

<https://researchcommons.waikato.ac.nz/>

Research Commons at the University of Waikato

Copyright Statement:

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

The thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognise the author's right to be identified as the author of the thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from the thesis.

Machine Learning for Nociception Assessment from Single-Channel Electroencephalography During General Anaesthesia

A thesis
submitted in partial fulfilment
of the requirements for the Degree
of
Master of Science (Research) in Artificial Intelligence
at
The University of Waikato
by
Christopher Mayo



THE UNIVERSITY OF
WAIKATO
Te Whare Wānanga o Waikato

2026

Abstract

Nociception is defined as the neural encodings of noxious stimuli, and is a key dimension of general anaesthesia that clinicians seek to control. The use of machine learning (ML) to predict intraoperative nociception from single-channel electroencephalography (EEG) data is poorly established, with existing indices relying on opaque proprietary algorithms or limited validation. A key challenge is the lack of reliable ground truth: many prior studies evaluate models solely on their ability to predict postoperative pain scores, which may not accurately reflect intraoperative nociceptive processes.

This thesis addresses this problem through two complementary contributions. Firstly, a hypothesis-driven unified validation framework was created to evaluate nociception indices against multiple physiologically motivated criteria, including responsiveness to surgical incision and analgesic administration, rather than relying solely on postoperative pain scores. Secondly, a range of ML approaches were developed to predict nociception from EEG. A dataset comprising 1,541 total hours of intraoperative EEG across 567 patients from three previous studies was collated, preprocessed, and aligned with drug dosage and surgical event data. This was used to develop and evaluate putative nociception indices using a range of ML approaches, including unsupervised approaches, representation learning methods to reduce confounding, and a novel EEG-informed pharmacodynamic non-linear mixed effects model (EEG-PD-NLME) that exploits fentanyl pharmacology to provide a supervisory signal.

The best-performing candidate indices were a principal component-based index (PCA-6) and an EEG-PD-NLME model with additional adversarial training (NLME+adv.), satisfying most of the criteria in the proposed framework. However, despite this, both models performed poorly when predicting postoperative pain (AUROC 0.511 and 0.388, respectively). This discrepancy suggests a mismatch between intraoperative nociception and postoperative pain, challenging the use of postoperative pain scores as a validation endpoint.

A well-validated and well-described index of nociception could be utilised clinically, allowing for improved titration of analgesic medications during surgery. This may lead to improved postoperative outcomes such as reduced postoperative pain. More broadly, this thesis demonstrates the importance of comprehensive evaluation when learning latent physiological variables in the absence of ground truth data.

Acknowledgements

This thesis is built upon substantial efforts from multiple individuals and research groups over more than a decade. Without these efforts, this thesis would not be possible. I would like to give my sincerest thanks to all who have been involved in this research up to this point, along with those who have supported my journey over the last 12 months.

Firstly, I would like to express my gratitude to my supervisors:

- Dr Jason Kurz of the Department of Mathematics at the University of Waikato. Thank you for your technical guidance and support in this project, along with your critical mathematical, statistical, and ML insights as I navigated the uncharted waters of EEG nociception assessment.
- Professor Jamie Sleight of the Department of Anaesthesiology at the University of Auckland (Waikato Clinical School), who has led the overall research direction of this project for many years. Thank you for your insights into anaesthesia, EEG analysis, pharmacology, and causality, and thank you for providing the datasets and clinical questions that have underpinned this work.

I would also like to acknowledge the research groups who Professor Sleight has led over the years who have collected the datasets used in this study. A special thanks to those who led the collection and initial analysis of the datasets: Dr Corrine Law, Dr Darren Hight, Dr Amy Gaskell, Dr Rebecca Pullon, and Dr Antara Banerji, as well as all the other anaesthetists and clinical staff involved in the collection of this data.

I would also like to acknowledge the Anaesthesia Neuroimaging Group from the University of Oxford who inspired some of the approaches used in this thesis. A special thanks to Dr Minh Tran, who led some of the initial work in applying ML to postoperative pain prediction in this dataset, and Dr Marco Fabuš for his work on hidden Markov models and EEG microstates that inspired some of the work in Chapter 5.

Finally, I would like to thank my friends, family, and my partner Kato for their ongoing support throughout this work. I would like to give a special thanks to my father, Dr Michael Mayo, for his help with refining some of the ideas in this thesis.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Aims and Contributions	2
1.3	Structure of Thesis	3
2	Background	5
2.1	Anaesthesia	5
2.2	Nociception and Pain	6
2.3	Pharmacokinetics and Pharmacodynamics	8
2.3.1	Causality and the Drug Titration Paradox	11
2.4	EEG Monitoring in Anaesthesia	13
2.5	EEG Analysis	14
2.5.1	EEG Transforms	14
2.5.2	EEG Cleaning	15
2.5.3	Frequency Domain	17
2.5.4	Other EEG Features	18
2.5.5	EEG Microstates	18
2.5.6	EEG Variability	19
2.5.7	Single-Channel EEG	19
2.6	Deep Learning in EEG Analysis	20
2.7	EEG and Nociception	21
2.8	EEG Drug Effects	22
2.8.1	Propofol	22
2.8.2	Volatile Agents	23
2.8.3	Fentanyl	23
2.9	Non-EEG Nociception Indices	23
2.10	Literature Review: EEG Nociception Indices	25
2.10.1	Depth-of-anaesthesia Indices and Extensions	25
2.10.2	Novel Non-ML Indices	29
2.10.3	Novel ML Indices	30
2.10.4	Discussion	34

3	Dataset and Tools	39
3.1	Software and Libraries	39
3.2	Dataset	39
3.2.1	BISPain	39
3.2.2	AlphaGA	40
3.2.3	AlphaMax	40
3.3	Ethical Approval	40
3.4	Data Processing	41
3.4.1	PK Calculations	41
3.4.2	EEG Cleaning	41
3.4.3	OpenIbis	44
3.4.4	SingLEM Embeddings	45
3.5	Dataset Summary	45
3.6	Exploratory Data Analysis	45
4	Validation Framework and Baseline Predictors	50
4.1	Motivation	50
4.2	Train-Validation-Test Split	50
4.3	Baseline Postoperative Pain Prediction	51
4.4	Validation Framework	52
4.4.1	Postoperative Pain Prediction (Category A)	52
4.4.2	OpenIbis (Category B)	55
4.4.3	Start of Surgery (Category C)	55
4.4.4	Fentanyl Bolus Correlations (Category D)	56
4.4.5	Explainability Analysis (Category E)	56
4.5	Baseline Indices	57
4.5.1	Constant Index	57
4.5.2	OpenIbis	57
4.6	Pain Score Category Prediction	59
4.6.1	Methods	59
4.6.2	Rank-Based Transform	60
4.6.3	Results	61
4.6.4	Discussion	64
5	Unsupervised Learning Approaches	65
5.1	Motivation	65
5.2	Dimensionality Reduction	66
5.2.1	Background	66
5.2.2	Visualisation	67
5.2.3	Methods	68
5.2.4	Results	71
5.2.5	Discussion	74
5.3	Hidden Markov Models and Microstates	75
5.3.1	Background	75
5.3.2	Methods	76

5.3.3	Results	78
5.3.4	Discussion	79
6	EEG-Informed Pharmacodynamic Non-Linear Mixed Effects Model	83
6.1	Motivation	83
6.2	Mixed-Effects Models in Pharmacology	84
6.3	EEG-PD-NLME Model	85
6.3.1	Mathematical Formulation	87
6.3.2	Loss Functions	90
6.3.3	Practical Considerations	92
6.4	Methods	93
6.4.1	Dataset	93
6.4.2	Per-Epoch Validation	93
6.4.3	Training	94
6.5	Experiments	95
6.5.1	Experiment 1: Initial Model	95
6.5.2	Experiment 2: PD Model and Loss Function	96
6.5.3	Experiment 3: Input Type	98
6.5.4	Experiment 4: Model Type	100
6.5.5	Experiment 5: Dataset Type	101
6.5.6	Experiment 6: Input Length	101
6.5.7	Experiment 7: Learning Rate	102
6.5.8	Experiment 8: Weight Decay	103
6.5.9	Experiment 9: Baseline EEG Comparison	103
6.5.10	Final Evaluation	104
6.6	Pre-Surgical Data	108
6.6.1	Model Adjustments	108
6.6.2	Experiment	110
6.7	Discussion	114
7	Representation Learning	117
7.1	Motivation	117
7.2	Adversarial Learning	117
7.3	Formulation	119
7.3.1	Baseline Binary Classification Model	119
7.3.2	Adversarial Heads	120
7.3.3	Causal Analysis	120
7.4	Methods	122
7.5	Experiments	123
7.5.1	Experiment 1: Baseline Model	123
7.5.2	Experiment 2: Patient Identity Invariance	127
7.5.3	Experiment 3: Hypnotic Depth Invariance	130
7.5.4	Experiment 4: Combined Adversarial Heads	131
7.6	Hypnotic Invariance in the EEG-PD-NLME Model	138

7.6.1	Methods	138
7.6.2	Results	139
7.7	Discussion	142
8	Discussion	144
8.1	Overview of Findings	144
8.2	Comparative Analysis	145
8.3	Validation Criteria	147
8.3.1	Postoperative Pain Prediction (Category A)	147
8.3.2	OpenIbis (Category B)	150
8.3.3	Start of Surgery (Category C)	152
8.3.4	Fentanyl Bolus Correlations (Category D)	153
8.3.5	Explainability Analysis (Category E)	154
8.3.6	Validation Criteria Summary	156
8.3.7	Model Summary	156
8.4	Requirements for an Index of Nociception	157
8.5	Limitations	158
8.5.1	Dataset	158
8.5.2	Modelling	159
8.5.3	Validation Framework	160
8.5.4	Causality	160
9	Conclusion	162
9.1	Contributions	162
9.2	Future Work	163
A	Unsupervised Learning Supplementary Material	165
B	EEG-PD-NLME Loss Derivations	172
B.1	Proportional Model MLE Loss Derivation	172
B.2	Absolute Model MLE Loss Derivation	175
C	EEG-PD-NLME Supplementary Material	177
D	Representation Learning Supplementary Material	179

List of Figures

2.1	Pain transmission pathways: (a) is an uninterrupted pain pathway, (b) is a nociception pathway under general anaesthesia with insufficient analgesia, and (c) is a nociception pathway with sufficient analgesia.	8
2.2	Schematic of a two-compartment PK model.	9
2.3	Schematic of a two-compartment PK model with effect-site equilibration.	10
2.4	A DAG showing the primary confounding pathway in the drug titration paradox (the drug levels and observed outcome measurement are both affected by baseline outcome measurement). .	12
2.5	PRISMA-style flow diagram summarising the systematic literature review process used to identify EEG-based nociception monitoring studies.	26
3.1	‘No cleaning’ dataset variant.	42
3.2	‘Standard cleaning’ dataset variant.	43
3.3	‘EEGdenoiseNet cleaning’ dataset variant.	44
3.4	Distributions of key patient parameters in the combined dataset. .	46
3.5	Distributions of the mean values of key time-varying parameters. .	48
4.1	Training and validation curves for the best-performing pain score category prediction model.	62
4.2	The best-performing model applied to three example EEG recordings from the test dataset.	63
5.1	UMAP output for: Top row: 1 s STFT dataset, 10 s STFT dataset, 30 s STFT dataset. Bottom row: 60 s STFT dataset, 120 s STFT dataset, SingLEM dataset. Coloured by density. .	68
5.2	UMAP output for: Top row: 1 s STFT dataset, 10 s STFT dataset, 30 s STFT dataset. Bottom row: 60 s STFT dataset, 120 s STFT dataset, SingLEM dataset. Coloured by OpenIbis. .	69
5.3	UMAP output for: Top row: 1 s STFT dataset, 10 s STFT dataset, 30 s STFT dataset. Bottom row: 60 s STFT dataset, 120 s STFT dataset, SingLEM dataset. Coloured by patient ID (for a sample of five patients).	70

5.4	PCA component 6 applied to three example EEG recordings from the test dataset.	73
5.5	PCA component 6 loadings (inverted).	74
5.6	HMM state characterisation across all patients: fractional occupancy (left) and transition probability matrix (right).	78
5.7	Inverted fractional occupancy of state 5 applied to three example EEG recordings from the test dataset.	80
6.1	Overview of the EEG-PD-NLME model.	85
6.2	Bolus prediction examples and training curves from Experiment 1.	97
6.3	Bolus prediction examples and training curves from 120-second ResNet-18 model.	106
6.4	ResNet-18 model applied to three example EEG recordings from the test dataset.	107
6.5	GradCAM visualisation of EEG spectrograms.	109
6.6	Violin plots for random effect distributions across each pain category.	110
6.7	Bolus prediction examples and training curves from 120-second ResNet-18 model with pre-surgical regularisation ($\lambda_b = 0.5$). . .	112
6.8	ResNet-18 model with additional training on pre-surgical data applied to three example EEG recordings from the test dataset. .	113
7.1	Overview of model with two adversarial heads.	121
7.2	Directed acyclic graph illustrating the hypothesised causal relationships during general anaesthesia.	121
7.3	UMAP projection of the learned latent representations for ten randomly selected patients. Each point corresponds to a single sample, with colours indicating patient identity.	126
7.4	UMAP projection for partial patient-invariant model with $\lambda_p = 1$. Each point corresponds to a single sample, with colours indicating patient identity.	129
7.5	Train and validation loss, validation AUROC, and adversarial classification accuracy.	130
7.6	Train and validation loss, validation AUROC, and adversarial regression loss.	131
7.7	Baseline discrete classifier applied to three example EEG recordings from the test dataset.	136
7.8	Discrete classifier with adversarial training ($\lambda_p = 0$, $\lambda_d = 100$) on the same EEG recordings.	137
7.9	GradCAM visualisation of EEG spectrogram predictions with adversarially trained discrete classifier.	138
7.10	EEG-PD-NLME model with adversarial training ($\lambda_d = 100$) example plots.	141

8.1	Comparison matrix of each of the evaluated indices, with pair-wise Pearson r calculations.	146
C.1	Experiment 1 random effect distribution.	177
C.2	Random effect distribution for 120-second ResNet-18 model. . .	178
C.3	Random effect distribution for 120-second ResNet-18 model with pre-surgical regularisation ($\lambda_b = 0.5$).	178

List of Tables

2.1	Core components of general anaesthesia and representative agents.	5
2.2	EEG frequency bands and their anaesthetic relevance.	17
2.3	Ramsay Sedation Scale (RSS).	29
2.4	Summary of ML EEG approaches for nociception detection, including dataset, training target, model, and validation status. Internal validation denotes models that were trained with an internal train-test split.	36
3.1	PK parameter values used for effect-site concentration modelling.	41
3.2	Summary of constituent datasets included in this study, combined to give a total dataset with 1,541 hours of EEG data across 567 patients.	47
3.3	Baseline characteristics stratified by pain group.	49
4.1	Performance of random forest (RF) and logistic regression (LR) models for postoperative pain classification.	52
4.2	Validation metrics and expected outcomes for assessment of the proposed pain index.	53
4.3	Summary of test set results for constant index.	58
4.4	Summary of test set results for OpenIbis.	58
4.5	Best validation loss (MAE in brackets) across segment sizes and preprocessing pipelines. The pipeline with the lowest MAE is in bold .	61
4.6	Summary of test set results for the best-performing pain score category prediction model.	62
5.1	Summary of train/validation set results for PCA component 6 from 60-second dataset.	72
5.2	Summary of test set results for PCA component 6 from 60-second dataset.	72
5.3	Summary of train/validation set results for inverted fractional occupancy of state 5.	79
5.4	Summary of test set results for inverted fractional occupancy of state 5.	81

6.1	Experiment 2 results for the proportional and absolute models across loss functions.	98
6.2	Experiment 3 results across input types.	99
6.3	Experiment 4 results across f_ϕ architectures.	100
6.4	Experiment 5 results comparing standard and EEGdenoiseNet-cleaned datasets.	101
6.5	Experiment 6 results across EEG segment lengths, using a ResNet-18 f_ϕ architecture.	102
6.6	Experiment 7 results across learning rate configurations for fixed effects (ϕ) and random effects (η), denoted $\text{lr}_\phi/\text{lr}_\eta$	102
6.7	Experiment 8 results across weight decay configurations applied to the random effects η and fixed effects ϕ	103
6.8	Experiment 9 results across baseline EEG integration strategies.	104
6.9	Summary of test set results for EEG-PD-NLME with ResNet-18 f_ϕ and 120-second input.	108
6.10	Pre-surgical regularisation experiment results across values of the balancing hyperparameter λ_b , using the ResNet-18 f_ϕ architecture with 120-second STFT input.	111
6.11	Summary of test set results for EEG-PD-NLME with ResNet-18 f_ϕ and 120-second input, trained with additional pre-surgical data.	111
7.1	Experimental results across different values of C_f	125
7.2	Validation probe performance for continuous variables. Lower MAE and MSE indicate better predictive performance.	126
7.3	Validation probe performance for categorical variables.	127
7.4	Effect of adversarial patient-invariance strength (λ_p) on task performance and patient-identification probe performance. Results are reported as mean $\pm$ standard deviation over ten random seeds. Lower probe metrics indicate greater patient invariance.	128
7.5	Effect of adversarial hypnotic-depth invariance strength (λ_d) on task performance and hypnotic-depth probe performance. Results are reported as mean $\pm$ standard deviation over 10 random seeds. Higher probe errors indicate greater invariance to hypnotic depth.	132
7.6	Patient identity and hypnotic depth probe performance for each adversarial configuration. Lower patient probe metrics indicate greater patient invariance; higher depth probe errors indicate greater depth invariance.	134
7.7	Summary of test set results for baseline classifier with no adversarial training.	134
7.8	Summary of test set results for discrete classifier with $\lambda_p = 0$, $\lambda_d = 100$	135

