, C, T)	
	Conv2D	F1	(1, 64)	(F1, C, T)	activation = Linear, mode=same
	BatchNorm			(F1, C, T)	
	DepthwiseConv2D	D × F1	(C, 1)	(D × F1, 1, T)	mode = valid, depth = D, max norm = 1, activation=Linear
	BatchNorm			(D × F1, 1, T)	
	Activation			(D × F1, 1, T)	activation = ELU
	AveragePool2D	(1, 4)		(D × F1, 1, T //4)	
	Dropout			(D × F1, 1, T //4)	p = 0.25 or p = 0.5
2	SeparableCon2D	F2	(1, 16)	(F2, 1, T //4)	mode= same, activation=linear
	BatchNorm			(F2, 1, T //4)	
	Activation			(F2, 1, T //4)	activation=ELU
	AveragePool2D		(1, 8)	(F2, 1, T //32)	
	Dropout			(F2, 1, T //32)	p = 0.25 or p = 0.5
	Flatten			(F2 × (T //32))	
Classifier	Dense			N	max norm=0.25, activation = softmax

TABLE 10: EEGnet parameters, as presented in [57]

EEG-Inception [59]

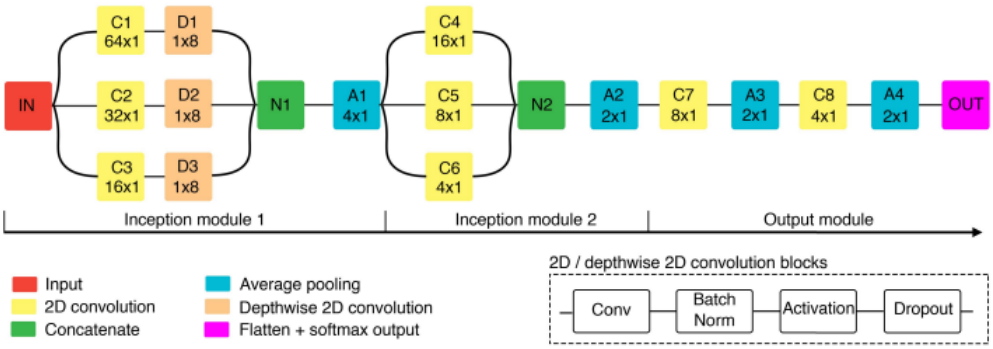


FIGURE 25: Architecture for EEG-Inception.

Block	Type	# filters	Depth	Size	Padding	Output	Connected to	Role
IN	Input	–	–	–	–	128 × 8 × 1	–	Input
C1	Conv2D	8	–	64 × 1	Same	128 × 8 × 8	D1	Temporal analysis
D1	DepthwiseConv2D	–	2	1 × 8	Valid	128 × 8 × 16	N1	Spatial analysis
C2	Conv2D	8	–	32 × 1	Same	128 × 8 × 8	D2	Temporal analysis
D2	DepthwiseConv2D	–	2	1 × 8	Valid	128 × 8 × 16	N1	Spatial analysis
C3	Conv2D	8	–	16 × 1	Same	128 × 8 × 8	D3	Temporal analysis
D3	DepthwiseConv2D	–	2	1 × 8	Valid	128 × 8 × 16	N1	Spatial analysis
N1	Concatenate	–	–	–	–	128 × 1 × 48	A1	Concatenation
A1	AveragePooling2D	–	–	4 × 1	–	32 × 1 × 48	C4, C5, C6	Concatenation
C4	Conv2D	8	–	16 × 1	Same	32 × 1 × 8	N2	Temporal analysis
C5	Conv2D	8	–	8 × 1	Same	32 × 1 × 8	N2	Temporal analysis
C6	Conv2D	8	–	4 × 1	Same	32 × 1 × 8	N2	Temporal analysis
N2	Concatenate	–	–	–	–	32 × 1 × 24	A2	Concatenation
A2	AveragePooling2D	–	–	4 × 1	–	16 × 1 × 24	A3	Dimension reduction
C7	Conv2D	8	–	4 × 1	Same	16 × 1 × 8	N3	Temporal analysis
A3	AveragePooling2D	–	–	4 × 1	–	8 × 1 × 24	C8	Dimension reduction
C8	Conv2D	6	–	4 × 1	Same	8 × 1 × 12	A4	Temporal analysis
A4	AveragePooling2D	–	–	2 × 1	–	4 × 1 × 6	C7	Dimension reduction
OUT	Dense	–	–	–	–	2	–	Softmax output

TABLE 11: EEG-Inception Architecture Details according to [59]

P3Cnet [66]

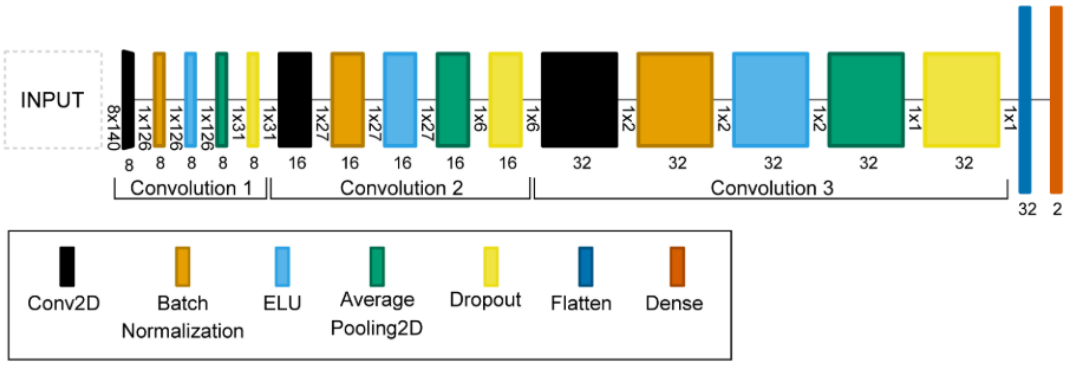


FIGURE 26: Architecture for P3Cnet.

Block	Layer	# filters	Size	Output	Options
1	Conv2D	8	(8, 15)	(1, 126, 8)	strides = (1,1), activation = linear, max norm = 1
	BatchNorm			(1, 126, 8)	
	Activation			(1, 126, 8)	activation= ELU
	AveragePool2D		(1, 4)	(1, 31, 8)	
	Dropout			(1, 31, 8)	p = 0.35
2	Conv2D	16	(1, 5)	(1, 27, 16)	strides = (1, 1), activation=linear
	BatchNorm2D			(1, 27, 16)	
	Activation			(1, 27, 16)	activation = ELU
	AveragePool2D		(1,4)	(1, 6, 16)	
	Dropout			(1, 6, 16)	p=0.2
3	Conv2D	32	(1, 5)	(1, 2, 32)	strides = (1, 1), activation = linear
	BatchNorm2D			(1, 2, 32)	
	Activation			(1, 2, 32)	activation = ELU
	AveragePool2D			(1, 1, 32)	
	Flatten			(1, 32)	
Classifier	Dense			2	max norm=0.25, activation = softmax

TABLE 12: Parameters related to the implementation of P3Cnet, according to [66]

P3Net [67]

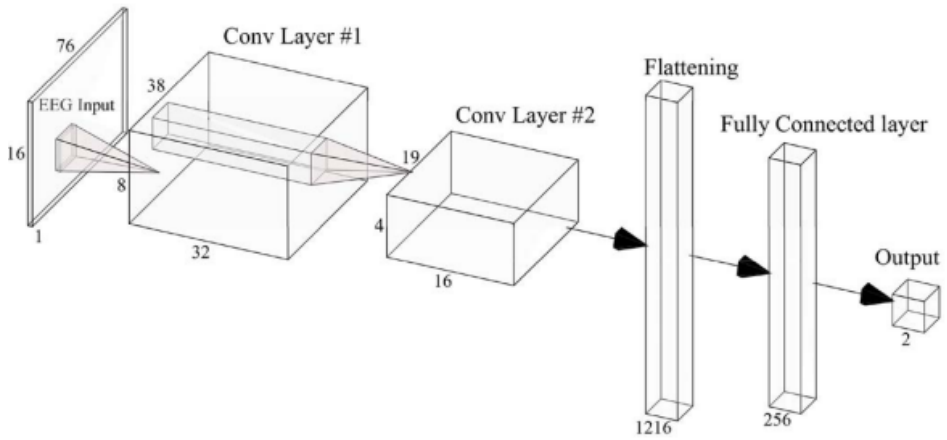


FIGURE 27: Architecture for P3net.