7.9	Ridge probe MAE and MSE on OpenIbis predictions as a function of hypnotic depth adversarial weight λ_d .	139
7.10	Summary of test set results for EEG-PD-NLME model with $\lambda_d = 100$ adversarial head.	140
8.1	A1. Pain prediction (index): AUROC with 95% CI on the test set. Significant results where the entire CI is above 0.5 are in bold .	147
8.2	A2. Pain prediction (nested): Δ AUROC relative to a baseline linear model. Significant results at $p < 0.05$ with positive Δ AUROC are in bold .	148
8.3	B1. OpenIbis inter-patient correlation: Pearson r and Spearman ρ between model output and OpenIbis score, aggregated across patients.	150
8.4	B2. OpenIbis intra-patient correlation: mean within-patient Pearson r and Spearman ρ between model output and OpenIbis score.	151
8.5	C1. Start of surgery: mean model output in the pre-surgical and surgical periods, and their difference (Surg.–Pre.). Any results where the difference is greater than 5 points are in bold .	152
8.6	D1. Fentanyl bolus: mean within-patient Pearson r and Spearman ρ between model output and fentanyl effect-site concentration following bolus events. Any correlations < -0.05 are in bold .	153
8.7	E. Explainability: regression coefficients for delta, alpha, and beta band power in a linear model predicting model output from EEG spectral features. All coefficients significant at $p < 0.001$.	154
8.8	Regression coefficients for delta, alpha, and beta band power in a linear model predicting model output from EEG spectral features, stratified by OpenIbis level. All coefficients significant at $p < 0.001$ unless otherwise noted.	155
A.1	Complete exploration for PCA components 1–10 for one-second STFT.	166
A.2	Complete exploration for PCA components 1–10 for ten-second STFT.	166
A.3	Complete exploration for PCA components 1–10 for 30-second STFT.	166
A.4	Complete exploration for PCA components 1–10 for 60-second STFT.	167
A.5	Complete exploration for PCA components 1–10 for 120-second STFT.	167
A.6	Complete exploration for EFA factors 1–10 for one-second STFT.	168
A.7	Complete exploration for EFA factors 1–10 for ten-second STFT.	168
A.8	Complete exploration for EFA factors 1–10 for 30-second STFT.	168

A.9	Complete exploration for EFA factors 1–10 for 60-second STFT.	169
A.10	Complete exploration for EFA factors 1–10 for 120-second STFT.	169
A.11	Complete exploration for Fractional Occupancies and HMM statistics.	170
A.12	Complete exploration for transition matrix statistics.	171
D.1	Baseline binary classification model.	180
D.2	Binary classification model with varying adversarial head strength (λ_p and λ_d).	180
D.3	NLME with varying adversarial head strength.	180

Glossary

anaesthesia A deliberately induced and reversible loss of sensation and/or awareness.

analgesia The reduction or elimination of nociceptive or painful sensation, typically achieved through analgesic drugs.

effect-site concentration The concentration of a drug at the point where it exerts its biological activity.

electroencephalography A non-invasive technique for measuring electrical activity in the brain using electrodes placed on the scalp.

fentanyl A potent synthetic opioid analgesic used for severe pain.

haemodynamic Relating to the dynamics of blood flow throughout the body.

hypnotic The dimension of anaesthesia that mediates loss of consciousness and suppression of memory formation.

nociception The neural encoding of noxious (painful) stimuli, which can occur when a patient is unconscious. It is distinct from pain, which is the full sensory and emotional experience that occurs in response to noxious stimuli.

opioid A class of analgesic drugs, typically used for moderate to severe pain.

pharmacodynamics The study of the relationship between effect-site concentration and drug effect.

pharmacokinetics The study of how drug concentration varies over time within the body.

titration The process of adjusting the dosage of a drug to achieve a desired effect.

Acronyms

AUROC area under the receiver operating characteristic curve.

BIS bispectral index.

EEG electroencephalography.

EEG-PD-NLME EEG-informed pharmacodynamic non-linear mixed effects.

EFA exploratory factor analysis.

EMG electromyography.

HMM hidden Markov model.

MAC minimum alveolar concentration.

MAE mean absolute error.

ML machine learning.

MSE mean squared error.

PCA principal component analysis.

PD pharmacodynamics.

PK pharmacokinetics.

STFT short-time Fourier transform.

Chapter 1

Introduction

1.1 Motivation

Surgical care is a critical component of healthcare, with over 300 million surgical procedures occurring worldwide each year [1]. Surgical procedures inevitably involve painful stimuli and tissue damage. General anaesthesia enables patients to tolerate this [2]. There are three primary dimensions of anaesthesia: the hypnotic dimension (controlling level of consciousness and ability to form retrievable memories), the nociceptive dimension (controlling nociception, the response of a patient to painful stimuli), and the paralytic dimension (controlling the ability of a patient to move). Each dimension is controlled by different drugs given by an anaesthetist during surgery [3].

An anaesthetist is able to control the state of a patient along each dimension via drug titration. Drug titration involves increasing or decreasing the dose of a drug in response to clinical signs that allow a clinician to infer the status of the patient [4]. Electroencephalography (EEG) is a non-invasive technique for the measurement of neural activity [5], and it is frequently used as a monitoring tool during surgery for the purposes of drug titration [6]. In particular, single-channel EEG is typically used in anaesthesia. It requires only a few electrodes on the patient's forehead, which is non-invasive and easy to deploy in an operating theatre setting [7].

Although inspection of the raw EEG waveform can give clinically useful information, substantial work has been done on generating processed indices that take the EEG as an input and output a numeric value accurately reflecting the status of the patient. The most popular of these indices is the bispectral index (BIS) monitor, which outputs an index representing the hypnotic state of the patient. It ranges from 0 to 100, with 0 indicating no brain activity, and 100 indicating full consciousness [8].

Much work has been done by commercial and research groups on developing a reliable index of hypnotic depth. Relatively less work has been done on developing an index of the nociceptive dimension, which represents a patient's responses to painful stimuli while anaesthetised. Although it is correlated

with hypnotic depth, it is a distinct dimension and can move separately to the hypnotic dimension [3]. This means that it requires its own titration process.

The nociceptive dimension of anaesthesia is primarily mediated by analgesic drugs given during surgery, such as the opioid fentanyl. Accurate titration of these drugs is desirable due to the deleterious effects of under- and over-titration of analgesic agents. Inadequate analgesia leads to increased postoperative pain [9], a higher risk of developing chronic pain, prolonged opioid use, and an increased risk of heart strain and immune suppression due to excessive sympathetic activation [10]. Excessive analgesia can lead to delayed emergence from general anaesthesia, delirium, nausea and vomiting, ileus, and potentially opioid-induced hyperalgesia (where opioids paradoxically enhance pain) [11].

Haemodynamic parameters such as heart rate and blood pressure are typically used to titrate levels of analgesia, but these are highly confounded endpoints [3]. It is known that nociception has an effect on EEG in anaesthetised patients [12]. An EEG-based monitor of nociception thus has the potential to provide more accurate opioid titration during surgery, and therefore reduced postoperative pain and improved postoperative outcomes for patients. Existing EEG-based nociception algorithms either lack comprehensive evidence supporting their use [13, 14, 15, 16], or are opaque proprietary algorithms [17]. This study aims to fill this research gap by creating transparent algorithms for nociception assessment, and comprehensively validating them.

Artificial intelligence and machine learning (ML) are increasingly being applied to EEG data across a broad range of applications [18], such as brain-computer interfaces [19], seizure detection [20], diagnosis of mental illnesses [21], and sleep staging [22]. Similar methods could be applied to intraoperative EEG in order to generate models that can estimate the levels of nociception during surgery. ML methods are able to learn complex and non-linear representations of EEG data. Deep learning (DL) methods in particular have shown substantial promise for creating useful feature representations for downstream tasks [23]. This has the potential to outperform traditional EEG spectral feature extraction techniques.

1.2 Aims and Contributions

There is a current lack of comprehensive and principled evaluation methods for intraoperative nociception, as well as a lack of well-described and well-validated algorithms for this purpose. The two key aims of this thesis are therefore:

- Develop and evaluate a unified validation framework for potential indices of intraoperative nociception.
- Apply and develop ML modelling approaches for estimation of intraoperative nociception, and then evaluate these approaches against the unified validation framework.

The key difficulty with estimation of intraoperative nociception is the lack of ground truth data describing the amount of nociception a patient is experiencing while anaesthetised. One measure that is frequently used to evaluate nociceptive indices is their ability to predict the pain score that a patient will report when they emerge from anaesthesia, by comparing pain score category to the value of the index at the end of surgery [16, 24, 25]. However, postoperative pain is only one of several possible surrogate measures, and may not be a strong reflection of intraoperative nociception. Other criteria, such as response to intraoperative events and response to analgesic drug levels, may also provide useful validation information. This thesis therefore aims to establish a unified validation framework that combines these factors to provide a more comprehensive and consistent method to evaluate putative nociception indices.

The lack of gold-standard labels also poses a difficulty for training models. Surrogate measures are often highly confounded by factors such as patient identity or hypnotic depth. Novel approaches are therefore required to develop models that can reliably estimate levels of intraoperative nociception.

The dataset used in this study consists of 1,541 hours of intraoperative EEG data across 567 patients. It is the combination of data from three previous studies, with data collected from surgeries at Waikato Hospital and North Shore Hospital. The dataset also includes drug dosage data and surgical event data (most importantly the timing of the start of surgery).

The results of this thesis yielded two key contributions:

- Two indices (PCA-6 and NLME+adv.) were identified that performed well on the unified validation framework, and could be taken forward for further development and prospective validation. These are well-described models, and in the case of PCA-6, inherently interpretable.
- A unified validation framework was constructed, and it was found that postoperative pain could not be predicted well by any nociception index investigated in this study, including PCA-6 and NLME+adv., despite these models satisfying all other validation criteria. This suggests that postoperative pain scores may not correlate well with end-surgical nociception.

1.3 Structure of Thesis

The thesis is structured as follows:

- Chapter 2 provides some background context to anaesthesia, nociception, and pharmacology. The current state of EEG monitors in anaesthesia is reviewed, along with current techniques used in EEG analysis. A comprehensive literature review is conducted analysing current evidence

around using EEG to estimate nociception, including existing ML models made for this purpose.

- Chapter 3 describes the dataset and tools used in this thesis. It describes how the EEG data were cleaned and drug concentrations were calculated. A brief exploratory data analysis is performed.
- Chapter 4 introduces the unified validation framework used in this thesis. It demonstrates how this framework can be used on some basic predictors of nociception from EEG, including a model that uses postoperative pain scores.
- Chapter 5 explores unsupervised learning techniques for generating an index of nociception. This includes principal component analysis and exploratory factor analysis. Hidden Markov models are also applied for estimation of EEG microstates.
- Chapter 6 introduces and evaluates a novel supervised learning method, the EEG-informed pharmacodynamic non-linear mixed effects model, which combines standard ML techniques with modelling techniques from pharmacology in order to estimate nociception from EEG using opioid concentrations.
- Chapter 7 explores the use of adversarial heads for generation of unfounded representations of EEG, which can then be used for estimation of nociception.
- Chapter 8 compares and discusses the models, and discusses the unified validation framework.
- Chapter 9 concludes the thesis.

Appendix A contains supplementary tables for Chapter 5. Appendix B contains the loss function derivations for Chapter 6, and supplementary graphs for this chapter are in Appendix C. Appendix D contains supplementary graphs for Chapter 7.

Chapter 2

Background

2.1 Anaesthesia

General anaesthesia is a controlled, reversible loss of consciousness and movement which enables patients to tolerate surgery [2]. The key dimensions of the general anaesthetic state are hypnosis (including impairment of memory formation), analgesia, and immobility. Different agents are used to induce each component of general anaesthesia, as each component is modulated by a partially distinct but overlapping set of receptors and neuronal pathways (some of which have not yet been precisely elucidated) [26]. Although there are correlations between these dimensions (for instance, analgesic drugs often have a contribution towards the hypnotic effect), it is possible for dissociation to occur between components (for instance, a state where a patient has a sufficient hypnotic state but insufficient analgesia) [3]. Table 2.1 summarises the tripartite model of anaesthesia and some of the common agents involved.

Table 2.1: Core components of general anaesthesia and representative agents.

Dimension	Common agents	Effect
Hypnotic	Propofol, desflurane, sevoflurane	Unconsciousness, amnesia
Analgesic	Fentanyl, remifentanyl	Reduces pain and nociception
Immobilisation	Rocuronium, atracurium	Immobilisation

Unconsciousness in surgery is induced and maintained by hypnotic agents, which can come in both intravenous and volatile (inhalational) forms [27]. The most frequently used intravenous anaesthetic is propofol, which is thought to act primarily by modulating gamma-aminobutyric acid (GABA) receptors [28]. Frequently used volatile anaesthetics include desflurane, sevoflurane, and isoflurane, which also act via multiple targets including potentiation of GABA receptors and inhibiting the N-methyl-D-aspartate (NMDA) receptor [27]. The result is disruption of cerebral cortical activity, both through direct inactivation, and indirect inactivation via shutdown of ascending arousal networks in the brain stem and hypothalamus. The thalamus is also significantly affected

by general anaesthesia. In normal physiology, the pathway from thalamus to cortex is responsible for the transmission of sensory information to the cerebral cortex (and is therefore responsible for awareness of sensory information). However, in anaesthesia, the thalamus is placed into a suppressed state, effectively isolating the cortex from any sensory information [29]. Hypnotic agents also interfere in hippocampal memory consolidation by inhibiting long-term potentiation, therefore inhibiting synaptic plasticity. The result of this is anterograde amnesia, where new memories cannot be formed while anaesthetised [30].

Adverse effects of hypnotic agents (especially if dosing is excessive) can include hypotension, respiratory depression, and postoperative delirium. Conversely, if dosing is insufficient, patients may experience intraoperative awareness (i.e. form retrievable memories). Anaesthetists therefore aim to titrate hypnotic agents such that the risk of these negative outcomes is minimised [28, 31].

The analgesic dimension of general anaesthesia is typically driven by opioid analgesics, such as fentanyl and remifentanyl [3]. Opioids act via mu, delta, and kappa opioid receptors, which cause a reduction in presynaptic neurotransmitter release. Opioids have several undesirable side effects, such as respiratory depression, constipation, and postoperative nausea and vomiting, which worsen if used in greater quantities. However, insufficient opioid administration can result in excessive intraoperative nociception and postoperative pain. The levels of analgesic therefore are titrated by the anaesthetist (typically using haemodynamic responses such as blood pressure and heart rate) in order to ensure efficacy whilst avoiding adverse effects [3].

The other key dimension of anaesthesia is immobility, which is achieved via the administration of a muscle relaxant. These are nicotinic anticholinergic receptor antagonists such as rocuronium and atracurium, which result in a neuromuscular blockade [27].

2.2 Nociception and Pain

Pain is a complex, multidimensional phenomenon consisting of four processes: transduction, transmission, perception, and modulation. In transduction, peripheral nociceptive sensory fibres (either A δ myelinated fibres or C unmyelinated fibres) detect any noxious stimuli that could result in tissue damage. This information is transmitted to the dorsal horn of the spinal cord, after which it is transmitted to the brain via ascending spinal tracts (primarily the spinothalamic tract). Signals are then transmitted to subcortical structures, such as the brainstem and thalamus. Conscious perception of pain then results from transmission of these signals to cortical regions of the brain. At each stage, various factors can modulate the pain processing and transmission process, such as endogenous and exogenous opioids [32].

Many noxious stimuli occur during surgery that can trigger nociception.

These stimuli are primarily mechanical, such as intubation at the start of surgery, skin incision, retraction, or drilling. Patients can be exposed to thermal noxious stimuli through the heat generated by a diathermy probe, and they can be exposed to chemical noxious stimuli through mediators released due to tissue damage [33, 34].

However, under general anaesthesia, the pain perception pathway is altered. There is some attenuation of nociceptive signals at the level of the spinal cord, although hypnotic agents do not necessarily suppress ascending signals entirely (particularly for more severe noxious stimuli). A more significant disruption occurs at the cortical level (particularly with regard to the thalamocortical tract). The result is that a patient will lack awareness of the sensation of pain (due to a lack of cortical processing of the signal), but subcortical processing will remain and may manifest as an autonomic response (such as a rise in heart rate or blood pressure) [35]. In contrast, the analgesic component of general anaesthesia causes suppression of ascending pathways at the level of the spinal dorsal horn and the spinothalamic tract [27]. For the purposes of this thesis, the term nociception will be used to refer to the neural encoding of noxious stimuli (at peripheral, subcortical, and cortical levels), whereas pain will refer to the subjective sensory and emotional experience in response to noxious stimuli. Although nociception shares many pathways with conscious pain processing, it cannot be described as equivalent to pain due to a lack of awareness and cortical processing [35]. Figure 2.1 summarises the differences between unimpeded pain transmission, nociception under general anaesthesia without sufficient analgesia, and nociception under general anaesthesia with sufficient analgesia. Note that this is an idealised model, and depending on the level of hypnotic and the noxious stimulus, the response to noxious stimulus under pure hypnotic will lie somewhere between (b) and (c).

Despite nociception not being strictly equivalent to pain, excessive intraoperative nociception is nevertheless associated with several deleterious effects. Uncontrolled nociception can cause an excessive surgical stress response. This causes excessive activation of the hypothalamic-pituitary-adrenal axis, as well as the sympatho-adrenomedullary axis, causing the release of cortisol, catecholamines, and inflammatory cytokines [36]. An excessive surgical stress response is associated with cognitive dysfunction, cardiovascular instability, increased postoperative complications, and increased risk of death [37].