Block	Layer	# filters	Size	Output	Options
1	Conv2D		(7, 7)	(32, 16, 76)	
	BatchNorm			(32, 16, 76)	
	Activation			(32, 16, 76)	activation = ReLU
	MaxPool2D			(32, 8, 38)	
2	Conv2D		(7, 7)	(16, 8, 38)	
	BatchNorm2D			(16, 8, 38)	
	Activation			(16, 8, 38)	activation = ReLU
	MaxPool2D			(16, 4, 19)	
3	Dropout			(16, 4, 19)	p=0.1
	Flatten			256	
Classifier	Dense			2	max norm=0.1, activation = softmax

TABLE 13: Details of the implementation of P3Net presented in [67]

RNN-EEGNet [58]

Block	Layer	# filters	Size	Output	Options
1	Reshape			(1, 8,350)	
	ZeroPadding		(0, 32)	(1, 8, 414)	
	Conv2D	30	(1, 125)	(30, 8, 290)	activation=linear, mode=valid
	BatchNorm			(30, 8, 290)	
	DepthwiseConv2D		(8, 1)	(60, 1, 290)	D = 2, max norm=1, activation=linear, mode=valid
	BatchNorm			(60, 1, 290)	
	Activation			(60, 1, 290)	activation = ELU
	AveragePool		(1, 4)	(60, 1, 72)	
	Dropout			(60, 1, 72)	p = 0.25
2	ZeroPadding		(0,8)	(60, 1, 88)	
	SeparableConv2D	50	(1, 16)	(50, 1, 73)	activation = linear, mode=valid
	BatchNorm			(50, 1, 73)	
	Activation			(50, 1, 73)	activation = elu
	AveragePooling		(1, 8)	(50, 1, 9)	
	Dropout			(50, 1, 9)	
3	Reshape			(50, 9)	
	GRU			(50, 128)	Bidirectional, units=64, return_seq=True, activation = elu
	GRU			(50, 128)	Bidirectional, units=64, return_seq=True, activation=elu
	BatchNorm			(50, 128)	
	Activation			(50, 128)	activation=ELU
	Flatten			6400	
Classifier	Dense			2	max norm=0.25, activation = softmax

Table 14: Parameters proposed for in the implementation of RNN-EEGnet as shown in [58]

PLNet [54]

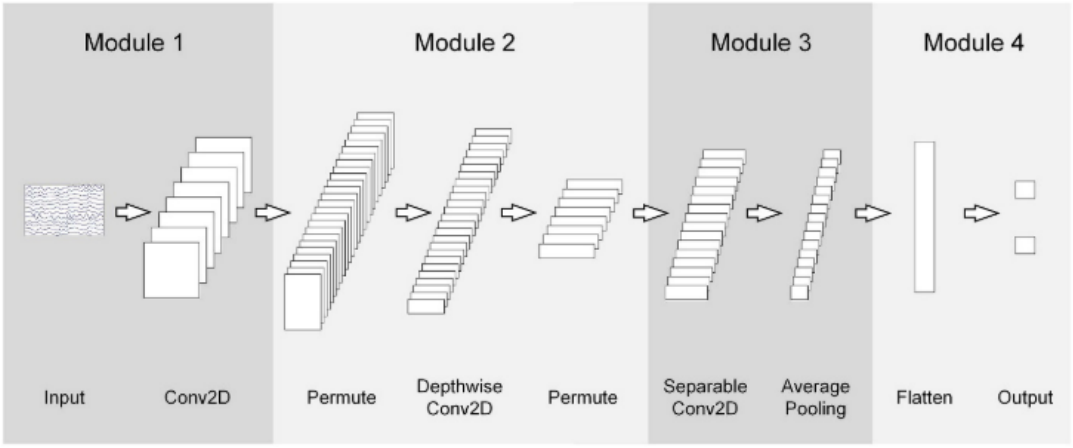


FIGURE 28: Architecture for PLnet.

Block	Layer	# filters	Size	Output	Options
1	Input			(60, 128)	
	Reshape			(60, 128, 1)	
	Conv2D	8	(1, 32)	(60, 25, 8)	strides = (1,4), max norm = 0.5
	BatchNorm			(60, 25, 8)	
	Activation			(60, 25, 8)	activation=linear
2	Permute			(60, 8, 25)	
	DepthwiseConv2D	25	(60, 1)	(1, 8, 25)	stride = (1,1), max norm=0.5, depth = 1
	Permute			(1, 25, 8)	
	BatchNorm			(1, 25, 8)	
	Activation			(1, 25, 8)	activation=ELU
	SpatialDropout2D			(1, 25, 8)	p = 0.25
3	SeparableConv2D	16	(1, 9)	(1, 17, 16)	strides = (1,1), max norm=0.5
	BatchNorm			(1, 17, 16)	
	Activation			(1, 17, 16)	activation = ELU
	AveragePool2D		(1, 17)	(1, 1, 16)	
	SpatialDropout2D			(1, 1, 16)	p=0.5
Classifier	Flatten			16	
	Dense			2	max norm=0.1, activation = softmax

TABLE 15: Parameters proposed for PLnet in [54].

B Subject information

Feasibility of channel reduction in real-time estimation of perceptual nociceptive threshold based on EEG measurements.

Information on technical scientific research for human participants

We provide the following information to inform your decision before participating in this study. Please take your time to thoroughly read and discuss it with your partner, friends, or family, if you wish. Should you have any questions after reviewing the information provided here, you are welcome to contact the researcher. Contact details are provided at the end of this information sheet.

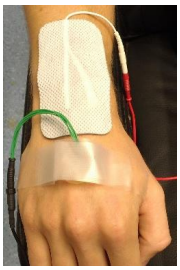
What is the objective of this study?

Chronic pain is an impairing condition that affects around 100million people in Europe. Its effects are found in different areas from economical to societal for not only patients, but also family members, caregivers, and the healthcare system. However, the understanding of the underlaying mechanisms related to pain and its maintenance is poorly understood leading to suboptimal treatment. Because of this, the development of accurate diagnostic tools is of great importance. Recently, a method for measuring properties related to the development and maintenance of pain has been developed. This project focuses on improving the current method by decreasing the time needed for the measurements and increase the reliability by using EEG responses instead of patients reports.

How will this research be conducted?

Figure 1: example of an EEG cap like the one used for the experiments.

The experiment will take place at the University of Twente and take approximately 2 hours. At the beginning of your visit, we will have an introductory conversation in which you can ask all your questions. After this we will ask you to give a written informed consent form. Next an EEG cap, such as the one shown in Figure 1 will be fitted. Gel will be injected into the electrodes. To improve conduction, the skin will be scratched a bit with a blunt needle, this should not hurt. Then the electrodes for stimulation will be attached to the back of the hand as shown in Figure 2. The stimulation electrode has very small (0.2mm) pins which will not penetrate the skin, but only stimulates the upper layer of the skin. The bigger electrode serves as a ground.



Once the set-up is complete, you will be given a response button; the measurement will not start until you press it. You can pause or stop the experiment by releasing the response button. When the stimuli are stronger than your detection threshold, you will feel a light pain sensation which can be described as tingling or little pinpricks.

The first part of the experiment consists of a familiarization session for you to get acquainted with the stimuli. A second measurement will be then conducted to get an initial estimation of the threshold. Afterwards, two measurements will take place. First, stimulations with different amplitudes around your perceptual threshold will be given and you will have to briefly release the button if you feel the stimulus. For the second one, you must keep the button pressed to receive the stimuli and try to count each stimulus you receive. If you need a short break, you can release the button.

What is expected of you?

Any person with a normal health condition can participate in this study. If you have skin abnormalities, abnormal blood pressure, heart problems, diabetes, chronic pain, implanted stimulation devices or if you are pregnant, you cannot participate.

To be able to participate in this study, there are a few recommendations:

- Don't use stimulants or narcotics (e.g. alcohol, drugs or painkillers) within 24 hours before the experiment.
- Before participating in the experiment have a good night of sleep, breakfast, and/or lunch.
- Contact the researcher if you have pain complaints at the day of the experiment.

Are there side effects and/or risks?
There are no expected side effects and/or risks by participating in this study. The needle electrode may cause some irritation to the skin or redness, which will fade in a day.

What are the possible (dis)advantages of participating in this study?
Other than contributing to science, the experiment performed in this study will not benefit or harm you in your personal situation. For us your participation can yield very useful data from which we can ultimately get a better understanding of chronic pain.

Can I retreat from participating?
You can decide freely if you want to participate in this study, so it's completely voluntary. You can change your mind at any moment and stop with the experiment without sharing a reason.

What happens with my personal information?
When you participate in this study, data about you is gathered during the experiments. This information will only be used for research and will never be disclosed to any person other than the researcher(s). Personal data will be stored as long as prescribed by legal regulations or scientific codes of conduct and destroyed afterwards. Your name will never be mentioned in any publication or presentation.

For further questions you can contact the researcher, Marcela Martínez, via m.l.martinezpinedo@student.utwente.nl

C Experimental protocol

Procedure based on B. van den Berg and L. Vanwinsen, adapted by M. Martínez.

Smart acquisition of nociceptive detection Thresholds based on an 8 channel EEG set-up.

Scope

Experimental protocol for “Feasibility of channel reduction in real-time estimation of perceptual nociceptive threshold based on EEG measurements.”

Background

The evaluation of perceptual thresholds plays a crucial role in the assessment of different sensory systems such as vision, hearing touch and nociception. Evaluation of the perceptual thresholds in pain research can be used to assess altered nociceptive processing and sensitivity of ascending pathways in the nociceptive system (van den Berg, B., et al., 2022).