Inadequate intraoperative analgesia is also associated with more severe pain postoperatively [9]. There are many sequelae of poorly controlled postoperative pain. It can increase the risk of myocardial infarction and ischaemia, it can cause hypoventilation and therefore increase the risk of pneumonia, it can reduce gastric motility and cause ileus, and it can impair immune function and wound healing. It also has significant psychiatric sequelae such as increased risk of anxiety and low mood. Longer term effects include central sensitisation and an increased risk of developing chronic pain, and prolonged opioid use (which can lead to a variety of adverse effects, including opioid dependence)

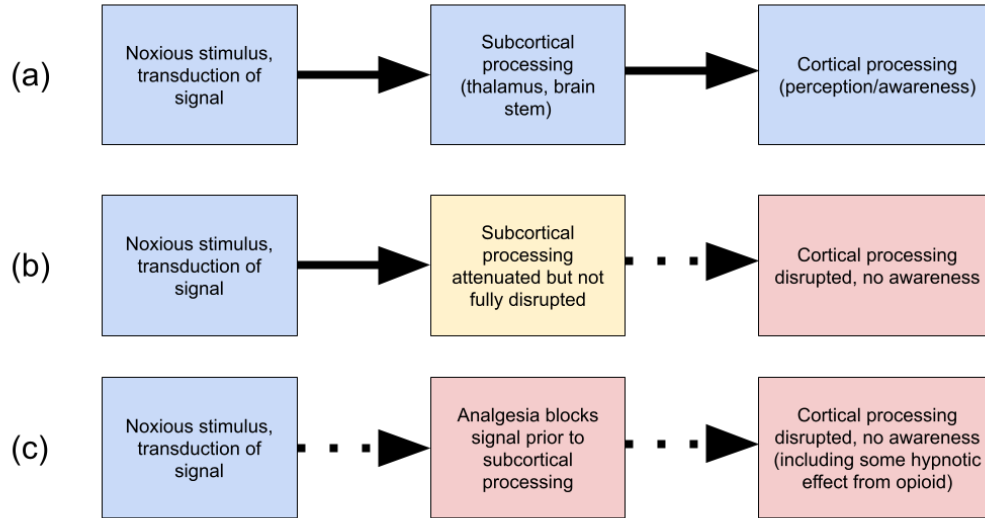


Figure 2.1: Pain transmission pathways: (a) is an uninterrupted pain pathway, (b) is a nociception pathway under general anaesthesia with insufficient analgesia, and (c) is a nociception pathway with sufficient analgesia.

[10]. Postoperative pain is ultimately associated with a reduction in quality of life, therefore it is in the best interests of clinicians to reduce it as much as possible [38]. At the same time, it is important to avoid giving analgesia where it is not required. This is due to the adverse effects of opioids and other analgesic agents, such as respiratory depression, nausea, and vomiting [3].

There is significant variability in how different patients perceive the same pain stimulus [39], mediated by differences at the level of the cerebral cortex [40]. It has been shown that pain perception varies with age [41], sex [42], and psychiatric or cultural factors [39]. Standard pain scoring (a scale where a patient is asked to rate their pain on a scale from 0 to 10) is therefore noisy and subjective [43]. It has been shown that the same patient may report varying pain scores when exposed to the same stimulus [44].

2.3 Pharmacokinetics and Pharmacodynamics

Pharmacokinetics (PK) and pharmacodynamics (PD) are a paradigm that allows for the estimation of drug effects based on dosages and dose timing (for instance, allowing the antinociceptive effect of fentanyl to be estimated from the size and timing of given intravenous doses). PK describes how the levels of drug vary within the body, usually with the goal of estimating the concentration of drug at the effect-site. PD describes what the effects of a drug are, given a particular effect-site concentration. PK and PD parameters vary from person to person, depending on age, weight, organ function, and other

subject-specific features [45].

The simplest PK model is a single compartment model, which assumes the body consists of a single well-mixed space, and drug instantaneously distributes throughout the body upon administration. Given a clearance Cl (which represents how quickly drug is removed from the body), and a volume of distribution V , the concentration C of drug at a given time t after administration can be modelled by:

$$\frac{dC}{dt} = -\frac{Cl}{V}C \quad [46]$$

However, this model is not sufficient for modelling fentanyl. It has been shown that fentanyl also redistributes into the periphery [47]. Rather than modelling the body as a single compartment, a two-compartment model can be used, where the body consists of a central and a peripheral compartment, as shown in Figure 2.2. Drug is considered to be administered into and eliminated from the central compartment.

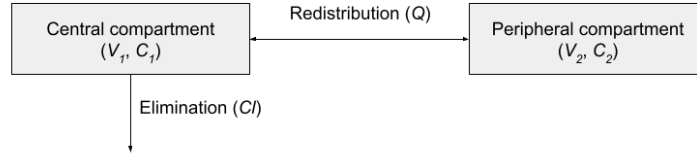


Figure 2.2: Schematic of a two-compartment PK model.

Given clearance from the central compartment Cl , a redistribution constant Q , a central compartment with volume V_1 and a peripheral compartment with volume V_2 , the concentrations in each compartment C_1 and C_2 at time t are modelled by the following:

$$\begin{aligned} \frac{dC_1}{dt} &= -\frac{Cl}{V_1}C_1 - \frac{Q}{V_1}(C_1 - C_2) \\ \frac{dC_2}{dt} &= \frac{Q}{V_2}(C_1 - C_2) \end{aligned} \quad [48]$$

Finally, it is necessary to model the time that it takes for fentanyl to reach the effect-site after it has been administered to the central compartment [47]. We therefore introduce a further constant, the effect-site equilibration rate constant k_{e0} to model the delay between administration and drug effect. The final effect-site concentration C_e is given by the following:

$$\frac{dC_e}{dt} = k_{e0}(C_1 - C_e)$$

Figure 2.3 summarises the two-compartment model with additional effect-site equilibration modelling [49].

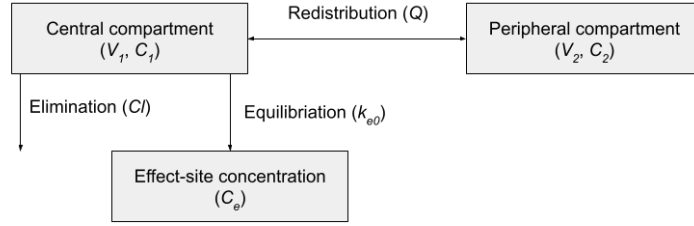


Figure 2.3: Schematic of a two-compartment PK model with effect-site equilibration.

This modelling is standard for opioid PK, as it appropriately models the multi-exponential decay in effect-site concentration that is observed with opioids [47, 50]. Exact solutions to these systems of ordinary differential equations (ODEs) become computationally costly to calculate with increasing model complexity, so numeric integration methods are preferred to provide estimates of effect-site concentration at each time point after drug administration [51].

After determining an effect-site concentration at each time step, the actual effect of a drug is typically modelled by a sigmoid PD model [50]. The Hill model is used for this, and it is parameterised by three values:

- E_{max} : the maximum possible effect that the drug can cause (representing the effect when all receptors have been activated).
- C_{50} : the effect-site concentration of drug required to generate an effect half as large as the maximal effect.
- γ : a constant determining the steepness of the sigmoid curve.

Given these, the effect of a drug (given some effect measure E) is given by:

$$E = \frac{E_{max} C_e^\gamma}{C_{50}^\gamma + C_e^\gamma}$$

A different PD paradigm is used for volatile anaesthetic agents. For these drugs, the primary metric used is the minimum alveolar concentration (MAC). One MAC is defined as the minimum concentration of volatile anaesthetic, at the alveoli, at which 50% of patients will not move in response to surgical stimuli. The MAC can therefore serve as a measure for the potency of volatile anaesthetics, as well as an agent-agnostic measure of volatile anaesthetic potency [52]. Amnesia occurs at approximately 0.25 MAC, unconsciousness is at approximately 0.5 MAC, and 1.2 MAC causes immobility in about 95% of patients [53].

In practice, end-tidal partial pressure of volatile anaesthetic is used to measure the MAC, drawing from the assumption that end-tidal partial pressures approximate the alveolar pressures [53]. Partial pressures in the alveoli

equilibrate with arterial blood, which then equilibrates with the brain. The end-tidal partial pressure of volatile anaesthetic therefore is used to determine the effect-site partial pressure [54]. The calculation of this can use a single equilibration constant k_{e0} for the entire diffusion process. The following equation shows how the end-tidal partial pressure C_{ET} can be used to model the effect-site concentration C_e . This can then be converted to a MAC, with the conversion factor depending on the potency of volatile anaesthetic being used [55].

$$\frac{dC_e}{dt} = k_{e0}(C_{ET} - C_e)$$

2.3.1 Causality and the Drug Titration Paradox

Analysis of pharmacometric data is often implicitly causal. It is therefore important to understand the causal assumptions underlying these analyses, particularly when ML is used to model pharmacometric data.

Consider, for instance, outcome Y_i , which refers to an outcome for patient i . This outcome is a random variable, and it implicitly encodes a probability distribution $P(Y_i = y)$. Now, consider a treatment $A_i = a$ which is applied to this patient. There are two distinct ways of interpreting the effect of this treatment:

- $Y_i|A_i = a$ refers to the outcome that occurs given that an observer sees that treatment a has been applied to patient i .
- Y_i^a refers to the outcome that occurs if a is imposed as an intervention. This is known as a potential outcome, and there is a direct causal link between Y_i^a and a .

$Y_i|A_i = a$ and Y_i^a are different outcomes. The former is affected by confounding, and any differences in the outcome are not necessarily solely attributable to the direct causal effect of a . The latter is affected only by a direct causal path [56].

In order to draw causal inferences about the effects of specific interventions, such as levels of drug, it is important to understand the causal confounders involved in a given scenario. Directed acyclic graphs (DAGs) can be used to express the confounders, treatments, and outcomes involved in a scenario, as well as any causal relationships between them (and the direction of those causal relationships) [56].

One important source of confounding that arises frequently in anaesthesia is titration. This is the process where the amount of drug given to a patient is varied such that a target outcome measure is achieved. Naive interpretation of data from drug titration scenarios leads to a drug titration paradox, where greater amounts of drug are associated with a reduced effect.

This negative correlation can be explained by imperfect titration: patients who receive greater levels of drug are more likely to have lower levels of the

outcome effect measure prior to drug administration (or there would be signs indicating that the levels of this outcome effect measure would soon fall). If the amount of drug given falls short of perfect titration, this negative correlation will persist in the observed data. Figure 2.4 is a DAG illustrating this scenario.

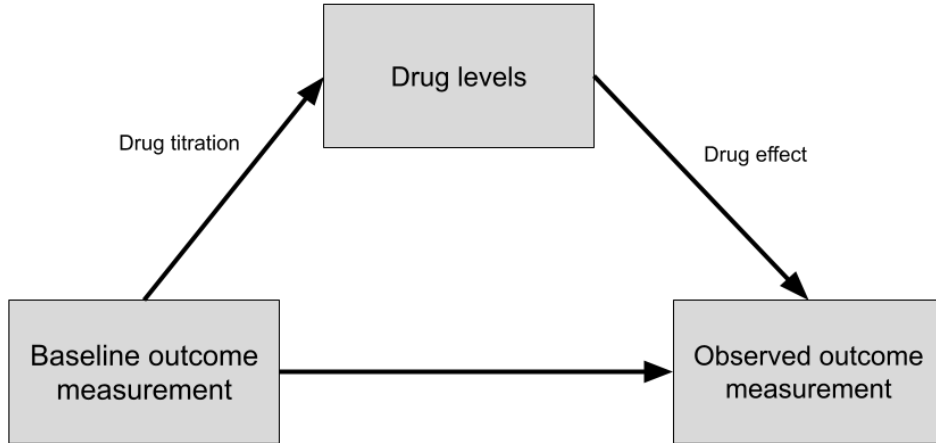


Figure 2.4: A DAG showing the primary confounding pathway in the drug titration paradox (the drug levels and observed outcome measurement are both affected by baseline outcome measurement).

However, titration does not always cause a negative correlation between drug and effect measure. The direction of correlation is dependent on how well-titrated administration of the drug is to the observed outcome. If titration is perfect, no correlation would be observed between the level of drug and the outcome. If titration is excessive, then a positive correlation would be observed between the level of drug and the observed outcome [57].

The drug titration paradox can exist at both an inter-individual and intra-individual level. At an inter-individual level, the drug titration paradox is akin to a selection bias, where individuals with a higher propensity to have low values of the outcome measure are favoured for higher doses of drug (and vice-versa). Where the outcome measure varies significantly within an individual, drug titration can occur (and therefore the drug titration paradox may arise) [58].

Mixed-effects modelling has been used to address the drug titration paradox, with inconsistent results when used on simulated data [59, 60]. A single intercept per patient is able to account for inter-individual variability, although in situations where the drug titration paradox exists at the intra-individual level, mixed-effects modelling is unable to uncover the true causal effect [58].

Other efforts to reduce the effect of the drug titration paradox include the addition of a second intercept to account for intra-occasion variability. A stochastic differential equation model with a Kalman filter has been used to model the underlying measure (in this case, pain) as a Wiener process.

Autocorrelation between temporally adjacent residuals has also been used to reduce the effect of the drug titration paradox [58]. There have also been efforts to use marginal structural models with inverse probability of treatment weighting (IPTW) to account for varying propensity of drug administration [59].

2.4 EEG Monitoring in Anaesthesia

Electroencephalography (EEG) is a non-invasive technique used for measuring neural activity. It involves the placement of one or more electrodes on the scalp. Postsynaptic action potentials (in addition to other neuronal electrical events) can produce voltage changes that are detectable from these electrodes [5], although these signals are attenuated as they move through the skull and scalp. Large and slow voltages, typically generated by large groups of synchronously firing neurons, are therefore more dominant in recorded EEG signals [29]. EEG can provide useful information about the neurophysiology of the subject, and is therefore used in a variety of applications across neuroscience and medicine, including in intraoperative monitoring of patients under general anaesthesia [5].

In anaesthesia, EEG is typically used as a measure for titrating depth of anaesthesia [29]. This is important to avoid under-dosing (which can lead to the formation of retrievable memories), or over-dosing (which can lead to excessive sedation). Cardiovascular parameters such as heart rate and blood pressure do have some use for monitoring depth of anaesthesia, although they are not reliable signs in isolation [7]. International guidelines therefore encourage the use of EEG monitoring during anaesthesia [6, 61], particularly for total intravenous anaesthesia [62]. EEG gives a more direct indication of cortical activity [7], and it has been shown in multiple studies that use of EEG monitoring during anaesthesia may be associated with lower risk of postoperative delirium and cognitive decline [63, 64].

Assessment of EEG during general anaesthesia is performed both by direct inspection of the waveform by anaesthetists, and by quantitative EEG monitors which process the waveform and output a single numeric index indicating the depth of anaesthesia [7]. The most commonly used depth of anaesthesia monitor is the bispectral index (BIS) monitor (Covidien). As a proprietary algorithm, the exact detail of how the BIS is calculated is not known publicly. It is known that it takes a single EEG channel as input, and combines information from the time domain, frequency domain, and other higher-order features in a multivariable model. The output is a value between 0 and 100, where 0 represents no brain activity, and 100 represents a fully awake patient [8]. A typical aim is to keep the BIS in the range of 40–60 during anaesthesia [65]. As a single index, the BIS is limited in its ability to appraise the state of a patient [66]. It also has weaknesses in particular patient subgroups, such as individuals with neurological impairment or hypothermia, and young children

[67].

There are several other proprietary depth of anaesthesia monitors. The Entropy Module (GE Healthcare) generates two indices: State Entropy (SE) and Response Entropy (RE). SE is calculated using just EEG, and RE is calculated using both EEG and facial electromyography (EMG) [68]. The qCON monitor (Quantum Medical) also provides two indices: qCON and qNOX. These are ML derived scores, which aim to predict anaesthetic depth and the likelihood of response to noxious stimuli [69]. Narcotrend (MonitorTechnik) adopts a classification approach to monitoring depth of anaesthesia, inspired by sleep staging. It classifies the given EEG into one of six stages ranging from A (awake) to F (no brain activity), and gives a BIS-like 0–100 index as a secondary output [70].

A single index is often too reductive for clinical use, so anaesthetists will supplement depth of anaesthesia monitors with their own direct inspection of the raw EEG waveform [7]. This can include recognising changes in certain frequency bands, recognising burst suppression patterns, and recognising EMG interference and other artefacts and noise. It has been shown that after only a small amount of training, anaesthetists can distinguish between awake, sedated, and anaesthetised patients with similar accuracy to a BIS monitor [66].

2.5 EEG Analysis

For the purposes of this study, the discussion will focus on EEG features that can be found in analysis of a single-channel EEG signal.

2.5.1 EEG Transforms

EEG can be analysed in the time, frequency, and time-frequency domains. Time domain analysis involves analysis of the raw waveform, and allows for direct analysis of amplitudes and time-varying patterns. Classical statistical measures, such as mean, variance, and kurtosis, can be applied to time domain EEG segments [71].

Frequency domain analysis of EEG signals typically involves the signal being transformed using a Fourier transform, usually using the Fast Fourier Transform (FFT) algorithm [71]. Given a time domain signal $x(t)$, the continuous Fourier transform is given by:

$$X(f) = \int_{-\infty}^{\infty} x(t)e^{-i2\pi ft} dt$$

The FFT is a fast algorithmic implementation of the discrete Fourier transform, which approximates the continuous Fourier transform. Note that the coefficients of the Fourier transform are complex numbers (i.e. they consist of both a magnitude and a phase) [72].

Other frequently used frequency domain transform methods include Welch’s method (which provides more stable power spectral density estimates), multi-taper methods, and autoregressive modelling [72].

Transforming a signal into the time-frequency domain results in a two-dimensional output which can show changes in the frequency composition of a signal over time. This output can be displayed as a spectrogram [71]. The short-time Fourier transform (STFT) adds a window function $w(t)$ (usually a bell-shaped Hamming window) to the standard Fourier transform, allowing for the Fourier transform to be calculated for a given temporal window. The equation for the STFT is:

$$X(f, \tau) = \int_{-\infty}^{\infty} x(t)w(t - \tau)e^{-i2\pi ft} dt$$

Other methods for transforming signals to the time-frequency domain include the continuous wavelet transform, the Wigner distribution function, and empirical mode decomposition (EMD) [72].

2.5.2 EEG Cleaning

EEG often requires substantial cleaning, as there are several artefacts that can contaminate EEG data. Physiological factors can cause artefacts of three main types:

- Eye movement or blinking, creating an electrooculography (EOG) signal.
- Facial muscle movement, creating an electromyography (EMG) signal.
- Cardiac activity, creating an electrocardiography (ECG) signal [73].

Extrinsic factors can also generate artefacts [73]. This includes artefact generated by electronic equipment and radiation from electronic devices. Motion artefact from breathing, surgery, sweating, or swallowing can also be a significant issue [74].

Many artefact removal techniques for EEG exist, but most are not appropriate for the single-channel setting. Regression methods require extra EOG and EMG reference channels. Blind Source Separation (BSS) techniques, such as Independent Component Analysis (ICA) and Canonical Correlation Analysis (CCA), require multiple-channel data. ICA is an extension of principal component analysis (PCA) designed to separate a source signal from statistically independent noise [73].

The discrete wavelet transform (DWT) and thresholding can be used to clean single-channel data. Firstly, the DWT decomposes the signal into wavelets. The wavelet coefficients are thresholded using a shrinkage rule such as Stein’s Unbiased Risk Estimate (SURE), and the signal is then reconstructed from the remaining wavelets. EMD similarly decomposes the signal into Intrinsic

Mode Functions (IMFs), which are thresholded, and the cleaned signal is reconstructed [75]. Hybrid BSS methods for EEG cleaning can be utilised with single-channel data by using transforms such as the wavelet transform and EMD. The signal is first decomposed into multiple channels using the wavelet transform or EMD. The resulting output is used as the input for a BSS denoising process, such as ICA. The signal then undergoes an inverse BSS and an inverse wavelet/EMD transform, giving a cleaned signal [73].

Adaptive filtering and Wiener filtering have both been used for cleaning of single-channel signals. However, adaptive filtering requires a reference channel, and Wiener filtering requires an estimate of the power spectral densities of the cleaned signal and the noise to be provided [73].

Several deep learning approaches for cleaning EEG data have been attempted. These approaches generally follow a supervised learning paradigm: a network f_θ with unclean input signals X outputs cleaned signals $\hat{Y}$. This is compared against ground truth clean signals Y , and a reconstruction loss function $\mathcal{L}(Y, \hat{Y})$ is minimised across the dataset [76]. Semi-synthetic datasets are often used to generate the clean and unclean pairs that constitute X and Y . For instance, EEGdenoiseNet was trained with a dataset consisting of clean EEG and synthetic noisy EEG deliberately contaminated with EOG and EMG [77].

Supervised deep learning methods for cleaning EEG have included convolutional neural networks (CNNs) (such as EEGdenoiseNet), recurrent neural networks (RNNs), and transformers. Generative adversarial networks (GANs) have also been used for denoising datasets. GANs consist of two networks: a generator (which reconstructs a clean signal from an unclean signal), and a discriminator (which is trained to discriminate whether the generator’s output is indistinguishable from a real clean EEG). The final loss function consists of a combination of the reconstruction loss and an adversarial loss from the discriminator. Autoencoders are an unsupervised approach to EEG denoising, which consist of an encoder that converts the input into a low dimensional latent representation (whilst also filtering out noise), and a decoder that reconstructs the signal. Autoencoders can be used to denoise EEG in the absence of a ground truth dataset [76].

It is important to remove baseline drift and power supply interference. This is typically achieved using a high-pass filter (and de-trending) to remove any low frequency drift, and using a notch filter to eliminate power supply interference [78]. In the context of anaesthesia, a bandpass filter is often utilised, effectively adding a low-pass filter. In anaesthesia, the frequencies of interest are often relatively low (in the alpha and delta range), so removing irrelevant high frequencies can improve the signal-to-noise ratio [79].

2.5.3 Frequency Domain

Frequency domain analysis is a critical part of the analysis of EEG in anaesthetised patients. In conventional EEG frequency domain analysis, the signal undergoes a Fourier transform and a power spectrum is calculated. This power spectrum is organised into several distinct bands, each of which has different anaesthetic applications. Table 2.2 summarises the bands frequently used in EEG analysis [80]. Of particular note are alpha oscillations, which are dominant at moderate levels of anaesthesia. As anaesthetic depth increases, delta waves become more dominant and alpha wave power decreases, reflecting a generalised slowing of brain activity [81, 82]. Higher frequency waves such as gamma waves tend to be associated with wakefulness [83].

Table 2.2: EEG frequency bands and their anaesthetic relevance.

EEG Band	Anaesthetic Applications
Slow wave (0.5–1 Hz)	Associated with deep anaesthesia [83]
Delta (0.5–4 Hz)	Associated with deep anaesthesia [81, 82]
Theta (4–8 Hz)	Associated with deep anaesthesia [82]
Alpha (8–13 Hz)	Dominant at moderate anaesthetic concentrations Decreases at deep anaesthesia levels [81, 82]
Spindle / Sigma (11–16 Hz)	Associated with moderate anaesthesia Sensitive to noxious stimuli [82]
Beta (13–30 Hz)	Increases in sedation [81]
Gamma (30–40 Hz)	Associated with wakefulness Suppressed under deep anaesthesia [83]

Several useful features can be extracted from the EEG bands. Typically, the power spectral density (PSD) is calculated for each band by taking the normalised squared magnitude of the Fourier transform output. Given a signal f , the Fourier coefficients $X(f)$ can be calculated. Given N samples, the PSD can be calculated as follows:

$$P_x(f) = \frac{1}{N} |X(f)|^2$$

Changes in EEG power can be decomposed into a periodic component (reflecting changes across specific frequencies), and a $1/f$ -like aperiodic component (contributing to all frequencies). Relative band power, the power of the band being analysed as a fraction of the power of the total spectrum, can be useful for mitigating the effect of the aperiodic component [84]. Band ratios (such as alpha-beta ratio, or theta-beta ratio) are also frequently used, although the value of these for mitigating the effect of the aperiodic component has been questioned [85].

Analysis within frequency bands can also be useful, particularly with regard to the alpha band. Peak alpha frequency decreases with increasing anaesthetic,

sometimes transitioning into the theta range [86]. The generalised slowing of EEG with increasing depth of anaesthesia can also be well represented by spectral edge measures. Spectral edge frequency at the 95th percentile is an important spectral edge measure. It represents the frequency below which 95% of total EEG power would be found. Increasing depth of anaesthesia results in reduction in spectral edge measures [87].

2.5.4 Other EEG Features

A phenomenon known as burst suppression occurs at very deep levels of anaesthesia, which involves alternating periods of high frequency oscillations and periods of no activity. It has been observed in coma, hypothermia, and deep general anaesthesia. Multiple potential neurophysiological processes have been proposed as underlying mechanisms [88]. One potential mechanism involves hypnotic agents generating a state of cortical hyperexcitability. Any subliminal stimuli generated by the subcortical brain or sensory system reaches the hyperexcitable cortex and causes a high frequency burst [89]. In the context of anaesthesia, burst suppression often indicates that a patient is too deeply anaesthetised. Burst suppression has been linked to increased incidence of postoperative delirium in some studies [90]. The proportion of time that a patient spends in burst suppression is known as the burst suppression ratio (BSR). The BSR is a useful measure of depth of anaesthesia at deeper levels, and is an important part of the BIS algorithm [91].

Phase-Amplitude Coupling (PAC) is a further measure that examines the extent to which the phases of low frequencies modulate the amplitudes of high frequencies, calculated by taking some correlation measure between the phases and frequencies [92]. PAC reflects hierarchical organisation of cortical dynamics (i.e. a more structured state), which can increase during anaesthesia [93]. Entropy measures, by contrast, measure how disordered the cortical state is. There are various entropy measures such as sample entropy, permutation entropy, and spectral entropy. Increased entropy is typically associated with wakefulness [94].

2.5.5 EEG Microstates

Microstates are semi-stable configurations of electrical activity in the brain, typically lasting 40–120 ms before switching to another configuration. In conscious subjects, microstates have been hypothesised as being the basic building blocks of conscious processing. Microstates usually require multiple-channel EEG data to classify. Once found, the durations, occurrence, transition dynamics, and percentage coverage of the microstates are analysed [95, 96].

Under anaesthesia, microstate duration increases, and complexity and diversity of transition dynamics falls [97]. It has been found that it is possible to achieve high accuracy on distinguishing consciousness and unconsciousness,

solely using features derived from a microstate analysis of EEG data [98]. This has led to the suggestion that microstate features could be used to aid in monitoring depth of anaesthesia [99].

2.5.6 EEG Variability

Substantial variability in EEG exists between subjects, and even within subjects between sessions. It is important that this variability is taken into account in any analysis of EEG data that involves multiple subjects. Sources of variation can be biological: EEG is affected by differences in skull thickness and the scalp-to-cortex distance, and tissue conductivity between the cortex and electrode. The neural generators that create EEG signals vary from person to person, with some genetic component: twin and family studies have shown moderate heritability of some EEG measures such as peak alpha frequency [100]. EEG is also seen to vary with age, usually characterised as a global slowing as an individual ages. This manifests as a decrease in alpha power and peak alpha frequency, and increases in delta and theta power [101]. A rise in alpha power is seen in children, reaching peak power in early adulthood before declining with ageing [102]. It has also been found that increasing age is associated with decreasing duration of microstates [99]. Sex differences in EEG have also been observed, primarily manifesting as heightened low alpha activity in older males and heightened low beta activity in older females (particularly in those with higher body mass index) [103].

Technical factors can also influence the EEG signal. Different placement of electrodes (or a different choice of reference) can lead to different spatial sampling of neural generators and therefore differences in the signal. Environmental factors such as temperature, humidity, and ambient electrical noise can also have an effect on EEG signal [100].

2.5.7 Single-Channel EEG

Single-channel EEG confers a number of advantages compared to multiple-channel EEG. Equipment and recording costs are lower, analysis tends to be less computationally expensive, dataset sizes are smaller, and single-channel EEG is less invasive [78]. The low cost and ease of measurement makes single-channel EEG ideal for use in the operating theatre. Most commercial EEG monitors use single-channel EEG [7].

However, single-channel EEG does have significantly lower information content than multiple-channel EEG, as it provides only a single spatial sampling of electrical information. Single-channel EEG cannot capture network dynamics, and EEG changes cannot be localised [104]. This makes it difficult to identify microstates in the traditional sense [99]. BSS EEG cleaning techniques (such as ICA or CCA) cannot be utilised on raw single-channel EEG signal, potentially leading to lower signal-to-noise ratio compared to multiple-channel

EEG. Although the proportion of EEG studies that use single-channel EEG has been rising in recent years, there remains a relative lack of public datasets and literature analysing the use of single-channel EEG [78].

Due to these weaknesses, some adaptations need to be made in the signal processing pipeline used for single-channel EEG data. Different artefact cleaning mechanisms need to be utilised, as previously discussed. If BSS methods are to be used, they require some mechanism to transform the single-channel data into multiple channels. Otherwise, besides lacking network analyses and localisation, the analysis process for single-channel data is very similar to that of other EEG data. It has been shown that ML-enabled signal processing can help mitigate some of the weaknesses of single-channel EEG data, both at the point where the EEG is cleaned and the point where the EEG is utilised [78].

2.6 Deep Learning in EEG Analysis

Deep learning (DL) has been employed in EEG analysis across a variety of fields. Inputs to DL models have included statistical measures, frequency domain features (such as band power), and spectrograms. Spectrograms, which represent a time-frequency domain representation of an EEG signal, enable the use of image-based DL methods. Raw EEG signal has also been used as input to DL models, allowing feature extraction and classification/regression to be learned end-to-end.

Network types that have been commonly used in EEG analysis include CNNs: both one-dimensional CNNs analysing raw signal and two-dimensional CNNs analysing spectrograms. Deep belief networks (DBNs), recurrent neural networks (RNNs), stacked auto-encoders, and multi-layer perceptrons (MLPs) have also been used on EEG data. In a 2019 review, CNNs, RNNs, and DBNs often outperformed stacked auto-encoders and MLPs on EEG data [105].

Transformers have also been useful for EEG data. Time-series transformers with positional encoding have been used to process EEG data in the time domain. Much like two-dimensional CNNs, vision transformers have been used to process spectrogram representations of EEG data. Graph Attention Transformers have been useful for multichannel EEG data: connections between different electrodes are represented as a graph and the transformer is able to learn complex patterns that take into account this structure. A variety of hybrid models also have been attempted using EEG data. For instance, CNNs are often used as feature extractors, and transformers are used for temporal sequence modelling and classification [106].

A more recent development in EEG deep learning has been the creation of several EEG foundation models. These are typically transformer encoder models trained to predict a self-supervised output, usually reconstructing masked segments of EEG. They are then used as feature extractors from raw EEG, or fine-tuned for a specific task. EEG foundation models are at an early stage of development, with many models only having been trained on a few hundred

hours of EEG data. Most are designed for multichannel EEG data, many with spatial encoding to take into account the layout of electrodes on the scalp [107]. SingLEM is a recent single-channel EEG foundation model trained on a larger dataset (357,000 hours) and optimised for feature extraction from single-channel EEG [108].

2.7 EEG and Nociception

Considerably more research has examined EEG responses to pain in awake patients than EEG responses to nociception in anaesthetised patients. Although no universally reliable EEG biomarker of pain has been identified, there are some clearly identifiable trends. Painful stimuli have generally been associated with increases in delta, beta, and gamma power. Decreases in alpha power have also been observed, although this finding is inconsistent. There have been no clear changes in theta power [109]. The increase in gamma power may be confounded by facial contractions (grimaces) made in response to painful stimuli [110].

Due to the blockades in neural pathways that occur in general anaesthesia, the EEG responses to noxious stimuli in anaesthetised patients differ from those in awake patients (most notably in the gamma frequency range). Although findings are inconsistent across studies, there are three spectral changes that have been generally observed:

- An increase in beta power.
- An increase in delta power, considered to be paradoxical.
- A decrease in alpha power.

Noxious stimuli typically have an arousal effect on the hypnotic dimension of anaesthesia. We would therefore expect noxious stimuli to be associated with a generalised increase in frequencies, with an increase in high frequency power and a decrease in delta power. The increase in beta power that has been frequently associated with nociception has been speculated to be linked to this arousal mechanism. It possibly represents movement into a more dream-like or wakeful stage, potentially reflecting widespread thalamocortical depolarisation. The beta arousal pattern tends to be associated with relatively lower doses of hypnotic agents (0.5–1.0 MAC), and insufficient opioid analgesia. It is also associated with a patient moving intraoperatively in response to pain.

By contrast, the increase in delta power that has been observed in some studies is incongruent with widespread arousal on the hypnotic dimension. It therefore has been referred to as paradoxical delta arousal, and it suggests some component of nociceptive EEG response which is orthogonal to the EEG changes associated with hypnotic agents. This increase in delta power is inconsistent and does not appear to occur similarly across all patients and studies.

It appears to be more pronounced in more invasive surgeries, such as those involving laparotomies. It tends to occur at higher doses of anaesthetic agents with insufficient opioid analgesia. It is thought that this increase in delta power may be caused by a subcortical negative feedback loop involving the brain stem and thalamus [12].

Another reported EEG change that has been noted is the decrease in alpha power, centred around 10 Hz [111]. This decrease is associated with high doses of anaesthetic agents and insufficient opioid analgesia. It has been hypothesised that alpha waves derive from thalamic pacemaker neurons, and a decrease in alpha power represents an increase in the complexity of communication between thalamus and cortex. However, all of these EEG changes have been fairly inconsistent between studies [12].

Nociception has also been associated with a decrease in high frequency variation, potentially related to a decrease in burst suppression-like patterns. This same study noted that the reduction in high frequency variation, as well as the decrease in 10 Hz power, were greater when fentanyl levels were lower [111]. Nociception has also been linked to changes in cross-frequency modulation. Modulation of an alpha carrier wave by theta and high delta increased with nociception, while modulation of the carrier wave by low delta decreased with nociception [112].

Most of the described patterns are transient and last only seconds. However, one study noted that in an animal model, a biphasic pattern that lasted up to 90 seconds was observed in response to a skin pinch. This featured an initial increase in alpha and delta, followed by a decrease in these frequencies and an increase in frequencies above 10 Hz [113].

It is unclear what effect, if any, intraoperative nociception would have on microstate transition dynamics. Although there has been extensive literature analysing the effect of hypnotic depth of anaesthesia on microstates, there has been little to no literature on microstates analysing the impact of the nociceptive dimension [97, 98, 99]. It has been noted that conscious pain perception is associated with a shift in the distribution of observed microstates [114].

2.8 EEG Drug Effects

2.8.1 Propofol

Propofol is a hypnotic agent, therefore its primary EEG effects are the typical changes observed with increasing depth of anaesthesia. This includes an increase in delta power (particularly slow wave power < 1 Hz), and an increase in alpha power, particularly in the frontal region. Propofol is also associated with a pattern of phase-amplitude coupling between delta phase and alpha amplitudes: at deep anaesthesia, high alpha amplitudes are accompanied by delta peaks, whereas at shallower anaesthesia, high alpha amplitudes are ac-

accompanied by delta nadirs [83].

2.8.2 Volatile Agents

Volatile agents broadly generate similar EEG changes to propofol. Again, volatile anaesthetics are associated with an increase in alpha power (particularly frontally), and slow delta power. Sevoflurane has been linked to a small increase in theta power at around 5 Hz [115].

2.8.3 Fentanyl

Although the primary aim of fentanyl administration during surgery is analgesic rather than hypnotic, it has long been known that high levels of fentanyl are associated with generalised EEG slowing. This includes the same slow delta oscillations that are observed with propofol and volatile agents, consistent with fentanyl having some hypnotic effect [116]. A rise in frontal theta power (4–8 Hz) has been proposed as a more specific biomarker for fentanyl, suggesting that fentanyl may have a direct effect on EEG separate from its hypnotic or antinociceptive action [117].

2.9 Non-EEG Nociception Indices

There have been many attempts at generating a measurable index of intra-operative nociception. Autonomic correlates of nociception such as increased blood pressure or heart rate have often been used as surrogate markers for nociception, but these signs are non-specific. Changes to these correlates can also occur in postural changes, bleeding, body temperature changes, and if cardiovascular agents are given [37]. Attempts at generating an objective index of nociception have used a wide variety of data, including indices requiring multiple modalities [118].

Most non-EEG-based nociception indices attempt to make a measurement of sympathetic or parasympathetic tone, which is then used as a proxy for surgical stress, which is partly driven by nociception [118]. The Analgesia Nociception Index (ANI), which ranges between 0 (extreme nociception) and 100 (no nociception), uses heart rate variability (HRV) to estimate parasympathetic tone. Nociception is associated with decreased parasympathetic tone and increased sympathetic tone, leading to a reduction in high frequency HRV and an increase in low frequency HRV. ANI has been shown to be moderately predictive of postoperative pain and may assist in appropriate intraoperative opioid titration [119].