For the assessment of the nociceptive function, intra-epidermal electric stimulation has been used to selectively stimulate nociceptive Fibers with currents lower than twice the detection thresholds (Mouraux, A., et al., 2010). Recent experiments at the University of Twente rely on a needle electrode placed on the right hand to administer the electrical stimuli. The electrical stimuli are generated with a custom-built stimulator (NociTRACK AmbuStim) with a uniformly randomized interstimulus interval of 3.5-4.5s.

In addition to this selective stimulation, participants must remain focused and consistently respond to the stimulation to accurately assess the perceptual thresholds. Nonetheless, perception report is not always equal to perception. Participants might lose focus due to long or boring procedures, might fail to report a perceived stimulus (lapsing) or might report stimulus perception while no stimulus was perceived (guessing). In some cases, subjects could be unable to communicate stimulus perception (van den Berg, B., et al., 2022).

Therefore, the use of a fully automated approach relying on electroencephalogram (EEG) has been explored to enable the accurate and objective assessment of perceptual threshold in a wide variety of patients groups. Nociceptive activity can be quantified by recording and analyzing the EEG signal time-locked to the stimuli. The average time-locked signal, referred to as the evoked potential (EP) describes transient synchronized activity of large neural networks within the cortex. In van den Berg et al., a deep learning classifier based on a 32 channel EEG was able to reliably analyze brain activity in response to nociceptive stimuli. The constructed neural network led to accurate estimates of the perceptual threshold and proved the possibility to use deep network classification to control the adaptive stimulus sequence and estimate the perceptual threshold in real time. Furthermore, the application of this method enables shorter procedures by removing the need to wait for the subject’s responses and limits the negative effects of poor task performance by the subject.

A next step in the development of the current method would be to reduce the number of EEG electrodes needed. A lower number of electrodes would reduce the time and effort needed to set up the procedures allowing for more time efficient studies and clinical applications on a larger scale. For this study, the electrodes (C3, Cz, C4, T8, F3, Fz and F4) were chosen based on the EPs observed using 32 electrodes with a 10-20 montage.

In this document, the experimental protocol for assessing the feasibility of an 8 channel EEG set-up will be described. The protocol consists of two measurements: one with the subject reporting their felt stimulus by pressing a button and a second where no concurrent physical task is performed.

Required Materials

Description	#	Specification
General		
Computers	1	One computer with LabView 2013 SP or higher, MTT-EP stimulation software (build in LabView), Matlab 2015b or higher, TMSi Polybench Designer, and a custom-made TMSi Polybench recording application. The computer should have at least the following specifications: - Windows 7 64-bit, or higher - Intel Core i5 1.6 GHz, or higher - Bluetooth adapter - 64-bit operating system - 8 GB RAM, or more
Tape	0.5 m	Non-allergenic skin-friendly medical tape. E.g., Leukofix.
Medical abrasive gel	2 cl	Medical abrasive gel to remove dead skin cells.
Cleansing liquid	2 cl	Alcohol, 70 % Ethanol
Cleansing tissues	2	Tissues (should preferable not release fibers)
Cleansing sticks	2	Cotton-top cleansing sticks
Multiple Threshold Tracking		
Stimulator	1	NociTRACK AmbuStim single-channel stimulator, capable of generating a minimum current of 8 µA and a maximum current of 16 mA. Shown in Figure 4.
Charger	1	NociTRACK charger for AmbuStim stimulators
Trigger generator	1	Arduino-based trigger generation system, which can be connected to the computer (input) via USB A to B cable, connected to the NociTRACK AmbuStim stimulator (output 1) via a BNC cable and connected to the EEG amplifier via a DB25 parallel cable. Shown in Figure 3 D-E-F.
USB A to B cable	1	Cable for connection of the trigger generator to the computer.
BNC to binder cable	1	Cable for connection of the trigger box (BNC) to the NociTRACK AmbuStim stimulator (Binder connector). Should have a length of at least 2 meters.
Custom cable with 37p sub-D connector	1	Cable for connection of the trigger generator to the EEG amplifier. Should have a length of 1-2 meters.
Stimulation electrode	1	Sterile IES-5 electrode for intra-epidermal electrocutaneous stimulation. Shown in Figure 9.
Grounding electrode	1	TENS disposable adhesive surface ground (40x50mm) electrode, which will serve as a ground during the stimulation. Shown in Figure 10.
Stimulator-to-electrode cable	1	A custom-made double cable that connects the stimulation and grounding electrode to the stimulator.
EEG Measurement		
Amplifier	1	TMSI SAGA32+/64+ amplifier with 32 or 64 unipolar, 4 bipolar, 9 auxiliary and 1 digital channel.
Medical power supply	1	TMSi power supply for EEG amplifier.
USB A to B cable	1	USB A to B cable used to connect the TMSi docking station to the PC
EEG caps	2	TMSi 32-channel low-noise actively shielded caps. A small-medium or medium-large size cap have to be available for different head sizes of the subjects.

Patient ground	1	TMSi wristbands for the patient ground (with its corresponding patient electrode cable)
Color-coded measurement tape	1	Tape for measuring the required EEG cap size.
Syringes		10 ml Luer-Lok tip syringe (B-D Plastipak) for injection of EEG gel into cap electrodes.
Blunt needles	1	Blunt needle (16G) for injection of EEG gel into cap electrodes and scratching the skin.
Wooden sticks	1	Small wooden sticks to move the hair after placing the electrode cap.
EEG electrode gel	200 ml	Electro-gel (ECI), for injection in EEG cap electrodes.
EEG electrode paste		Conductive EEG paste (Ten20).
Towel	2	Large size towel for covering the shoulders during application of electrode gel, and another medium size towel for subjects to dry the hair after measurement.
Power cable	1	Power cable for medical power supply.
Comfortable chair	1	Comfortable chair for participants to relax during the experiment. The chair should especially provide rest to the muscles around the head and neck since those might disturb the measurement.
Focus image	1	Small image or sign for subjects to look at during the experiment.

Procedure

A. General Preparation

Time: > 1 hour before session

1. Send potential participants an information e-mail containing the official patient information letter. After they confirm their participation, send them another e-mail with specific information about the experiment. These e-mails can contain a text like the one shown in Figure 1 The second mail should tell potential participants to:
    - The date, time, and location of the experiment.
    - Where to meet the researcher and contact information.
    - Bring a towel to the experiment, and possibly some shampoo/conditioner and hair gel.
    - Preferably wear lenses to the experiment if they have the option to choose between lenses and glasses (This is not consistent with the example text in Figure 1 but was later adapted).
    - No alcohol is allowed within 24h before the experiment.
    - Drink the same amount of coffee as they normally do.
2. If you expect participants with another language, make sure translations are available.
3. Plan an experimental date with potential participants via e-mail or telephone.
4. After a date has been planned, inform a possible emergency contact about the date and time of the experiment.
5. Regularly charge the electric chair. However, do not leave the charger connected for more than a few hours, since this might damage the battery.
6. **Always make sure that sufficient sterilized electrodes are available**. If this is not the case, sterilize a batch of electrodes according to "*SOP - Sterilization of IES-5 and BiModEl Electrodes*".
7. Make sure that you know in which lab the experiment takes place and which numbers to call in case of an emergency.

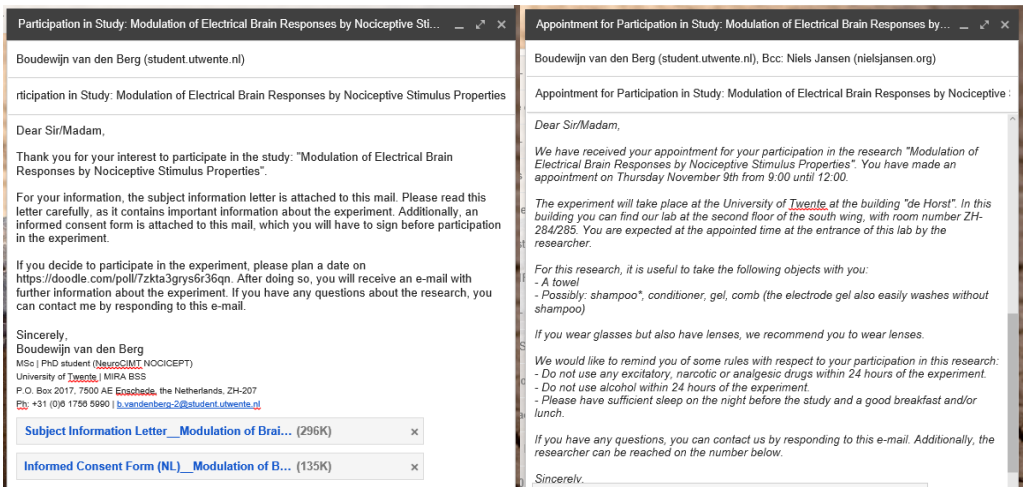


Figure 2: Example e-mails.