Skin conductance reflects filling of palmar and plantar sweat glands, which gives an indication of sympathetic status. This has been investigated as a marker of nociception, and it has been found to be highly sensitive. However, it is non-specific, as fever, agitation, and changes in sedation level can also

result in changes in skin conductance. Any factor that causes dryness of the palms, such as anticholinergic drugs or peripheral neuropathy, will reduce the sensitivity of skin conductance for detecting nociception [118].

Intraoperative pupillometry has been used as a marker of nociception, utilising the pupillary dilation reflex that occurs in response to noxious stimuli [120]. Pupillometers measure the pupil size and allow for estimation of nociception using the Pupillary Pain Index, an index from 1 to 9. Although it has shown some promise for titrating analgesia, it is nevertheless affected by multiple confounding factors. If the patient is not shielded from ambient light, constriction can occur. Drug effects can alter pupillary status, for instance opioids can cause pupillary constriction. It also requires access and ability to open a patient's eye, making it an episodic measurement rather than continuous. It may also be an inconvenient index to measure repeatedly intraoperatively [69, 118].

The nociception flexion reflex (NFR) has been utilised for estimation of the analgesic state of a patient. EMG activity of the biceps femoris muscle is monitored while applying electrical stimulation to the sural nerve. The threshold at which the NFR is elicited is used as an index of analgesia [69]. This has been shown to have fairly poor predictive ability for movement responses [121].

The Surgical Pleth Index (SPI) (GE Healthcare) is an index that uses a proprietary algorithm on the pulse oximetry waveform to calculate a numeric index between 0 and 100 reflecting sympathetic response to nociception. It is calculated by taking a weighted combination of the heartbeat interval (HBI) and the pulse photoplethysmography amplitude (PPGA), and subtracting this from 100. Increases in sympathetic tone, which occur in response to noxious stimuli, will cause the HBI and PPGA to fall, and the SPI therefore rises [118]. Evidence has been equivocal regarding the ability of the SPI to improve opioid titration intraoperatively. Use of the SPI has been associated with reduced intraoperative propofol dosage and tachycardia, but it was not associated with improved postoperative pain scores or reduced postoperative opioid consumption [24].

The Nociception Level Index (NOL) (Medasense) uses multiple sources to generate a numeric index between 0 and 100. It includes data from four sensors: pulse oximetry waveform, skin conductance, peripheral temperature, and accelerometry [118]. These are used as the inputs to a proprietary ML algorithm. It has shown promise in randomised controlled trial (RCT) evidence, where the experiment group using NOL monitoring had lower postoperative pain scores with comparable levels of intraoperative opioid administration [122].

The Cardiovascular Depth of Analgesia (CARDEAN) index utilises cardiovascular parameters to quantify nociception. Elevated blood pressure and heart rate are non-specific signs of autonomic response to nociception. However, an elevation in blood pressure associated with a transient rise in heart rate (which then decreases due to the baroreceptor reflex) is a more specific

marker of nociception. The CARDEAN index has been used to predict movement in anaesthetised patients, although the evidence supporting its use for nociception estimation is otherwise limited [123].

Most of these methods of measuring nociception rely heavily on sympathetic or parasympathetic tone varying predictably in response to nociceptive stimuli. However, these methods are highly confounded, as sympathetic and parasympathetic tone change in response to many factors beyond noxious stimulation and nociception [118].

2.10 Literature Review: EEG Nociception Indices

A systematic literature review was performed in April 2026 to identify any existing nociception estimation algorithms that use EEG data, with a specific focus on monitors utilising ML. PubMed, IEEE Xplore, and Scopus were searched using the following query:

```
( anaesthesia OR anesthesia OR intraoperative ) AND ( nociception OR pain ) AND ( "machine learning" OR "deep learning" OR "artificial intelligence" OR prediction ) AND ( EEG OR electroencephalography )
```

Of 249 articles identified from the databases, 21 articles were selected for inclusion, covering 12 different monitoring approaches. Figure 2.5 outlines the systematic review process.

Three broad themes were observed in the included studies:

- Approaches utilising existing depth-of-anaesthesia indices. This is either by directly using the index as an estimation of nociception, or using a transform of one or multiple indices (which may involve ML to weight different components).
- Purpose-built nociception indices that do not utilise ML.
- Purpose-built nociception indices that do utilise ML.

2.10.1 Depth-of-anaesthesia Indices and Extensions

BIS

Although the BIS primarily measures the hypnotic component of anaesthesia, multiple studies have analysed the BIS as a monitor of nociception [124, 125, 126]. The first of these studies compared the BIS against the RIII reflex (a component of their NFR) in terms of their ability to predict movement responses to noxious stimuli while anaesthetised. 12 healthy male volunteers were placed under general anaesthesia, and noxious stimuli were applied. A prediction probability P_K was calculated for each individual, quantifying the

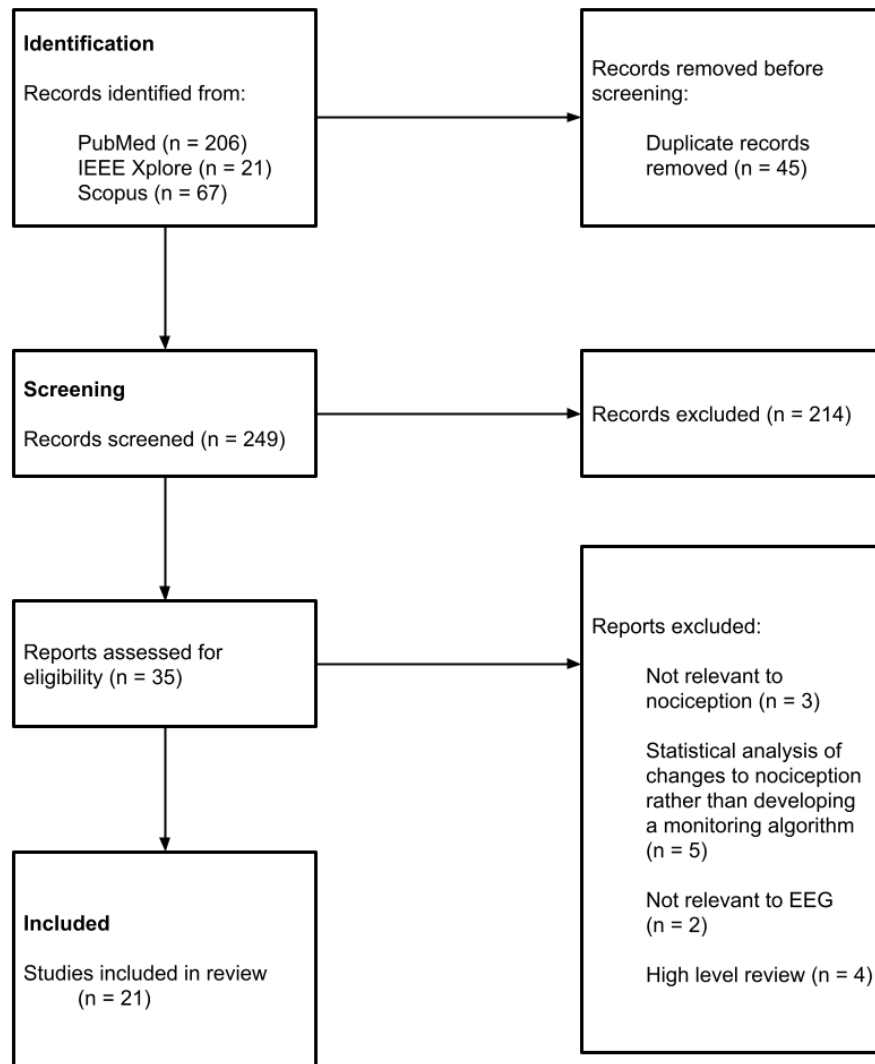


Figure 2.5: PRISMA-style flow diagram summarising the systematic literature review process used to identify EEG-based nociception monitoring studies.

discrimination ability of BIS or RIII to predict a movement response to noxious stimuli. The mean P_K was 0.955, indicating a strong ability of the BIS to predict whether someone would respond to a noxious stimulus in this small cohort [124].

Another study examined the effect of a standardised noxious stimulus (transcutaneous electrical nerve stimulation) on BIS, with a particular focus on the effect of age. The study population was 52 younger patients (aged 25–58), and 52 older patients (aged 65–78). For both younger and older patients, a significant increase in BIS was observed when the noxious stimulus was applied. For younger patients, the average increase was from 48.2 to 68.0 ($p < 0.0001$), and for older patients, the average increase was from 59.2 to 73.1 ($p < 0.0001$). Additionally, this study showed that when C_e of remifentanyl was increased, the BIS values fell, indicating responsiveness of BIS to opioids [125].

CVI

The sBIS, a metric quantifying the variability of the BIS, has been proposed as a measure for intraoperative nociception. It is calculated as the standard deviation of the BIS over the previous three minutes. It forms part of the proprietary Composite Variability Index (CVI) (Covidien), which combines BIS, sBIS, and sEMG (the standard deviation of EMG over the previous three minutes) in order to generate a 0–100 index intended to reflect nociceptive state. The training set for the CVI consisted of 153 cases with 59 movement events. The weights of the CVI a_1 , a_2 , a_3 , and a_4 were trained to provide the best discrimination between segments with movement events and segments without movement events. Given these weights, the BIS, sBIS, and sEMG, the CVI is calculated as follows:

$$\text{CVI} = 100 \left(\frac{1}{1 + e^{-(a_1 \text{sBIS} + a_2 \text{sEMG} + a_3 \text{BIS} + a_4)}} \right)$$

In one study, 105 surgical cases were divided into 862 ten-minute segments. 33 of these contained movement and 829 did not. The area under the receiver operator curve for discriminating movement was 0.83 for sBIS, 0.92 for sEMG, and 0.89 for CVI [127]. In another study, 120 participants were anaesthetised and subjected to noxious stimuli. CVI had superior ability to discriminate between the presence and absence of noxious stimuli compared to BIS, heart rate, or blood pressure. CVI did not significantly fall when remifentanyl increased, suggesting that the observed response to noxious stimuli may have been due to confounding with hypnotic depth [128].

GE Entropy

The Entropy Module (GE Healthcare) is primarily a depth-of-anaesthesia monitor measuring hypnotic state, and it outputs two indices: State Entropy (SE) and Response Entropy (RE). SE is calculated using EEG alone, and RE is

calculated using both EEG and EMG. The difference between SE and RE has been considered to be a biomarker for nociception as it isolates the facial EMG component of RE. One study involved 40 participants who were anaesthetised, and exposed to varying levels of electrical stimuli. It was found that as the size of the electrical stimulus increased, the difference between SE and RE also increased ($p < 0.0001$). The difference between SE and RE did not increase with increasing sevoflurane concentration, suggesting that this may be an appropriate marker of nociception unconfounded by hypnotic depth [129].

Another study analysed SE and RE in a different set of 40 patients, analysing whether SE or RE alone would serve as a meaningful marker of nociception. Spearman's ρ was calculated between each measure and the C_e of propofol and remifentanil. When correlating with C_e of propofol, the ρ values were -0.455 and -0.429 for maximum SE and RE respectively. However, when correlating with remifentanil, the ρ values were -0.139 and -0.123. This suggests that these indices are better as measurements of hypnotic depth and are poor correlates for nociception [130]. A similar study with fourteen laparotomy patients found that although SE and RE could evaluate analgesic depth, they did not vary significantly during nociceptive stimulation. It was also noted that due to its reliance on facial EMG, the usefulness of RE is limited in paralysed patients [131].

RN

A response index of nociception (RN) was developed, which includes heart rate variability, RE, the difference between RE and SE, and pulse oximetry waveform variability. The dataset for developing this index consisted of 55 female patients undergoing general anaesthesia. In order to get a ground truth measurement of nociception, a Clinical Signs-Stimulus-Antinociception (CSSA) score was constructed. This score consists of three parts: the clinical signs of nociception (i.e. movements) n_{signs} , the level of noxious stimulus n_{stimulus} (graded by an anaesthetist's expectations of how significant the stimulus would be), and the level of analgesic drug $n_{\text{remifentanil}}$. Each of these is scored between 0 and 4, and combined as follows:

$$\text{CSSA} = n_{\text{signs}} + n_{\text{stimulus}} - n_{\text{remifentanil}} + 4$$

Linear and sigmoid regression were used to fit factors to each term of the CSSA equation. Remifentanil effect was estimated with a linear regression on the change in RE, and the beat-to-beat RR interval (RRI) on an ECG. The estimate of the clinical signs of nociception was determined with a sigmoid regression on RRI, and the change in the difference between SE and RE. The estimate of stimulus was determined with a sigmoid regression using features from the pulse oximetry waveform. RN was finally calculated by combining the three factors in a final linear regression [132].

A validation study was carried out analysing the usefulness of RN to detect

nociception in a group of 54 female patients. It was found that increases in RN were associated with noxious stimuli, and higher RN correlated with increased chance of movement ($p < 0.05$) [133].

2.10.2 Novel Non-ML Indices

Time-Frequency Representation

A group of 378 patients undergoing endoscopy under general anaesthesia was used to develop several candidate indices predicting responsiveness to nociceptive stimuli. The endpoint of these approaches was a patient’s score on the Ramsay Sedation Scale (RSS) after nailbed compression, and the presence or absence of a gag reflex during endoscopy. Table 2.3 summarises the RSS.

Table 2.3: Ramsay Sedation Scale (RSS).

Score	Clinical Response
1	Anxious and/or restless
2	Cooperative, oriented, and tranquil
3	Responds to commands
4	Brisk response to light glabellar tap or loud auditory stimulus
5	Sluggish response to light glabellar tap or loud auditory stimulus
6	No response to light glabellar tap or loud auditory stimulus

One minute windows were extracted and processed by a time-frequency representation function. Various features from this representation, such as the mean, median, and maximum values of powers and frequencies, along with Shannon and Rényi entropies, were used as inputs into multivariate models. The best discriminator for the gag reflex had a P_K of 0.816, and the best discriminator for RSS discrimination had a P_K of 0.830 [134].

AMIF

An approach utilising the auto-mutual information function (AMIF) was developed, using the same patient database and endpoints as above. Several measurements, including spectral power and weighted mean frequency in each band (including very high frequency bands), and spectral edge frequencies were calculated. Then, AMIF was calculated as the mutual information between an EEG signal and time-delayed versions of itself, and a variety of AMIF metrics were calculated. While BIS had a P_K of 0.620 for distinguishing between RSS scores of 5 and 6, and a P_K of 0.799 to distinguish between an RSS below 5 and 6, a combined multivariate index using AMIF metrics had a P_K of 0.763 for 5 and 6, and a P_K of 0.810 for less than 5 and 6 [135]. This approach was then extended to include Refined Multiscale Entropy measures, which yielded similar results [136]. It is noted that these measures achieved the best results when

using very high frequencies, possibly capturing high frequency facial EMG as a correlate of nociception [135].

BARM

Composite cortical state (CCS) and cortical input (CI) are proprietary measures of hypnosis and nociception, respectively. They are the two key indices that are used in the Brain Anaesthesia Response Monitor (BARM), and unlike other indices, they are derived from physiological modelling of brain electrical activity rather than heuristic modelling using EEG data. The algorithm is proprietary and therefore unknown, but it appears to involve modelling the EEG as an autoregressive moving average (ARMA) process.

In one validation study, 25 patients undergoing coronary artery bypass graft surgery had CCS and CI measured during surgery, and these indices were correlated against propofol levels, fentanyl levels, and BIS. Although CCS, CI, and BIS all fell significantly with fentanyl boluses, it was found that CI could differentiate between different levels of fentanyl. CI also had poor ability to predict hypnotic state compared to CCS or BIS, suggesting that it measures some element of the EEG that is orthogonal to hypnosis [137].

A further evaluation study used data from 80 patients, and observed the effects of remifentanyl on CI and CVI. Both CI and CVI were able to differentiate patients that received remifentanyl from those that did not. These indices were also used to predict movement, with each nociception index being combined with a hypnosis index: CI was combined with CCS, and CVI was combined with BIS. It was found that CI + CCS outperformed CVI + BIS, with a higher sensitivity (75.8% vs 42%, $p = 0.006$) and specificity (52% vs 24%, $p = 0.016$) [126].

2.10.3 Novel ML Indices

qNOX

The qCON monitor (Quantum Medical) provides two indices: qCON and qNOX. Whilst qCON is an index of hypnotic depth, qNOX is an index of nociception. qNOX was developed using data from 450 patients undergoing endoscopy under anaesthesia with propofol and remifentanyl, 80 patients under anaesthesia with sevoflurane, and 50 awake volunteers, for a total dataset of 580 patients. It was developed in concert with qCON.

Although the exact details are proprietary, some high-level aspects of the training process of qNOX are known. The ground truth label was a binary outcome derived from the RSS. Nailbed pressure was applied to each patient, and if they moved their hand, they were considered to be at five or below on the RSS, whereas if they did not, they were considered to be at 6.

From the EEG, the powers from four frequency bands were calculated: 1–4 Hz, 8–13 Hz, 13–44 Hz, and 30–44 Hz. These were transformed into frequency

ratios F_n , where E_n is the power for a given frequency band n , and E_{tot} is the total power across 1–44 Hz:

$$F_n = 20 \times \log \frac{E_n}{E_{tot}}$$

The four frequency ratios were used as the input to an Adaptive Neuro Fuzzy Inference System (ANFIS). This is a hybrid of a five-layer neural network, and a fuzzy logic system. Standard backpropagation was used to train the weights of this network. There was some compensation to account for the effect that hypnotic anaesthesia has on the ability of a subject to move in response to noxious stimulus: if qCON was below 25, compensation was applied to the qNOX to account for the lack of movement that may also occur under deep anaesthesia. The output is then scaled to give an index between 0 and 100, with 0 being minimal probability of movement and 100 being maximal probability of movement.