B. EEG System Preparation

Time: > 1 hour before session

1. Attach the EEG amplifier to the power adapter.
2. Attach the EEG amplifier via a USB A to B cable to the EEG computer.
3. Connect the trigger generator to the computer using the USB A to B cable.

4. Connect the trigger generator to the EEG amplifier using the parallel cable.
5. Connect the cap to the amplifier.
6. Connect the unipolar snap cable of the grounding electrode to GND.
7. Attach a small sign or image in front of the chair, for subjects to focus on a single point during the experiment.

For more information about the connections, please refer to Figure 2.

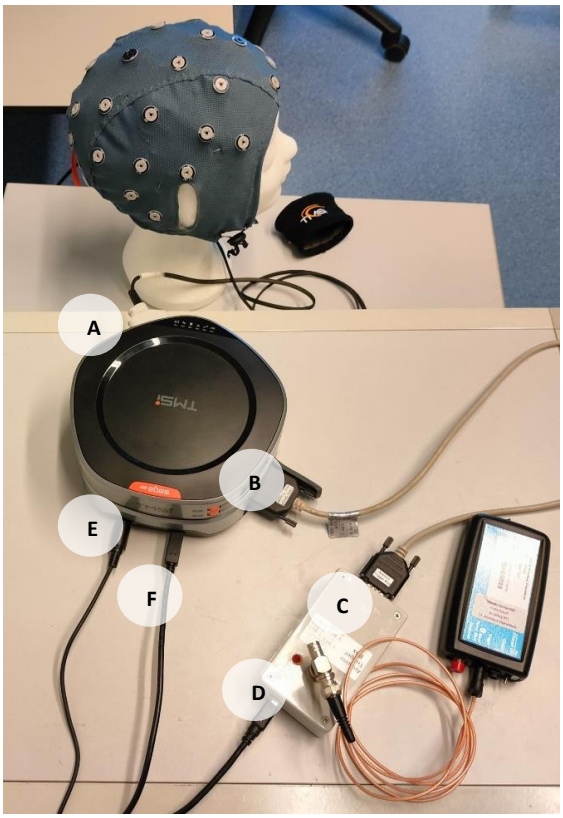


Figure 2. Connections between EEG system, trigger generator and EEG cap using a SAGA amplifier. A is the connector of the cap. B is the connection of the trigger cable which is connected to the trigger box C. The trigger box should also be connected to the PC through cable D. Additionally, the amplifier is powered by connecting cable E and cable F is the connects the amplifier to the same PC.

C. Stimulator System Preparation

Time: > 1 hour before session

1. Connect the trigger generator to the stimulator via the BNC cable.
2. Note the number of the used stimulator and locate the corresponding calibration file.
3. Connect the stimulator-to-electrode cable to the stimulator.
4. Charge the stimulator **with the stimulator turned off**, using the NociTRACK charger.

For more information about the connections, please refer to Figure 3.

Figure 3: Inputs and outputs of the NociTRACK AmbuStim stimulator. A is a response button, which is actuated by the subject to indicate if the stimulus is perceived. B is a connection to the stimulator cable. C is an input for the trigger signal. D is the power switch. E is the input for the charger. The stimulator should be switched off while charging.

D. Materials Preparation

Time: > 1 hour before session

Make sure that the following materials are ready for use and easy to reach:

1. Electrode caps.
2. Colored measurement tape.
3. EEG electrode gel and paste.
4. Two syringes.
5. Blunt needles (in the package) for gel injection into the electrodes.
6. Medical abrasive gel.
7. Tissues and cotton sticks.
8. Alcohol.
9. Unipolar snap cable for the grounding electrode of the EEG
10. One TENS electrode.
11. One sterilized IES-5 electrode.

E. System Start-up

Time: > 20 minutes before session

1. Turn on both computers.
2. Turn on the power supply of the EEG amplifier.
3. Turn on the EEG amplifier.
4. Turn on the NociTrack.
5. Note the NociTrack number and check for an existing calibration file.

F. Labview Initialization

Time: > 10 minutes before session

On the LabView computer:

1. Edit the config.ini by copying the configuration for the first part of the experiment from ‘task1_MM’
2. Open the LabVIEW program NDTEP.vi.
3. Wait until FrontPanel.vi started. Press the arrow button.
4. A new screen will open (Figure 4b). Fill in the Patient ID (do not add personal information) and press “Continue”.
5. A new screen will open (StimCom Bluetooth Control).
    a. Click on “Search” to start searching for a AmbuStim stimulator (Figure 4a).
    b. The number of the AmbuStim is indicated on the stimulator.
    c. If the AmbuStim does not appear on the screen, check if the stimulator is turned on and press “Search” again.

d. Click on the AmbuStim which is to be connected, but do not press “Connect” yet (Figure 4a).

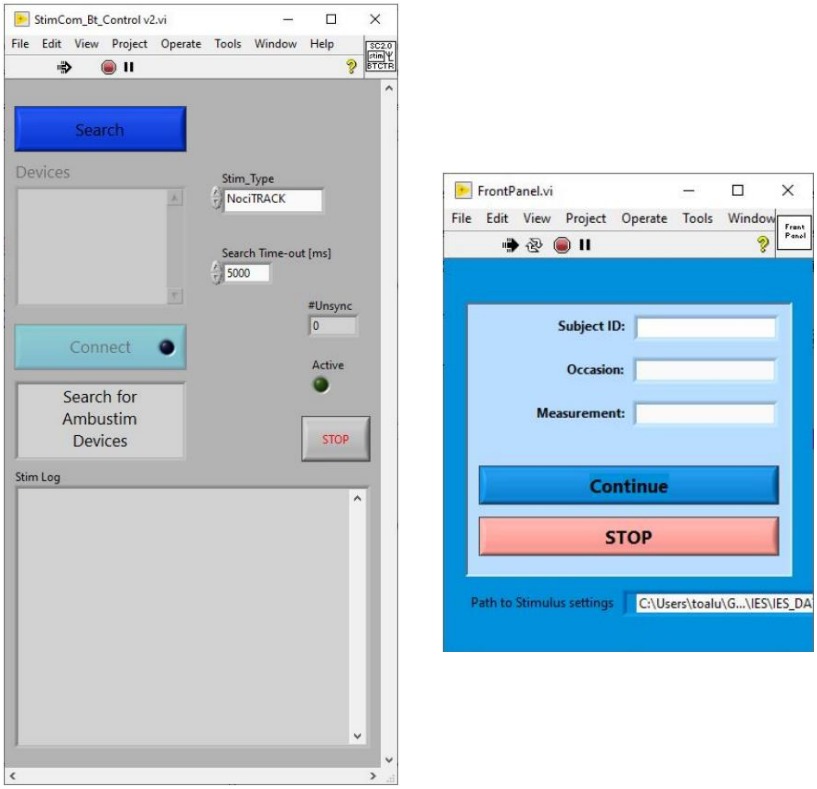


Figure 4: Left, the Bluetooth connection interface for the NociTRACK AmbuStim stimulator. First, press ‘Search.’ If the correct stimulator device has been found (doublecheck with the number on the stimulator), press ‘Connect.’ Right, the initial interface of the experimental application. Fill in the subject ID, confirm the COM port of the trigger generator in ‘Device Manager,’ and press ‘Continue.’

G. EEG initialization

Time: > 10 minutes before session

1. The impedance measurement interface in MATLAB will start after the device is connected.
2. Turn your phone to airplane mode.
3. Close the impedance measurement window and wait for MATLAB to automatically close.

H. Subject reception and preparation

Time: start of the session

1. Meet with the subject at the entrance of the lab. To do so, be present at the entrance 10 minutes before the start of the session and leave the lab door open.
2. Give the subject a hard copy of the information letter and the informed consent (preferably, the subject has already received and read the information letter in advance, via e-mail). Ask the subject:
    - **“Do you have any questions?”**
    - **“Would you like to participate in the study?”**
    - **“Do you want to sign the informed consent?”**
3. Explain to the subject what is going to happen. Ask the subject:
    - **“Please, set the mobile phone to airplane mode.”**

- **“Do you need to go to the toilet? The session will take approximately 2 hours.”**