A validation study was performed using a dataset of 60 patients undergoing propofol and remifentanyl anaesthesia. qNOX was measured before and after laryngoscopy, laryngeal mask airway (LMA) insertion, and tracheal intubation. It was also recorded whether a patient moved in response to these stimuli. It was found that, for all groups, the noxious stimulus led to a statistically significant increase in qNOX ($p < 0.05$ for all). It was also found that patients who moved in response to stimuli had a higher qNOX than patients that did not [138]. Another validation study used a larger dataset of 140 patients, also undergoing propofol and remifentanyl anaesthesia. Again, individuals with higher qNOX were more likely to respond to LMA insertion [139]. It should be noted that both of these studies were conducted by researchers associated with Quantum Medical.

Independent validation studies have been more mixed on the clinical value of qNOX. A further study involved a dataset of 150 patients undergoing surgery with sevoflurane anaesthesia. In this study, qNOX was measured every five minutes at the end of surgery. Patients were then asked to rate their pain on a scale from 0 to 10 once they were awake, with 0 being no pain and 10 being severe pain. qNOX showed almost no ability to discriminate between those with no or mild pain (0–3) from those in moderate-to-severe pain (4–10): the area under the receiver operating characteristic curve (AUROC) was 0.501 (95% CI: 0.406–0.597). This was worse than using heart rate as a discriminator (AUROC: 0.605; 95% CI: 0.511–0.700) or blood pressure (AUROC: 0.591; 95% CI: 0.497–0.685) [16].

Another independent study analysed the ability of qNOX to detect noxious stimuli. 83 patients had qNOX, heart rate, blood pressure, and BIS monitored during surgery. Each of these indices were recorded before and after a variety of stimuli: skin incision, tracheal intubation, and electrical tetanic stimuli. Significant increases ($p < 0.001$) in qNOX were observed after intubation and skin incision, however no significant differences were observed with tetanic

stimuli. qNOX had an AUROC of 0.65 for discrimination of intubation from a non-noxious stimulus, and an AUROC of 0.67 for discrimination of skin incision from a non-noxious stimulus. However, the ability to discriminate tetanic stimuli was lower: 0.54 and 0.52 for two different levels of tetanic stimuli and a non-noxious stimulus. This was worse than blood pressure or heart rate, indicating the value of qNOX to detect nociception is low [140].

Both independent studies questioned the usefulness of qNOX as a nociception index. It was noted that correlation between qNOX and qCON was high, indicating that qNOX potentially serves as a secondary marker of hypnotic depth rather than a marker of nociception [16, 140].

Pain Threshold Index

The pain threshold index (PTI) (Beijing Easymonitor Technology) is a ML-based algorithm for estimation of nociception. It is the counterpart to the wavelet index (WLI), which estimates depth of sedation. It originates from a broader proprietary EEG analytics platform designed for estimating 24 different indices [17].

The construction of the PTI can only be understood at a high level due to its proprietary nature. As an input, the PTI takes two channels of EEG data. This data is transformed with three transforms: a continuous wavelet transform, a discrete dyadic wavelet transform, and a FFT. Candidate features are extracted from each of these transforms, including the power and phase in the alpha band, delta band, theta band, and beta band, as well as a spectral edge frequency and central frequency. These features are combined in a multi-variable regression, although the details of the training target, the exact input variables, and validation process are not made clear [141].

One validation study of the PTI featured 134 patients undergoing general or urological surgery under general anaesthesia. Heart rate, blood pressure, PTI, and WLI were recorded before and after intubation and skin incision events. Whether there were significant changes to the heart rate and blood pressure was used as a target outcome, and the ability of the PTI to predict haemodynamic changes was evaluated. It was found that the PTI had fairly strong ability to discriminate episodes associated with haemodynamic response, with an AUROC of 0.81 for intubation and 0.82 for skin incision [141].

Another validation study with 76 patients undergoing urological surgery under general anaesthesia evaluated the ability of PTI to predict postoperative pain. The mean PTI values from the last ten minutes of surgery were taken for each patient, and used as a predictor for postoperative pain scores. Postoperative pain, initially measured on a scale from 0 to 10, was divided into no or mild pain (0–3) and moderate-to-severe pain (4–10). For predicting this binary outcome, the PTI had an AUROC of 0.772. This was superior to SPI, heart rate, and blood pressure for predicting postoperative pain [142]. A similar validation study was conducted with 80 children undergoing laparoscopic surgery under general anaesthesia, again aiming to predict moderate-to-severe

pain scoring 4–10 on a 0–10 scale. The AUROC for the PTI was 0.796, again suggesting moderate-to-strong ability to detect patients at risk of moderate-to-severe postoperative pain [25].

An RCT was conducted to observe the effect of using PTI-guided analgesia during anaesthesia. 122 females undergoing gynaecological surgery were randomised to either a PTI-guided analgesia group, or standard care. The outcome was whether or not the patient required analgesia postoperatively. 40% of patients in the control group required analgesia, compared to 14% in the PTI group ($p = 0.002$) [143].

Altés *et al.* 2021

The aim of this study was to predict whether a patient would move in response to standardised tetanic stimuli, using EEG. The dataset used consisted of 77 female patients undergoing gynaecological surgery under general anaesthesia with different propofol and remifentanyl levels. At various points during anaesthesia, the patients would be exposed to a standardised tetanic stimulus, and whether or not they moved in response to the stimulus was recorded. Single-channel EEG was gathered from each of them, with 45-second segments prior to tetanic stimulation being used as input for all predictive models.

The 45-second segments were split into five-second epochs, and a variety of features were extracted from each epoch. This included a variety of entropy measures, the Hjorth parameters, the power spectral density, and mean powers in each band (delta, theta, alpha, and beta). A two-dimensional scalogram was also generated, using a continuous wavelet transform to create a time-frequency domain representation.

Using these extracted features and transforms, five different models were tested, with the endpoint being whether or not a patient would move in response to tetanic stimulation:

- A one-dimensional CNN architecture (EEGNet) using the raw signal, based on Inception.
- Using the tabular data with extracted features, a random forest classifier, support vector classifier, and XGBoost classifier were trained.
- A two-dimensional CNN architecture (Inception V3) using the time-frequency scalogram.

Two other pieces of information were also gathered: the pupil diameter, and the calculated effect-site concentration of remifentanyl. Ensembles including various combination of the trained models, the pupil diameter, and the remifentanyl concentration were tested. Five-fold cross-validation was performed to validate each model or combination of models.

The best model for predicting movement was an ensemble using pupil diameter, the effect-site concentration of remifentanyl, a random forest classifier

on extracted features, and a one-dimensional CNN. This had an AUROC of 0.84 for predicting movement in response to tetanic stimulus [13].

Abdel Deen *et al.* 2025

This study trained multimodal models for prediction of nociception. The dataset consisted of 100 patients undergoing surgery under general anaesthesia. During each surgery, metrics such as heart rate, blood pressure, volatile anaesthetic concentration, and opioid dosages were measured, along with events such as intubation. Five anaesthetists were given these series, and each made an independent assessment of how they anticipated nociception would vary over time. These were aggregated into a single nociception assessment score (NOA) for each patient over time, varying between 0 and 100. The NOA served as the target label for ML.

The input data for this study were multimodal. For each patient, EEG, ECG, and photoplethysmography (PPG) data were combined and used as the input. Each patient recording was divided into 64-second segments. For EEG data, the mean power was calculated across delta, theta, alpha, beta, and gamma bands. This was combined with features extracted from the ECG and PPG data, normalised, and used as the input for the models.

This study trained two models: a two layer MLP and a long short-term memory network (LSTM). The MLP had an R^2 value of 0.80 when comparing the validation predictions against the NOA reference labels, whereas the LSTM had an R^2 value of 0.81. It was found that the NOA provided accurate responses to rapid changes to painful stimuli, whereas the LSTM was less sensitive to these changes but provided smoother output trajectories [14].

Abdel Deen *et al.* 2026

This study used the same NOA dataset as Abdel Deen *et al.* (2025), with the same multimodal signals (EEG, ECG, PPG) and aggregated anaesthetist NOA target. However, in this model, a quantum multilayer perceptron (QMLP) was used. The simulated entanglement mechanism in a QMLP can be used to represent connections between parameters within layers. The QMLP was benchmarked against a transformer, a regular MLP, and several classical ML models. The R^2 value for the QMLP was 0.673, which exceeded the performance of alternative models tested on the same validation sample [15].

2.10.4 Discussion

Table 2.4 summarises the purpose-built intraoperative nociception indices created with a ML approach. The only monitor with RCT evidence of improvement in patient outcomes is the PTI, however the least is known about how it was trained. qNOX performs poorly in independent trials. The remain-

ing models remain at an early stage with no independent studies or external evidence supporting their use.

Establishing a reliable, objective, and unconfounded measure of ground truth (whether it is intended to be used as a ML training target or a validation target) is a significant challenge in the development of any index of nociception.

Many studies use movement in response to noxious stimulus (such as nailed pressure in RSS) as an objective target, either for training ML models for nociception detection [13, 138], or for validating other methods of nociception detection [124, 125, 133, 134, 135]. However, this is a confounded target. Both the hypnotic and paralytic components of general anaesthesia also contribute to whether or not a patient moves in response to painful stimuli [128]. qNOX does compensate for this [138], but the high levels of correlation between qNOX and qCON (a related index of hypnotic depth) indicate that the compensation is insufficient for qNOX to be considered a pure measure of nociception [16, 140].

A more simple approach, typically used for validation studies, involves determining whether or not an index can distinguish between the presence and absence of noxious stimuli. In many cases, this is an electrical tetanic stimulus applied to the anaesthetised patient [125, 128, 140]. Tetanic stimuli come with the advantages of being standardised. It is also possible to vary the size of the stimulus, which can be used as a further method of validation: one would expect a dose-response relationship between the size of the stimulus and the nociception index [129]. However, electrical tetanic stimuli are not reflective of the actual stimuli that occur during surgery. Although these stimuli are more variable and unpredictable, some studies have opted to observe the response of a candidate nociception index to standard surgical noxious stimuli, such as skin incision and intubation [138, 140].

The response of an index to opioids has been used as a method for validating candidate nociception indices. In the presence of a constant nociceptive stimulus (such as an electrical tetanic stimulus), one would anticipate that varying levels of opioid would lead to varying levels of nociceptive blockade. Appropriate response to different levels of opioid is an indication that an index may be capturing nociception-associated EEG changes [125, 130, 137]. However, it is important to note that opioids also have some hypnotic effect which could plausibly drive changes in an index of hypnotic effect [144].

Haemodynamic reactivity (i.e. changes in blood pressure and heart rate) has also been used as a ground truth against which a candidate nociception index has been validated [141]. However, as previously mentioned, haemodynamic parameters are non-specific and can vary in response to many factors that are not nociception [37]. Blood pressure and heart rate are the current standard clinical monitoring tools that an anaesthetist would use to monitor nociception, and the objective of developing an EEG nociception index is to outperform them. These parameters are therefore often used as a comparison against which nociception indices are evaluated [16, 128, 140].

Table 2.4: Summary of ML EEG approaches for nociception detection, including dataset, training target, model, and validation status. Internal validation denotes models that were trained with an internal train-test split.

Name	Year	Training dataset	Target	Model	Proprietary	Validation
qNOX	2014	530 anaesthetised patients, 50 awake volunteers (EEG only)	Movement in response to nailbed pressure	Adaptive Neuro-Fuzzy Inference System (ANFIS)	Yes (Quantium Medical, Spain)	Successful internal validation [138, 139]. Independent studies reported limited ability to predict postoperative pain [16] and inconsistent detection of noxious stimuli [140].
Pain Threshold Index	2017	Unknown (EEG only)	Unknown	Multivariable regression	Yes (Beijing Easymonitor Technology, China)	Strong ability to detect haemodynamic changes [141], postoperative pain [142, 25]. Validated prospectively in a randomised trial [143].
Altés <i>et al.</i>	2021	77 anaesthetised patients (EEG and pupil diameter)	Movement in response to standardised stimuli	Ensemble of CNN, random forest classifier	No	Internally validated only [13]
Abdel Deen <i>et al.</i>	2025	100 anaesthetised patients (EEG, ECG, PPG)	Anaesthetist-graded NOA score	MLP / LSTM	No	Internally validated only [14]
Abdel Deen <i>et al.</i>	2026	100 anaesthetised patients (EEG, ECG, PPG)	Anaesthetist-graded NOA score	QMLP	No	Internally validated only [15]

Subjective anaesthetist assessment of nociception has been used as a ground truth target for ML. This has been performed in a structured way with the CSSA score, where intraoperative movement, level of noxious stimuli, and level of opioid were combined into a single score (with weightings and grading of movement and stimuli being subjectively determined) [132], or an unstructured way with the NOA score [14]. This is a very subjective target and is reliant on the intuition of the graders. One study attempted to improve objectivity by using a group of anaesthetists and aggregating their independent scores [14]. However, this would not account for any consistent biases in the scoring.

Prediction of postoperative pain is a more clinically meaningful target for validating these indices. Pain scoring is subjectively graded on a scale of 0–10 by the patient once they are awake from surgery. The pain score is often divided into a binary target, separating no/mild pain (0–3) from moderate-to-severe pain (4–10). The mean score of a candidate nociception index from some period towards the end of surgery is taken, and its ability to discriminate whether or not a patient would have moderate-to-severe pain is determined, using metrics such as AUROC. If the predictive ability is superior to some baseline index (such as haemodynamic parameters or BIS), the index could be considered useful [16, 142, 25].

The strongest evidence for evaluating a candidate nociception index is an RCT. This involves study participants (scheduled for surgery under general anaesthesia) being randomly assigned to either usual anaesthetic care, or anaesthetic care with opioid administration guided by an index. Patient-centred outcomes are used as the primary endpoint of these studies, such as amount of analgesia required postoperatively, or pain scores. RCTs generate evidence showing how the nociception index performs when in use by anaesthetists, and demonstrate whether use of the index creates tangible benefits for patients. The use of RCTs for evaluation of EEG nociception indices has been very limited [143].

Disentangling the antinociceptive and hypnotic dimensions is a non-trivial problem. As shown by this review, many indices that were developed to capture the hypnotic dimension contain elements of the antinociceptive dimension. This includes BIS [124, 125], RE, and SE [130]. Likewise, indices intended to be measures of nociceptive depth, such as CVI [128] and qNOX [16] often appear to contain information about the hypnotic dimension. Any putative index for nociception detection needs to either be trained to detect a nociception-specific target (which is not confounded with hypnotic or paralytic targets), or there needs to be sufficient adjustment for this confounding. As part of the validation process, it is also important for candidate nociception indices to be correlated against measures of hypnotic depth or concentrations of hypnotic agents, as this can demonstrate that an index operates independently of hypnosis.

Another significant confounding factor mentioned across multiple studies is facial EMG. Nociceptive stimuli can cause facial contractions such as grimaces,

which creates an EMG signal. This EMG signal can contaminate the EEG signal (often at high frequencies within the gamma band), and can be used as a predictor of nociception. Some indices deliberately incorporate the EMG signal, such as GE Entropy’s RE (with the difference between RE and SE being a pure measure of facial EMG) [129] and CVI [127]. Some indices may also unintentionally incorporate the EMG signal. It has been speculated that the performance of AMIF and qNOX may partially be due to these algorithms making use of EMG contamination to predict nociception [135, 140]. Despite its use in these algorithms, facial EMG is not a reliable indicator of nociception, as it is highly confounded by paralysis. In paralysed patients, facial EMG becomes highly unreliable [129, 131]. In order to generate a nociception index appropriate for general use, it will therefore be necessary to focus on EEG signals originating from the brain, possibly by using a low pass filter to remove high frequency EMG signal [145].

There is currently heavy commercial involvement in this field. Many of the key indices associated with nociception estimation are proprietary, such as CVI, RE/SE [127], BARM [137], qNOX [138], and the PTI [17]. It is desirable for any artificial intelligence system in healthcare to be as transparent and explainable as possible [146]. However, for many of these proprietary indices, the way that the index is calculated or trained is not known in full. The most well-validated index of nociception from EEG data, the PTI, has very little publicly available information regarding how it was developed [141]. This can lead to reduced trust amongst clinicians and therefore reduced adoption [146]. Commercial interests can also lead to conflicts of interest and a lack of neutrality in published validation studies, which occurred in the case of qNOX. Validation studies carried out with the involvement of researchers from Quantum Medical reported more favourable results for qNOX [138, 139] as compared to independent validation studies [16, 140].

Chapter 3

Dataset and Tools

3.1 Software and Libraries

All analysis was completed in Python 3.10.12, with some use of MATLAB 2025b for data extraction. Key libraries for analysis were NumPy 1.26.4, Pandas 2.3.0, and SciPy 1.15.3. PyTorch 2.9.1 and Scikit-learn 1.7.1 were used for model building and ML. Matplotlib 3.10.3 was used to generate figures.

3.2 Dataset

The dataset used in this study is a combination of three previously collected datasets from Waikato Hospital and North Shore Hospital. The three studies are BISpain [147], AlphaGA [148], and AlphaMax [149]. All datasets have single-channel EEG for all patients, a 0–10 postoperative pain score for all patients when they awaken from anaesthesia, and time and dosage data for drugs given during surgery. All datasets also contain age, sex, and weight for all patients. The timing of the start of surgery was measured for all patients, defined as the time of first surgical incision. The timing of surgical induction, and the timing of the end of the surgery were also measured.

3.2.1 BISpain

BISpain was a randomised controlled trial conducted with the aim of determining if targeting lower intraoperative BIS would lead to lower postoperative pain scores. 135 patients were recruited, all between 18 and 65 with American Society of Anesthesiologists (ASA) scores of I or II, and all were undergoing surgeries less than 180 minutes in duration. The surgeries occurred at Waikato Hospital and North Shore Hospital, and all patients had continuous BIS monitoring intraoperatively. Patients were randomised to either a low BIS target (30–40) or a high BIS target (45–60), and the anaesthetist would titrate drugs intraoperatively in order to achieve these targets. Propofol was used for induction, and desflurane was used for anaesthetic maintenance. Analgesia was

achieved via a fentanyl infusion. This study found no statistically significant difference in postoperative pain scores between the two groups [147].