4. Instruct the subject:
    - **“Sit-down on the chair and set the chair to a comfortable position.”**
    - **“Make sure there is sufficient space for the legs.”**
    - **“The chair should be inclined slightly backwards to relieve the muscles around the neck.”**
5. Use the colored measurement tape to measure which size of the cap is required.
6. Measure the distance between the nasion and the inion.
7. Measure the distance between the pre-auricular points.
8. Place the electrode cap on the head of the subject, by pulling over the forehead towards the back.
9. Ask the subject:
    - **“Please adjust the cap to fit as tightly as possible on the head.”**
    - **“Please, attach the strap around the chin as tightly as possible, and pull the rope at the side of the cap for a better fit.”**
10. Using the measured distances, make sure that the cap electrode is exactly in the middle of those measured positions. If necessary, slightly adjust the position of the cap, and ask the subject for feedback on the fit. Tape the wires of the cap to the share or shirt to avoid movement artifacts.
11. Once the cap is adjusted, secure it by taping the wires to the chair or subject’s shirt.
12. Use medical abrasive gel and a cotton stick to clean the positions of the ground electrode.
13. Use a tissue with alcohol to clean the same location.
14. Explain the subject:
    - **“You will not feel anything from EEG measurement.”**
    - **“Gel will be injected into the electrodes. To improve conduction, the skin will be scratched a bit with a blunt needle. This should not hurt, if it does, pleas say so.”**
    - **“If you discomfort, you can indicate this at any time.”**
15. Make sure that the location of the ground electrode is dry and attach the ground electrode.
16. Take the needle and the syringe. Show to the subject you take a new needle from the package and fill the needle and syringe with gel.
17. Fill the selected electrodes (F3, Fz, F4, T7, C3, Cz, C4, T8) in the cap with gel by injecting it while scratching the skin by turning the blunt needle with a circular motion. Use the back end of a cotton stick for extra scratching if necessary.
18. Use the impedance display on the screen to make sure impedances are below 5 kOhm.
19. In the impedance interface of the EEG recording application, take a screenshot to save information about the impedances.
20. Explain to the subject that you will attach the electrodes for stimulation:
    - **“The first electrode is an electrode with small pins that will not penetrate the skin, but solely serve to stimulate the upper layer of skin.”**
    - **“The second electrode is a sticky electrode which serves as a ground.”**
21. Attach the electrodes as depicted in Figure 5. Ask the subject:
    - **“Please, hold the IES-5 electrode at the back of the right hand.”**
22. Attach the electrode with medical tape and ask the subject:
    - **“Is the pressure on the electrode needles painless?”**
23. Glue the TENS electrode right behind the IES-5 electrode on the wrist, as is depicted in Figure 5. Ask the subject:
    - **“Can you press the electrode firmly onto the skin?”**
24. Attach the stimulator-to-electrode cable to the electrodes.

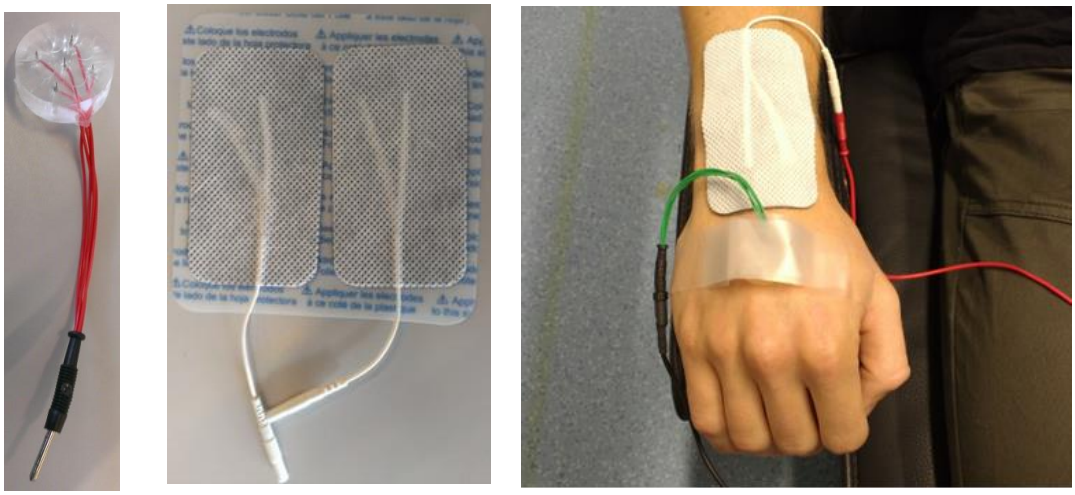


Figure 5: The IES-5 electrode (left) and the TENs electrode (center), which serves as ground for stimulation, and the example of the placement of the electrodes (right)

I. Familiarization

Time: 40 minutes after start of the session

1. Explain to the subject:
    - **"First, a measurement will be made for familiarization and initialization of the experiment."**
    - **"I will press on start in the application. However, the measurement will not start until you press the response button. You can pause or stop the experiment by releasing the response button."**
2. Ask the subject:
    - **"Please, hold the stimulator in the hand opposite to the side of stimulation."**
3. Explain to the subject:
    - **"The first measurement will just serve to get acquainted with the stimuli. Therefore, you should hold the response button as long as possible and release the response button if the stimuli start to hurt. If the sequence reaches 1 mA, the measurement will stop automatically."**
4. Press 'Connect' in the LabView interface to connect to the stimulator.
5. A screen will open for measurement of the initial detection threshold (Figure 6). When the subject is ready, start the first measurement via the LabView interface by pressing 'Stimulate' and tell the subject that he/she can start by pressing the response button.
6. Explain that:
    - **"The second measurement will serve to determine the initial detection threshold. Therefore, you should release the response button as soon as you feel a sensation that you ascribe to the stimulus."**
7. When the subject is ready, start the second measurement via the LabView interface and tell the subject that they can start by pressing the response button.
8. If necessary, the second measurement can be repeated by starting the measurement a third time via the interface. If the second measurement was successful, press 'Continue' in the LabView interface.
9. A new screen will appear (Figure 7), which is the control interface for multiple threshold tracking.

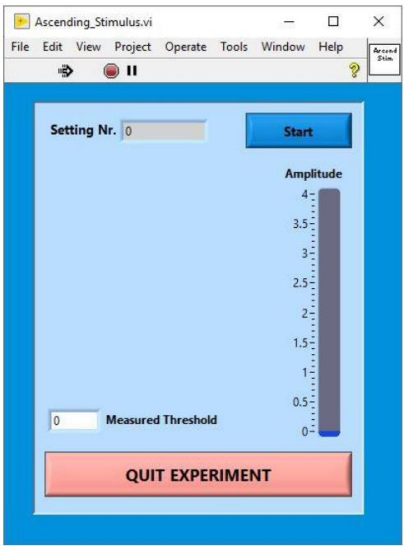


Figure 6: Interface for familiarization and measurement of the initial detection threshold.

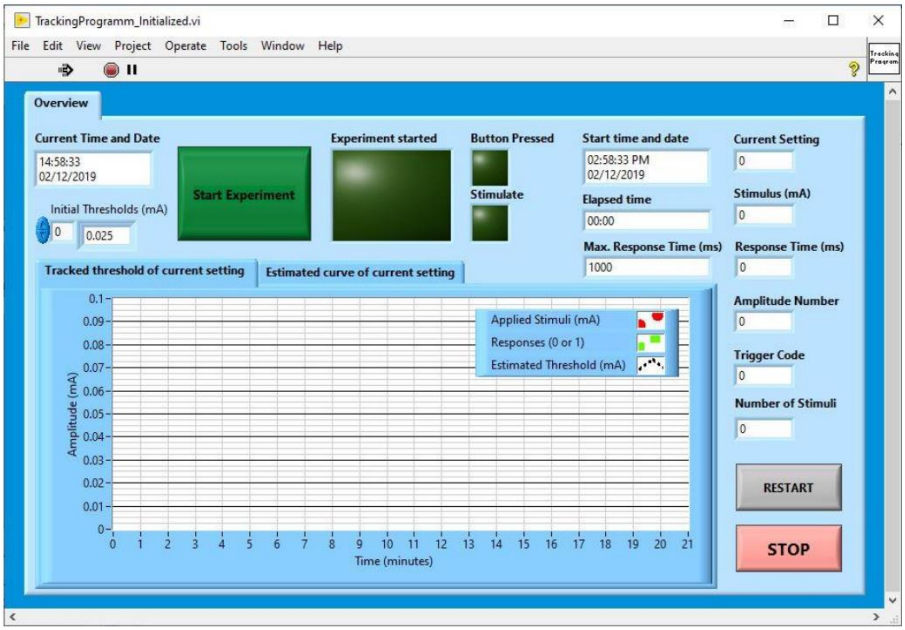


Figure 7: Interface for multiple threshold tracking.

J. Experiment

Time: 50 minutes after the start of the experiment

1. Open the LabVIEW program program “NDTEP.vi” .
2. Wait until FrontPanel.vi starts. Press the arrow button.
3. A new screen will open (Figure 4b). Fill in the Patient ID (do not add personal information) and press “Continue”.
4. A new screen will open (StimCom Bluetooth Control)
    - Click on “Search” to start searching for an AmbuStim stimulator (Figure 4a)
    - The number of the AmbuStim is indicated on the stimulator
    - If the AmbuStim does not appear on the screen, check if the stimulator is turned on and press “Search” again.
5. Click on the AmbuStim which is to be connected, and press “Connect” (Figure 4a).

6. Perform the two tasks explained below in random order.

J.1. Task 1

In goal of this task is to evaluate the performance of the deep learning classifier in comparison to the subject report (button-press)

1. Edit the file config.ini by copying the configurations for the second part of the experiment from **config**
2. Explain the experimental procedure to the subject:
    - **"To receive stimuli, you have to press the button."**
    - **"You have to release the button immediately when you feel a sensation that you prescribe to the stimulus."**
    - **"After releasing the button, you can re-press the button after approximately one second."**
    - **"If you need a short break, you can wait longer before re-pressing the button."**
3. Ask the subject:
    - **"Please, blink as few times possible while holding the response button."**
    - **"Keep looking towards the focus image on the wall while holding the response button."**
    - **"Try to relax and not move while holding the response button."**
    - **"Do not talk while holding the response button."**
    - **"Keep your attention focused on the detection of stimuli."**
    - **"Doing this will greatly enhance the signal quality."**
4. Press "Start Experiment".
5. Indicate the subject may now press the button and start the procedure.
6. A message will pop-up after 200 stimulations indicating the end of the session.

J.2. Task 2

In goal of this task is to evaluate the performance of the deep learning classifier when no concurrent physical task is performed.

1. Edit the file config.ini by copying the configurations for the second part of the experiment from "task2_MM"
2. Explain the experimental procedure to the subject:
    - **"To receive stimuli, you have to press the button."**
    - **"Keep the button pressed and focus on the stimuli. Try to count each stimulus you receive."**
    - **"If you need a short break, you can release the button."**
3. Ask the subject:
    - **"Please, blink as few times possible while holding the response button."**
    - **"Keep looking towards the focus image on the wall while holding the response button."**
    - **"Try to relax and not move while holding the response button."**
    - **"Do not talk while holding the response button. If you have an urgent question, you can release it before speaking."**
    - **"Keep your attention focused on the detection of stimuli."**
    - **"Doing this will greatly enhance the signal quality."**
4. Press "Start Experiment".
5. Indicate the subject may now press the button and start the procedure.

6. A message will pop-up after 200 stimulations indicating the end of the session.

K. Round-up

1. Inform the subject:
   - **"The experiment was completed successfully."**
2. Turn-off the stimulator and disconnect the subject from all cables.
3. Instruct the subject:
   - **"You can take off the EEG cap."**
   - **"You can wash your hair in the sink in the lab, or downstairs in the shower, (ZH-109, go down the stairs close to the red couches, in front of the stairs.)."**
4. When the subject is ready to leave, tell the subject:
   - **"Thank you for your participation in the experiment."**
   - Give the student the financial compensation for participation in the experiment.
   - Ask the subject if he/she would like to be informed about the result of the experiment.
   - Provide the subject with contact information in case he/she has any questions.

L. Clean-up

1. Turn-off the software, and the EEG amplifier.
2. Make a note regarding the adverse event on the informed consent form.
3. Clean the cap electrodes directly after the experiment following the instruction next to the sink.
4. Dry the cap on the ventilator.
5. Put all equipment back where it belongs.

References

[1] Inui K, Kakigi R. Pain perception in humans: use of intraepidermal electrical stimulation. J Neurol Neurosurg Psychiatry 2012;83:551–6. https://doi.org/10.1136/jnnp-2011-301484.

[2] Steenbergen P, Buitenweg JR, Trojan J, van der Heide EM, van den Heuvel T, Flor H, et al. A system for inducing concurrent tactile and nociceptive sensations at the same site using electrocutaneous stimulation. Behav Res Methods 2012;44:924–33. https://doi.org/10.3758/s13428-012-0216-y.

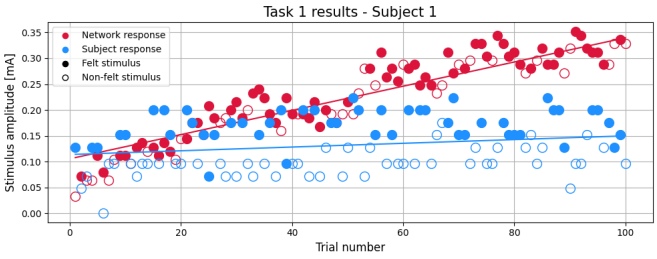
[3] Doll RJ. Psychophysical methods for improved observation of nociceptive processing. University of Twente, 2016. [4] Sur S, Sinha V. Event-related potential: An overview. Ind Psychiatry J 2009;18:70. https://doi.org/10.4103/0972-6748.57865

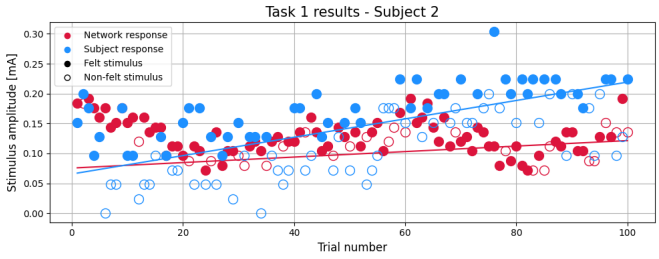
D Individual results from the threshold tracking procedure

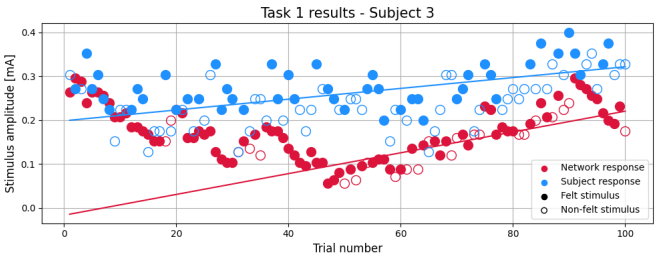
The obtained threshold tracking results from the conducted experiments are shown below:

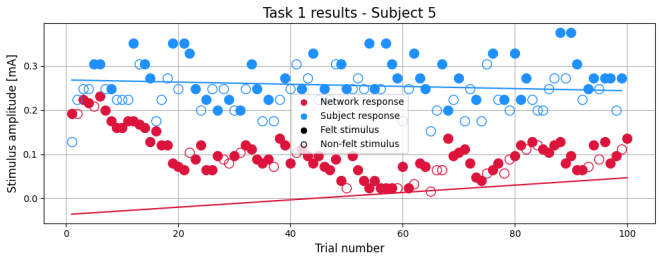
D.1 Task 1

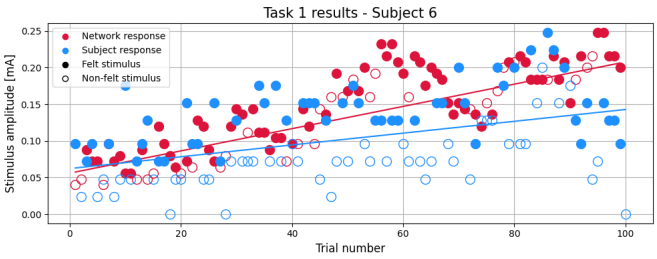
Due to a malfunctioning electrode in the EEG cap, the data from subjects 3 and 5 were excluded from this task. Additionally, the measurement from subject 4 was not recorded properly, as the configuration intended for Task 2 was mistakenly applied during this part of the experiment.