3.2.2 AlphaGA

AlphaGA was an observational study which aimed to model the causes of reductions in alpha power during surgery. 305 patients were recruited, of which 237 were included in the final analysis after excluding excessively noisy EEG or patients with missing monitoring data. These patients had ASA scores between I and IV, and were aged between 18 and 90. All surgeries occurred at Waikato Hospital, with frontal EEG being measured using either the BIS or GE Entropy module. All patients received standard anaesthetic care. Induction was achieved with propofol, and sevoflurane or desflurane was used for maintenance. Patients received fentanyl and/or morphine for analgesia. A generalised linear model was fit to explain variation in alpha power, using effect-site concentrations of propofol, volatile anaesthetic, and opioid as inputs. The residuals from this model were then analysed. It was found that reductions in alpha power were associated with body cavity surgery and longer operations, leading to the hypothesis that a reduction in alpha power could provide a reliable measure of nociception [148].

3.2.3 AlphaMax

AlphaMax was a randomised controlled trial aiming to determine whether deliberate intraoperative maximisation of alpha power would lead to reductions in postoperative delirium. 200 patients were recruited, all aged over 60, and undergoing elective surgery greater than 90 minutes in duration. 100 patients were allocated to the control group (standard anaesthetic care), and 100 patients were allocated to the intervention group, where the anaesthetists aimed to titrate anaesthetic agents in order to maximise oscillatory alpha power. Patients were induced with propofol, and then desflurane and fentanyl were used for hypnotic maintenance and analgesia, respectively. All recordings were completed at Waikato Hospital, with single-channel EEG data collected from the Entropy module. There was ultimately no difference in delirium levels found between the two groups [149].

3.3 Ethical Approval

Ethical approval was obtained separately for each study, and all patients gave written consent for the use of their data. All data were anonymised after collection and prior to use. BIS Pain was approved by the Northern X Regional Ethics Committee (NTX/09/06/047), and registered on the Australian New Zealand Clinical Trials Registry (ANZCTR) (ACTRN12609000350224) [147]. AlphaGA was approved by the New Zealand Health and Disability Ethics

Committee (HDEC) (12/CEN/56) [148]. AlphaMax was approved by HDEC (17/NTA/56) and registered on ANZCTR (ACTRN12617001354370) [149].

3.4 Data Processing

There were three key processing steps involved in the construction of the dataset: the calculation of drug concentrations at each time point during surgery, the cleaning and processing of EEG data, and the calculation of the OpenIbis index.

3.4.1 PK Calculations

During each surgery across all three studies, the times and dosages of fentanyl, morphine, and propofol were recorded. A two-compartment model with an effect-site equilibration constant k_{e0} was used to model the effect-site concentration of each of these agents. Reference values from the literature were used for each of the PK parameters in the two-compartment model. Table 3.1 gives the values of these PK parameters used to construct the model [150, 151, 152, 153]. The model was then solved using numeric methods in order to determine the effect-site concentrations of each agent at each point during the surgery. Effect-site concentrations of each of these agents were included in the dataset with one measurement every second during surgery. Morphine was converted to fentanyl equivalents for the purposes of this study, where 2 mg of morphine was considered equivalent to 100 μg of fentanyl [148].

Table 3.1: PK parameter values used for effect-site concentration modelling.

Parameter	Fentanyl [151]	Morphine [153]	Propofol [152]
Model structure	Two-compartment	Two-compartment	Two-compartment
k_{e0} (s^{-1})	0.147/60	0.418/3600	1/150
Q (Ls^{-1})	2.3/60	101/3600	0.989/60
V_1 (L)	$13 \times \text{weight}/70$	$14.2 \times \text{weight}/70$	$24 \times \text{weight}/70$
V_2 (L)	$358 \times \text{weight}/70$	$258 \times \text{weight}/70$	$112 \times \text{weight}/70$
Clearance (Ls^{-1})	0.62/60	102/3600	2.64/60

The concentrations of volatile anaesthetics were calculated using end-tidal measurements and a k_{e0} of 0.0048 s^{-1} based on a reference k_{e0} half-time of 144 s [154]. The concentrations of volatile anaesthetics were converted to MAC values for consistency between desflurane and sevoflurane.

3.4.2 EEG Cleaning

Three variants of the complete EEG dataset were generated during the cleaning process:

- No cleaning.
- Standard cleaning.
- Cleaning using an EEG denoising transformer [155].

The AlphaGA and AlphaMax EEG data were recorded at a sampling rate of 100 Hz. The BISPain data, however, were recorded at a sampling rate of 128 Hz. The BISPain data were downsampled to 100 Hz using the polyphase method. The combined EEG data at this stage formed the ‘No cleaning’ variant of the dataset. Figure 3.1 shows the appearance of three example spectrograms at this stage.

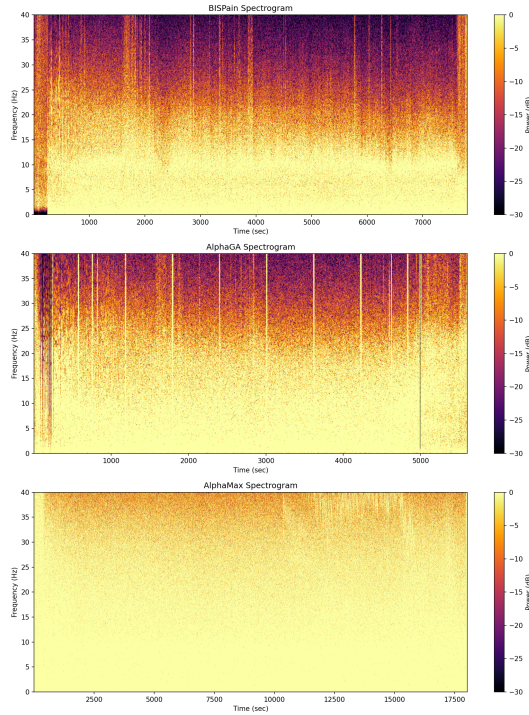


Figure 3.1: ‘No cleaning’ dataset variant.

All EEG data were then high-pass filtered with a Butterworth forward-backward filter with a cut-off at 0.5 Hz. This removed any low frequency drift and effectively de-trended the data.

A slope-based artefact detector was then applied. The first backward difference was calculated, and then smoothed with a three sample convolution. At any time points where the smoothed slope exceeded a predefined threshold (500 units for AlphaMax and 200 units for AlphaGA and BISPain), the time points were removed and replaced with a Not a Number (NaN) flag. Small gaps in each signal (up to five samples in duration and 200 units in amplitude) were then linearly interpolated. Larger gaps were retained.

The data were bandpassed with a Butterworth forward-backward filter, with frequencies from 0.5–40 Hz being retained. The EEG data were then

finally standardised using online exponential moving standardisation. Given EEG measurement x_t at time step t , a previous mean μ_{t-1} and variance σ_{t-1}^2 , and a constant α , the current mean and variance at time step t are calculated as follows:

$$\mu_t = (1 - \alpha)\mu_{t-1} + \alpha x_t$$

$$\sigma_t^2 = (1 - \alpha)\sigma_{t-1}^2 + \alpha(x_t - \mu_t)^2$$

The normalised EEG measurement z_t is then calculated as follows, applying a bias correction term c_t . $\alpha = 0.01$ and $\varepsilon = 10^{-8}$ were used for the calculations for this dataset.

$$c_t = 1 - (1 - \alpha)^t$$

$$\hat{\mu}_t = \frac{\mu_t}{c_t}, \quad \hat{\sigma}_t^2 = \frac{\sigma_t^2}{c_t}$$

$$z_t = \frac{x_t - \hat{\mu}_t}{\sqrt{\hat{\sigma}_t^2 + \varepsilon}}$$

The output of this normalisation step constitutes the ‘standard cleaning’ variant of the dataset. Figure 3.2 demonstrates the appearance of the same three example spectrograms at this stage.

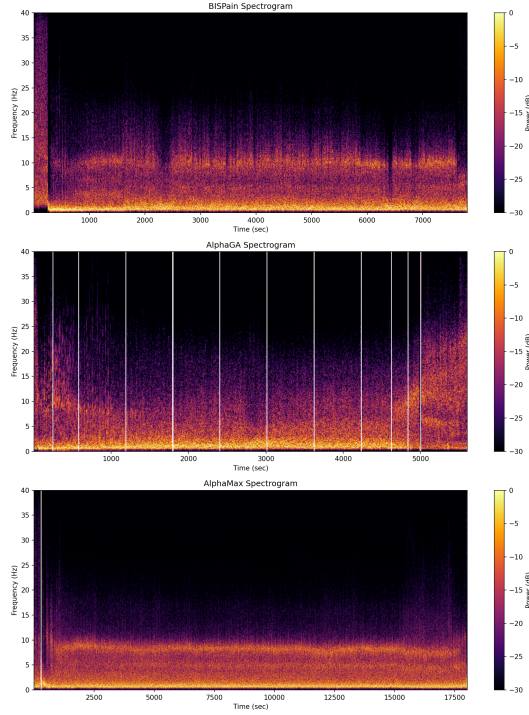


Figure 3.2: ‘Standard cleaning’ dataset variant.

Finally, the data were used as an input to a denoising transformer pre-trained on the EEGdenoiseNet dataset. The denoising transformer is an attention-based model that uses local and non-local self-similarity to detect noise and outliers. The denoising transformer was used with the aim of removing any residual noise in each EEG signal, including any EMG contamination [155]. The output of this was the third dataset variant, with the results shown in Figure 3.3.

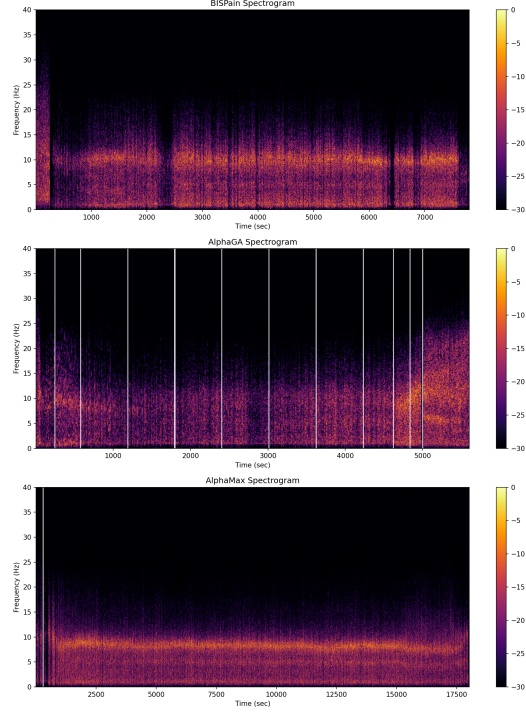


Figure 3.3: ‘EEGdenoiseNet cleaning’ dataset variant.

3.4.3 OpenIbis

The BIS is the most frequently used index of hypnotic depth. However, the algorithm for the BIS is proprietary. This has led to efforts to reconstruct the BIS algorithm. One such effort, known as the Ibis algorithm, resulted from the forensic disassembly of an A-2000 BIS monitor into machine code, and the creation of a BIS emulator. 36 hours of intraoperative EEG data were used as input to the emulator, and the final and intermediate calculations produced by the emulator were analysed. This analysis yielded the Ibis algorithm, which almost perfectly recreates the original BIS algorithm. When calculating the correlation between BIS and Ibis scores, the Pearson correlation coefficient was $r = 0.9970$.

OpenIbis is a simplified and publicly available version of Ibis, originally implemented in MATLAB. OpenIbis combines the burst suppression ratio and

spectral power ratios to give an index between 0 and 100. When comparing the OpenIbis index against the original BIS index, the correlation coefficient was $r = 0.9395$. OpenIbis tends to underestimate the depth of anaesthesia compared to BIS (i.e. the values generated from OpenIbis tend to be higher than those generated by BIS) [156].

OpenIbis was re-implemented in Python, and OpenIbis values were calculated for all patients in the dataset. The ‘no cleaning’ variant of the EEG dataset was used as the basis for calculation of OpenIbis scores, although small gaps (of less than or equal to five samples) were linearly interpolated prior to processing. The EEG dataset was resampled from 100 Hz to 128 Hz, and then used as the input to OpenIbis. The OpenIbis algorithm outputs one value every 0.5 seconds.

3.4.4 SingLEM Embeddings

SingLEM is a foundation model designed specifically for single-channel EEG data. The EEG dataset (with standard cleaning) was resampled to 128 Hz, and then rescaled by a factor of 0.5 to fit the expected range for SingLEM. Each EEG was then passed through SingLEM. The output of SingLEM is a 16-dimensional embedding for every second of single-channel EEG data [108].

3.5 Dataset Summary

Table 3.2 summarises the three constituent datasets of this study. Only individuals with adequate EEG data and pharmacological data were included in this study. Patients who did not have sufficient or aligned data were excluded from the analysis, as well as any patients who received total intravenous anaesthesia. The final dataset included 102 subjects from BISpain, 265 subjects from AlphaGA, and 200 subjects from AlphaMax, giving a total sample size of 567 patients, with 1,541 total hours of EEG. Compared to the other studies, AlphaMax subjects were older on average and more likely to be male (73.0%). BISpain subjects were younger and more likely to be female (81.3%). AlphaMax patients tended to have lower pain scores (median 0) compared to AlphaGA (median 1) and BISpain (median 5).

3.6 Exploratory Data Analysis

The three datasets were combined. Figure 3.4 shows the distributions of some key parameters. The mean patient age for the combined dataset was 60.9 years (standard deviation 17.7 years), the mean patient weight was 82.4 kg (standard deviation 17.8 kg), and the mean surgical duration was 121.4 minutes (standard deviation 88.5 minutes). The median pain score across all patients was 1 (interquartile range 0–5).

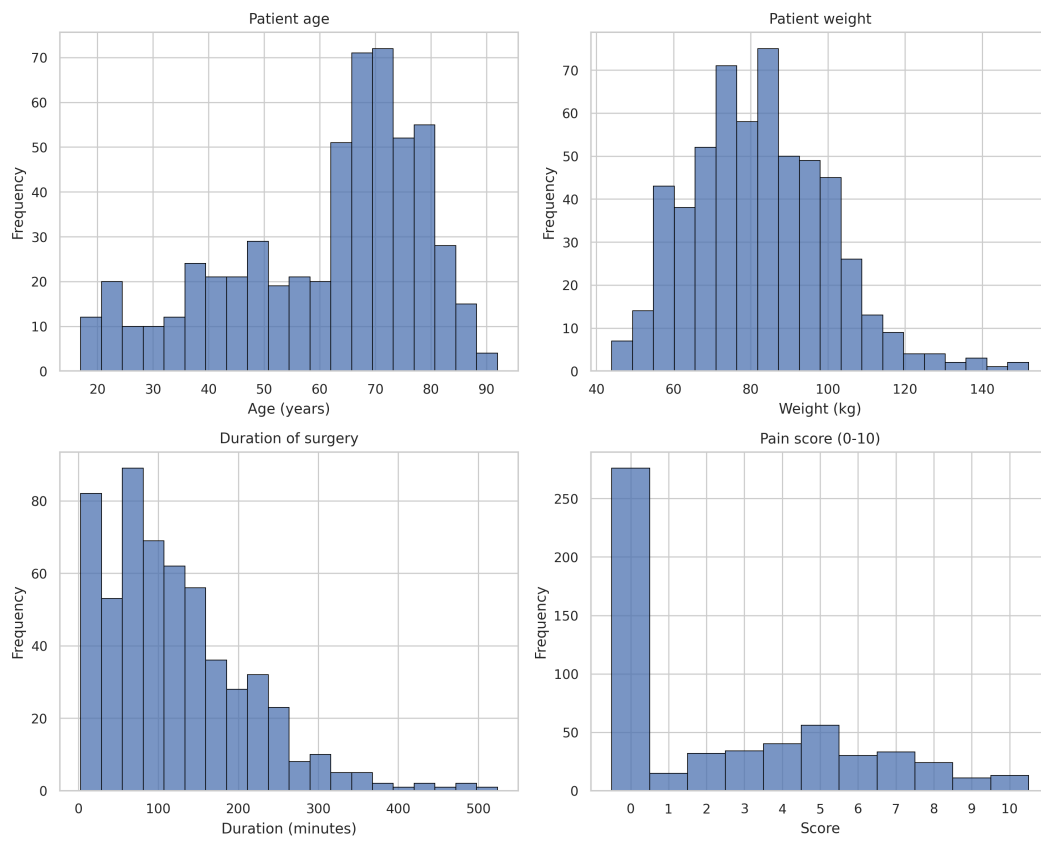


Figure 3.4: Distributions of key patient parameters in the combined dataset.

Table 3.2: Summary of constituent datasets included in this study, combined to give a total dataset with 1,541 hours of EEG data across 567 patients.

Study	BISPain	AlphaGA	AlphaMax
Recruited subjects	135	305	200
Included subjects	102	265	200
Publication year	2014	2019	2025
Location	Waikato Hospital; North Shore Hospital	Waikato Hospital	Waikato Hospital
Total EEG length (hours)	222	567	751
Hypnotic agents	Propofol; desflurane	Propofol; sevoflurane; desflurane	Propofol; desflurane
Analgesic agents	Fentanyl	Fentanyl; morphine	Fentanyl
Age criteria	18–65	18–90	≥ 60
Mean surgery length (minutes)	84.9	91.3	179.0
Mean age (std) (years)	42.1 (13.2)	59.1 (18.1)	72.7 (6.8)
Proportion female (%)	81.3%	49.4%	27.0%
Median pain score (IQR)	5 (0–7)	1 (0–5)	0 (0–4)

276 patients (48.7%) reported no pain postoperatively. 81 patients (14.3%) reported mild pain (pain score 1–3). 126 patients (22.2%) reported moderate pain (pain score 4–6). 79 patients (13.9%) reported severe pain (pain score 7–10).

Figure 3.5 shows the mean value of OpenIbis across all patients, as well as the mean effect-site concentration of fentanyl, volatile anaesthetic, and propofol.

For the purposes of this analysis, pain score was expressed as a binary value, grouping no/mild pain (pain score 0–3) and moderate/severe pain (pain score 4–10). Table 3.3 was then constructed, which shows how the distribution of covariates varies between different pain groups. Chi-squared tests were performed for categorical variables, t -tests were performed for normally distributed numeric variables, and Mann-Whitney U -tests were performed for other numeric variables.