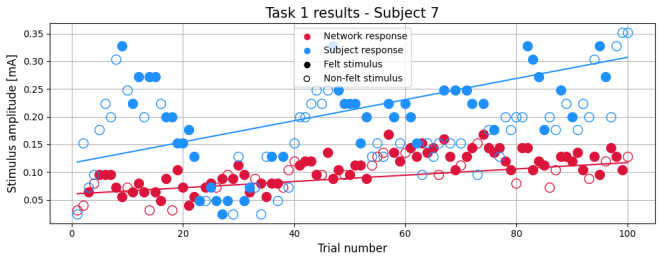


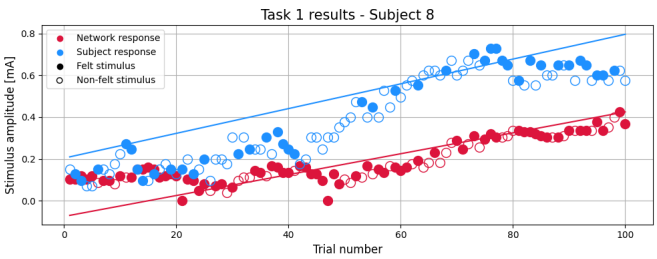


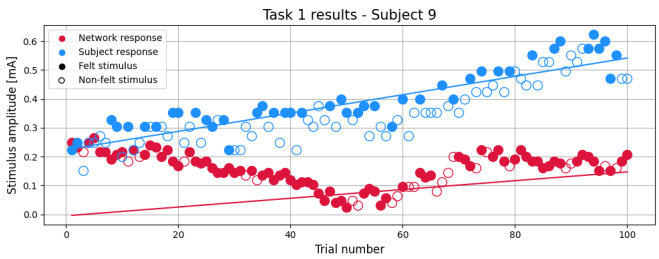


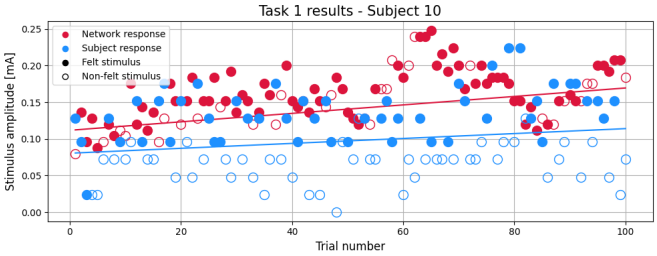


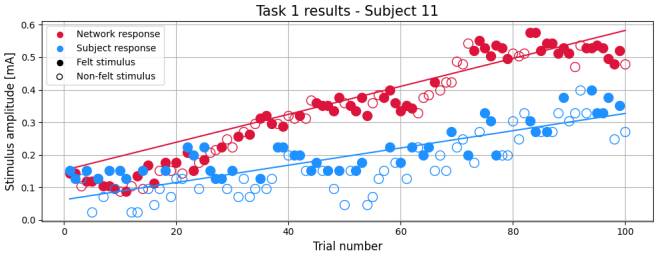


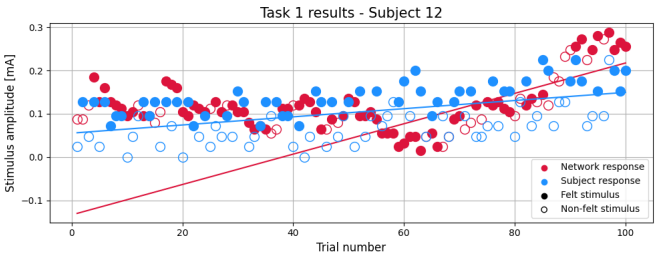


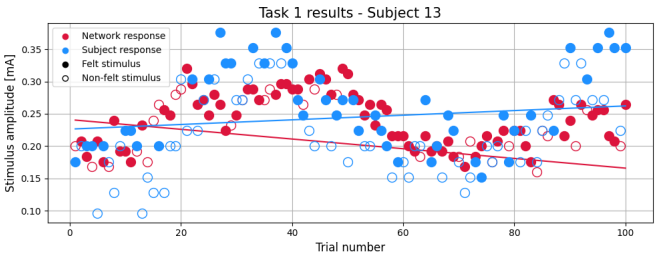


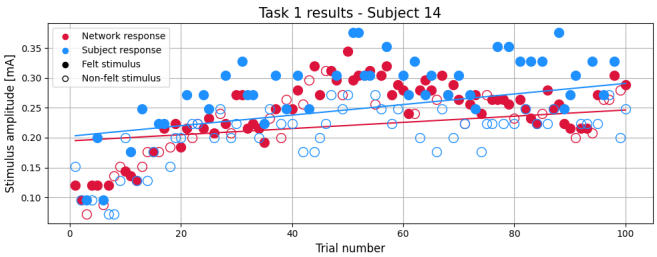


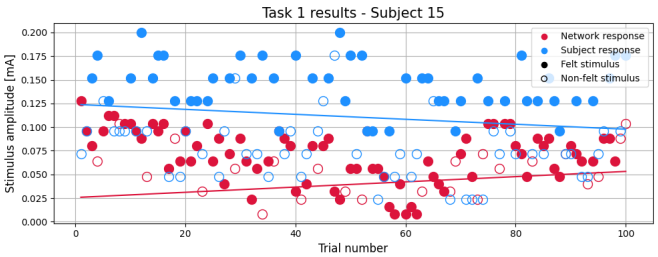






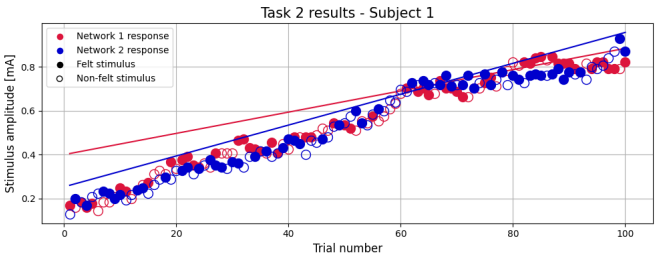


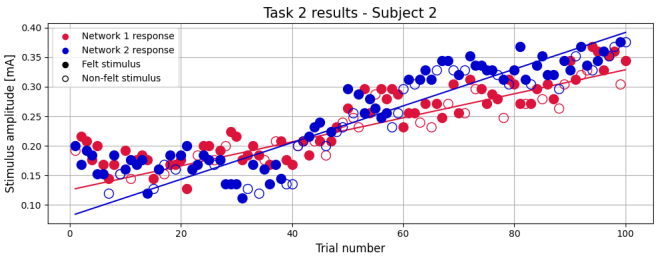


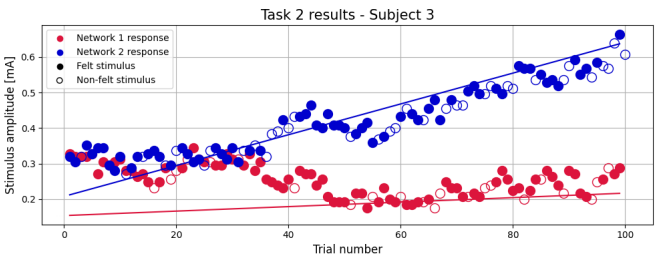


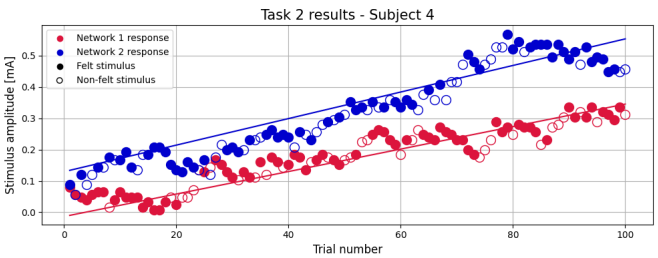
D.2 Task 2

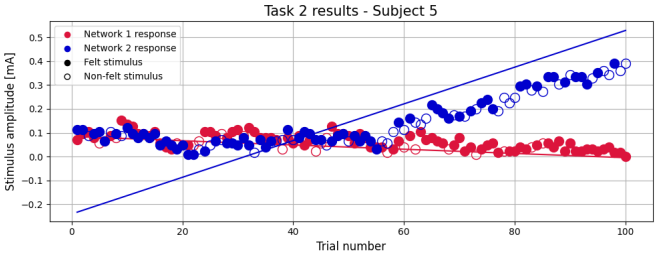
For this task, subjects 3 and 5 were excluded due to the same issue as in task 1. Furthermore, subject 1 was excluded because it is believed that a broken stimulation cable prevented a proper administration of stimuli, leading to unreliable results.

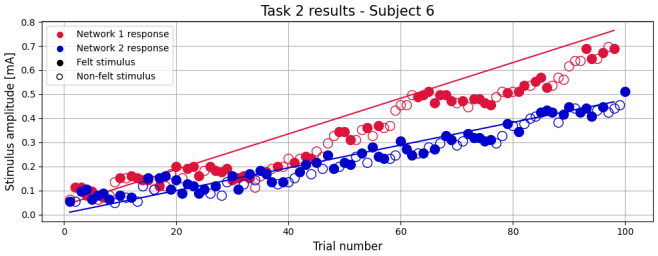


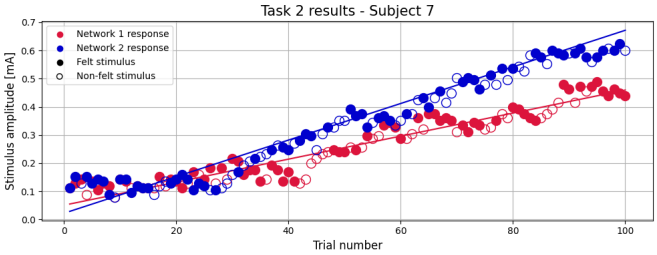


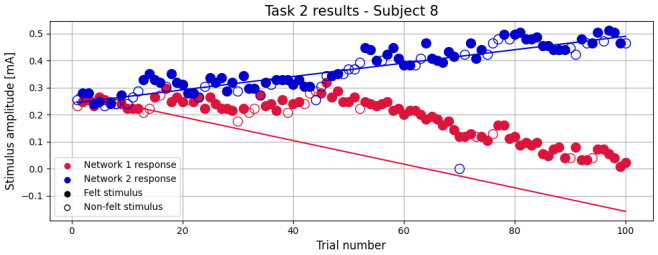


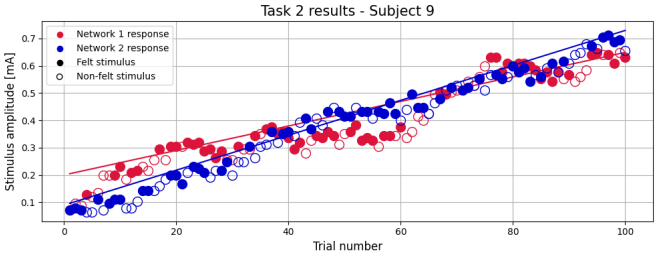


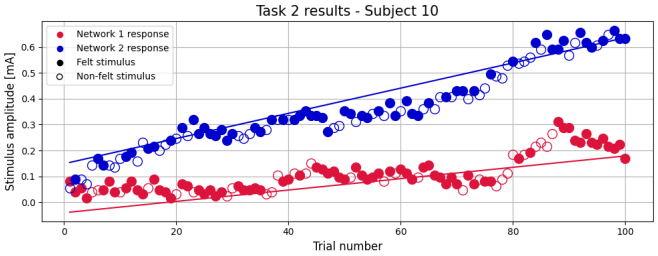

Task 2 results - Subject 5
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number

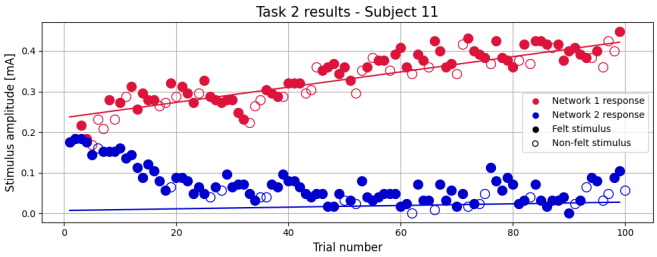

Task 2 results - Subject 6
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number

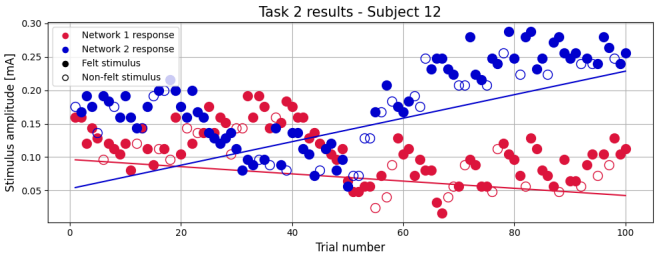

Task 2 results - Subject 7
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number

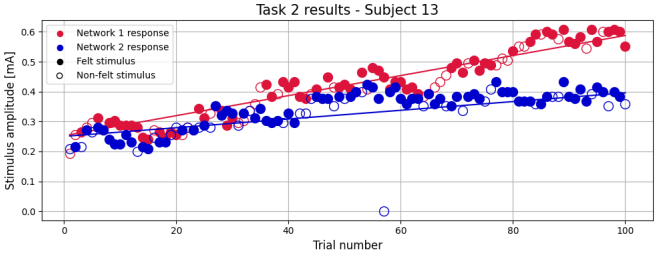

Task 2 results - Subject 8
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number

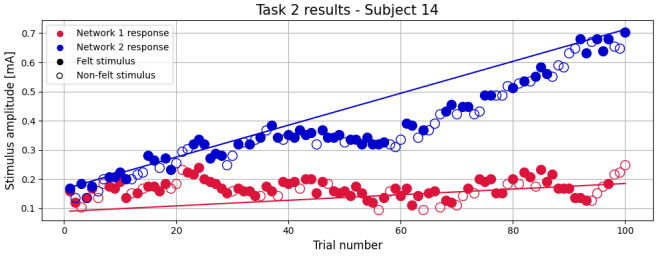

Task 2 results - Subject 9
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number

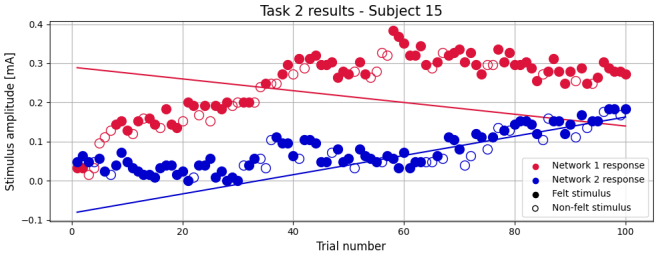

Task 2 results - Subject 10
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number


Task 2 results - Subject 11
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number


Task 2 results - Subject 12
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number


Task 2 results - Subject 13
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number


Task 2 results - Subject 14
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus
Stimulus amplitude [mA]
Trial number


Task 2 results - Subject 15
Stimulus amplitude [mA]
Trial number
Network 1 response
Network 2 response
Felt stimulus
Non-felt stimulus

E Ordinary linear regression

To explore factors potentially influencing the accuracy of threshold estimates generated by the neural network, an ordinary linear regression analysis was conducted, focusing only on Task 1, where a threshold derived from subjects' button press responses was available. The dependent variable was defined as the absolute difference between the estimated and actual thresholds for each subject. The following explanatory variables were considered:

- Order of tasks: A binary variable was created to indicate whether the subject performed Task 1 (coded as 0) or Task 2 (coded as 1) first.
- Gender
- Signal quality: The proportion of non-rejected epochs after applying Autoreject was used as a metric for signal quality.

Other parameters, such as detection rate and age, were initially considered. However, these were excluded from the final model due to their limited variability. Specifically, the detection rate for all subjects was approximately 0.5, and the age range of participants was narrow (between 22 and 27 years).

Continuous variables, including signal quality and the absolute difference between thresholds, were shifted by subtracting their mean. The correlation among the explanatory variables was assessed using the Variance Inflation Factor (VIF). All variables presented VIF values below 1.6, indicating no significant multicollinearity and making them suitable for inclusion in the regression analysis [73].

The resulting model was not statistically significant (p-value = 0.258), suggesting that the selected variables do not adequately explain the differences in threshold estimates. However, given the small sample size, it is unclear whether this finding is generalizable.

F Q-Q plots

The distribution obtained from the experimental data, was explored using a Q-Q plot in order to assess normality of the data. Figure 29 and 30 show the plots for task 1 and 2 respectively. From these figures it is clear that the assumption of normality is not met and therefore non parametric statistics are used.

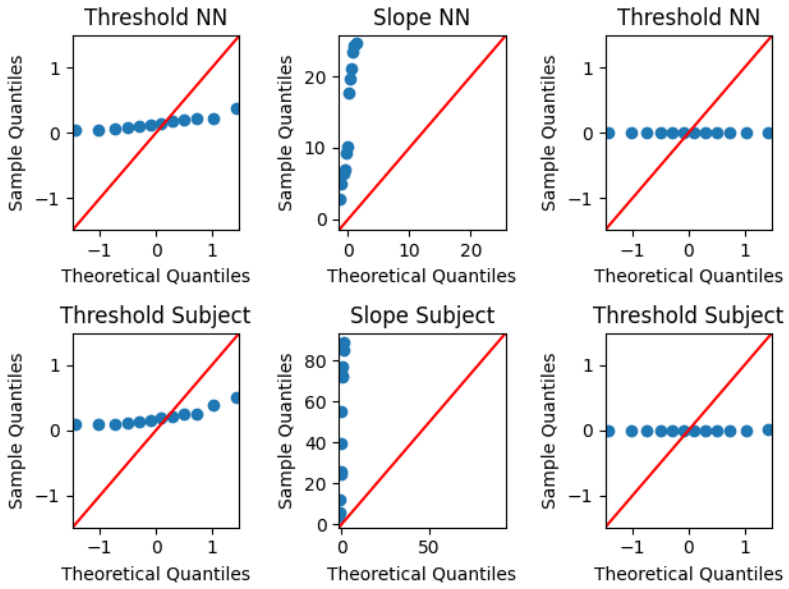


FIGURE 29: Q-Q plot for the metrics derived from task 1 to assess normality.

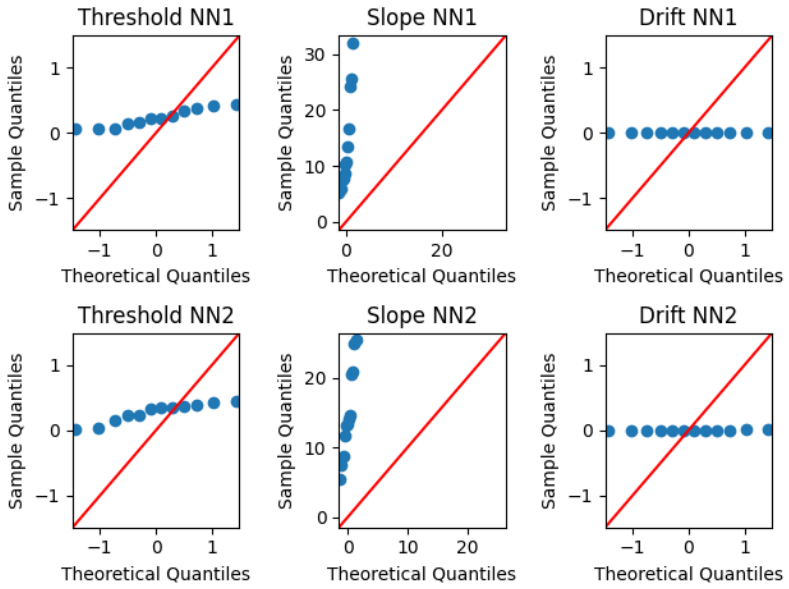


FIGURE 30: Q-Q plot for the metrics derived from task 1 to assess normality.