Patients who reported moderate or severe pain were more likely to be from the BISPain study ($p < 0.001$), they were younger ($p < 0.001$), they were more likely to be female ($p < 0.001$), and they also weighed less ($p = 0.019$). There were no significant differences in pain score grouping by duration of surgery.

The patients reporting more severe pain were also more likely to have greater hypnotic depth. They had lower OpenIbis values ($p < 0.001$) and higher concentrations of volatile anaesthetic ($p = 0.017$). They also had greater

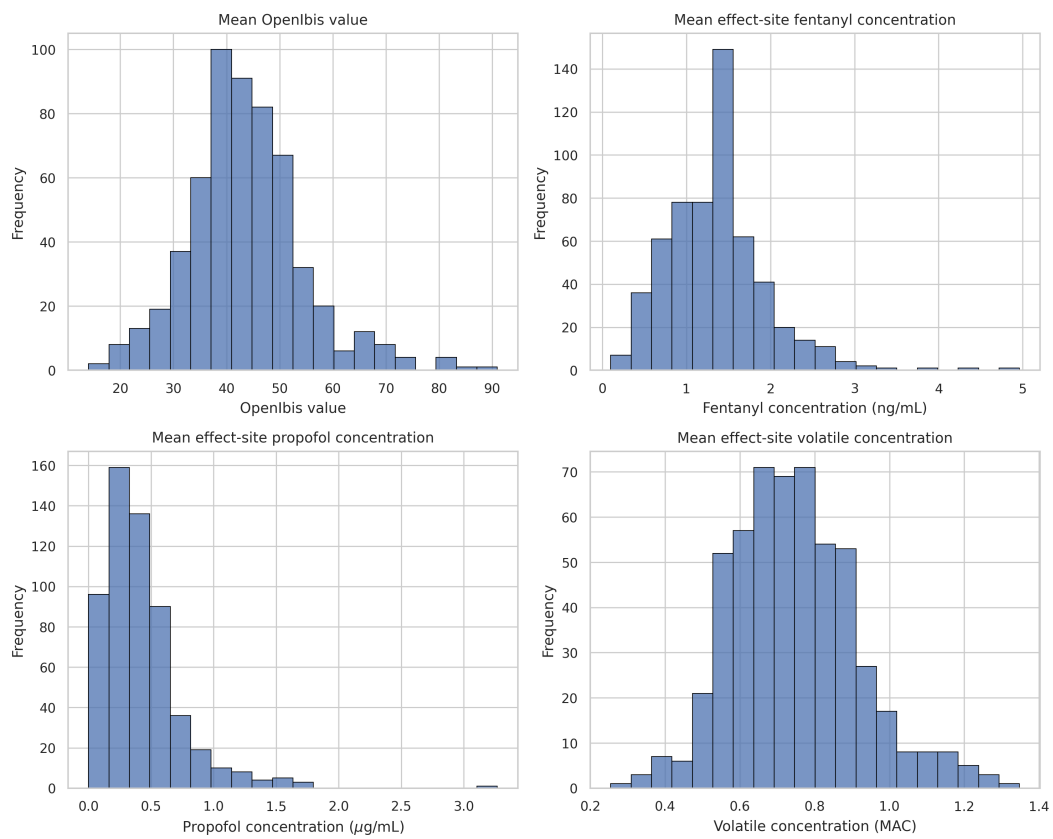


Figure 3.5: Distributions of the mean values of key time-varying parameters.

concentrations of fentanyl, although this difference was not statistically significant ($p = 0.197$). It is possible that these differences are a manifestation of the drug titration paradox. Patients with greater levels of nociception are more likely to demonstrate associated clinical signs, such as movement, changes in haemodynamic parameters, or changes in EEG parameters such as a rise in BIS (relevant to BIS_{Pain}) or a reduction in alpha power (relevant to Alpha_{Max}). Anaesthetists are likely to respond to this by increasing levels of volatile anaesthetic or fentanyl.

Table 3.3: Baseline characteristics stratified by pain group.

		Overall	No/Mild (0-3)	Moderate/Severe (4-10)	<i>p</i> -value
Study, n (%)	AlphaGA	265 (46.7)	174 (48.7)	91 (43.3)	< 0.001
	AlphaMax	200 (35.3)	142 (39.8)	58 (27.6)	
	BIS _{Pain}	102 (18.0)	41 (11.5)	61 (29.0)	
Age (years), median [Q1-Q3]		66.0 [48.0, 74.0]	68.0 [58.0, 76.0]	56.5 [40.0, 71.0]	< 0.001
Sex, n (%)	Female	268 (47.3)	141 (39.5)	127 (60.5)	< 0.001
	Male	299 (52.7)	216 (60.5)	83 (39.5)	
Weight (kg), mean (SD)		82.4 (17.8)	83.8 (17.5)	80.1 (18.2)	0.019
Duration (minutes), median [Q1-Q3]		104.3 [57.0, 169.9]	108.0 [57.8, 182.9]	95.9 [55.1, 148.2]	0.218
Mean OpenBis value, median [Q1-Q3]		43.1 [37.3, 49.2]	44.3 [39.4, 50.6]	39.7 [34.0, 46.8]	< 0.001
Mean fentanyl conc. (ng/mL), mean (SD)		1.4 (0.6)	1.3 (0.6)	1.4 (0.5)	0.197
Mean propofol conc. (μg/mL), median [Q1-Q3]		0.4 [0.2, 0.5]	0.3 [0.2, 0.5]	0.4 [0.2, 0.5]	0.479
Mean volatile conc. (MAC), mean (SD)		0.7 (0.2)	0.7 (0.2)	0.8 (0.2)	0.017
Total		567	357	210	

Chapter 4

Validation Framework and Baseline Predictors

4.1 Motivation

Validation of a putative index for intraoperative nociception is non-trivial due to a lack of reliable ground truth data about nociception levels at each point during surgery. There is no single metric that can validate such an index outside of a full prospective study [143]. Furthermore, in this study, several approaches for predicting intraoperative nociception were attempted, each with different loss functions. For these reasons, it is necessary to establish a set of validation metrics that can apply universally across different approaches. These metrics need to be able to validate the behaviour of indices both on an intra-patient level (i.e. how the index varies over time for a single surgery) and inter-patient level (i.e. how the index behaves between different patients). These metrics need to be physiologically grounded and comprehensive, providing a scaffold upon which the indices developed in this thesis and future works can be evaluated against.

4.2 Train-Validation-Test Split

Appropriate and consistent splitting of patients across train, validation, and test sets is critical for fair and unbiased comparisons of potential indices of nociception. It is important that splitting occurs at the patient level, in order to avoid leakage of patient-level data across splits. For consistency, it is also important that the same test set is used to evaluate different approaches.

Patients were categorised into four categories by pain score, based on standard, validated cut points [157]. These categories were no pain (0), mild pain (1–3), moderate pain (4–6), and severe pain (7–10). The 567 patients in the dataset were then split, with stratification ensuring even balance across pain categories between each of the three splits.

The splits were as follows:

- 70% train set (396 patients).
- 10% validation set (57 patients).
- 20% test set (114 patients).

The validation set was used for initial validation of each model variant. Hyperparameter tuning, early stopping for neural network training, and comparison of different network architectures were performed using the validation set, with each approach evaluated via its own loss function. In some cases, the train set and validation set were combined, either when a larger dataset was required for validation, or when no separated validation set was required.

The test set was used for validation of the best performing model variant of each approach. This was chosen to be larger than the validation set such that there would be sufficient patient data to reliably calculate inter-patient measures.

4.3 Baseline Postoperative Pain Prediction

Prediction of postoperative pain score has been frequently used as a method for evaluating indices of nociception [16, 24, 25]. A brief experiment was conducted to determine how effectively postoperative pain score can be predicted, using only demographic data and the length of surgery. This provides a baseline which can be improved upon with the addition of EEG data.

Three binary targets were defined for this experiment:

- Pain score > 0 (separating no pain from any pain)
- Pain score > 3 (separating no/mild pain from moderate/severe pain)
- Pain score > 6 (separating no/mild/moderate pain from severe pain)

The covariates for this experiment were patient age, weight, sex, and the duration of surgery. Any missing data were median-imputed. Numeric covariates were scaled to have variance 1 and mean 0. For each of the above targets, a random forest model (with 100 estimators and a maximum depth of 4) and a logistic regression model with an L2 penalty ($C = 1.0$) were trained. These models were trained using the combined train and validation set, and evaluated on the test set.

Table 4.1 shows the results from this baseline experiment. The area under the receiver operating characteristic curve (AUROC) ranges between 0.53 and 0.63, indicating that these models have poor discriminative power. Random forest models consistently outperformed the logistic regression models in terms of AUROC, but discriminative power was still weak. For the prediction of pain score > 6 , the modal class (0) was predicted for all patients due to the class imbalance.

Table 4.1: Performance of random forest (RF) and logistic regression (LR) models for postoperative pain classification.

Target	Model	AUROC	Accuracy	Precision	Recall	F1-score
Pain > 0	RF	0.627	0.575	0.582	0.561	0.571
Pain > 0	LR	0.594	0.566	0.567	0.596	0.581
Pain > 3	RF	0.602	0.619	0.469	0.366	0.411
Pain > 3	LR	0.561	0.664	0.556	0.366	0.441
Pain > 6	RF	0.614	0.858	0.000	0.000	0.000
Pain > 6	LR	0.538	0.858	0.000	0.000	0.000

4.4 Validation Framework

Table 4.2 summarises the unified validation framework that was used throughout this study. A nociception index was calculated at each point during surgery for each patient in the test set, and the metrics were then calculated using the calculated index. For the purposes of validation, lower values of an index are assumed to represent lower levels of nociception, and higher values are assumed to represent higher levels of nociception.

There is no reliable, open-source metric of intraoperative nociception. This validation framework demonstrates some assumptions about how one might anticipate an ideal index would behave, based on literature and clinical intuition. However, it is not a replacement for a gold-standard metric, and many of the components could be debated. As a result, violation of one metric does not necessarily invalidate the index in question, and may lead us to reconsider our assumptions about the nature of nociception. Of these metrics, those in Category A (pain score), Category C (responsiveness to start of surgery) and Category D (responsiveness to fentanyl) will be prioritised as they associate any potential index with real world factors outside of the EEG.

4.4.1 Postoperative Pain Prediction (Category A)

Pain score is a noisy, ordinal measure. It is an 11-point scale, ranging from 0 (no pain at all) to 10 (the worst pain imaginable). It is a highly subjective measure obtained simply by asking a patient to rate their present pain on the scale; thus it is affected by numerous patient-level factors. Due to this, pain score was reduced to the four aforementioned standard categories (no pain, mild pain, moderate pain, and severe pain) [157]. This is effectively equivalent to three binary classification tasks at each pain score threshold (> 0 , > 3 , > 6).

The ability of a putative nociception index to predict postoperative pain was assessed by taking the mean value of that index over the final 15 minutes of surgery (i.e. the part of surgery that is closest to when the postoperative

Table 4.2: Validation metrics and expected outcomes for assessment of the proposed pain index.

Category	Metric	Expected Outcome
A. Pain score	A1. AUROC of index for postoperative pain score prediction with 95% confidence interval	Higher AUROC; confidence interval does not contain 0.5
	A2. Comparison of nested logistic regression models with and without the index	Increased AUROC with inclusion of the index; significance accepted at $p < 0.05$
B. OpenIbis	B1. Inter-patient OpenIbis correlation	Weak positive correlation
	B2. Mean intra-patient OpenIbis correlation	Weak positive correlation
C. Start of surgery	C1. Difference between intraoperative and pre-surgical index	More positive
D. Fentanyl bolus correlations	D1. Correlations between fentanyl levels and index during fentanyl bolus periods	More negative
E. Explainability	E1. Regression coefficient of alpha power	Decreases
	E2. Regression coefficient of beta power	Increases
	E3. Regression coefficient of delta power	Increases

pain scores are obtained). In this case, the end of surgery was defined as the recorded end time of surgery (the emergence phase of surgery was not included, as EEG from this phase is often noisy). An AUROC was then calculated at each of the three pain score thresholds using the nociception index. A 95% confidence interval was then constructed via bootstrapping with 2000 resamples of the original data. The three AUROCs and associated confidence intervals constitute validation metric A1. An ideal index would be expected to have high AUROCs for prediction of postoperative pain. Each confidence interval should ideally not contain 0.5, indicating a model that has statistically significant discriminative ability.

It is also important to determine whether an index provides additional predictive ability beyond that which can be achieved using simple demographic factors. It is possible that an index could achieve weak predictive ability simply by encoding demographic factors such as age, weight, or gender.

For each pain score threshold, two logistic regression models were constructed using the combined train and validation set, and they were evaluated on the test set. The first model f_{base} was effectively identical to the logistic regression used in the baseline pain prediction, using age, weight, gender, and surgical duration as predictors. The second model $f_{\text{augmented}}$ used the same predictors, as well as the mean value of the index over the final 15 minutes of surgery. In order to evaluate whether an index provided predictive value, the AUROC of $f_{\text{augmented}}$ was compared to that of f_{base} . An ideal index would lead to a significantly raised AUROC when included in a predictive model for postoperative pain.

Significance testing for a difference between the two models was achieved using the likelihood ratio test [158]. Given covariates $\mathbf{x}_i$ for patient i , a binary outcome $y_i \in \{0, 1\}$, and a coefficient vector $\hat{\beta}$, the log-likelihood of the fit is given by:

$$\ell(\hat{\beta}) = \sum_{i=1}^n [y_i \log \hat{\pi}_i + (1 - y_i) \log(1 - \hat{\pi}_i)], \quad \hat{\pi}_i = \frac{1}{1 + e^{-\hat{\beta}^\top \mathbf{x}_i}}$$

The likelihood ratio is then calculated as follows, where $\ell(\hat{\beta}_{\text{base}})$ and $\ell(\hat{\beta}_{\text{augmented}})$ are the log-likelihoods calculated by fitting f_{base} and $f_{\text{augmented}}$:

$$G^2 = 2 \left[\ell(\hat{\beta}_{\text{augmented}}) - \ell(\hat{\beta}_{\text{base}}) \right]$$

G^2 can be used to generate a p -value using the chi-square test, exploiting that, under the null hypothesis (that the index does not contribute and has a coefficient of 0), $G^2 \sim \chi^2$ asymptotically. An ideal index would lead to an increase in AUROC, as well as a statistically significant improvement in fit according to this statistic. The change in AUROC between the two logistic regression models and the p -value for the difference in fit constitute validation metric A2.

4.4.2 OpenIbis (Category B)

Conflation of the hypnotic and nociceptive dimensions of anaesthesia has been noted as a frequent issue with previous attempts at developing an index of intraoperative nociception [16, 128]. An ideal nociception index should include the component of nociception which lies orthogonal to the hypnotic dimension, so it is necessary to compare any prospective index against an established index of hypnotic depth.

OpenIbis was compared against all prospective nociception indices, with both Pearson’s r (correlation coefficient) and Spearman’s ρ (rank-order correlation) being used to measure the correlation of each index against OpenIbis. This was conducted on two levels: the inter-patient level and the intra-patient level. On the inter-patient level, the mean value of OpenIbis and the mean value of the index were taken over the final 15 minutes of surgery. This gave one OpenIbis and one index value per patient, from which the correlations were then calculated, giving validation metric B1. On the intra-patient level, the correlation between the index and OpenIbis was calculated individually for each patient (using values from throughout surgery, calculated at a rate of one value per second). The mean Pearson’s r and Spearman’s ρ across all patients were taken as an overall measure of intra-patient correlation between OpenIbis and the prospective index, giving validation metric B2.

The BIS has demonstrated some weak ability to detect nociception, with higher BIS scores representing increased nociception (which is associated with wakefulness and shallower anaesthesia) and lower BIS scores representing reduced nociception (which is associated with deeper anaesthesia) [124, 125, 126]. Furthermore, hypnotic anaesthetics have some antinociceptive effect, particularly at high concentrations [159]. For these reasons, an ideal index of pure nociception would therefore be expected to have a weak positive correlation with BIS.

4.4.3 Start of Surgery (Category C)

All patients had a recorded time for the first surgical incision. It would be anticipated that the period immediately before this would have little to no nociceptive stimulus. Therefore, an ideal nociception index would be lower during this period compared to during surgery.

The pre-surgical period $\mathcal{P}$ was defined as the time period between 180 seconds after intubation, and the first surgical incision (intubation is a strong nociceptive stimulus [160] so it was excluded from the pre-surgical period). The surgical period $\mathcal{S}$ was defined as the time between the first surgical incision and the end of surgery. The mean index was calculated for each of these periods, and the difference $\Delta\hat{y}$ between them was calculated. Given index values $\hat{y}_t$ at each time point, this was calculated as follows:

$$\Delta \hat{y} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \hat{y}_t - \frac{1}{|\mathcal{P}|} \sum_{t \in \mathcal{P}} \hat{y}_t$$

It would be anticipated that, in an ideal index, this metric would be highly positive, representing the increase in nociception that one would anticipate with the commencement of surgery. This is validation metric C1.

4.4.4 Fentanyl Bolus Correlations (Category D)

As fentanyl is an analgesic, it would be anticipated that increasing effect-site concentration of fentanyl would be associated with a reduction in nociception. However, the interpretation of the overall correlation between fentanyl levels and any potential nociception index is complicated by the drug titration paradox (i.e. the anaesthetist is likely to give more fentanyl in response to anticipated or apparent increases in nociceptive stimulus throughout surgery). This analysis has therefore been limited to time points in surgery where fentanyl is increasing following a bolus. At these time points, the assumption is made that nociceptive stimuli are relatively constant, and any changes in nociception index are due to the increase in fentanyl. The increase in fentanyl in these periods is due to the time required for fentanyl to diffuse to the effect site following a bolus, rather than deliberate control on the part of the anaesthetist. This means that the interpretation of changes to EEG, at least within boluses, should be free from the drug titration paradox.

The time segments of interest were identified as any point where effect-site concentration of fentanyl was rising at a rate of at least 0.01 ng/mL per second for at least 30 seconds (and up to a maximum duration of 1000 seconds). For each bolus-associated segment, the effect-site concentration of fentanyl and nociception index level were taken every second, and the correlations (Pearson’s r and Spearman’s ρ) between them were calculated. The mean values of these correlation coefficients were calculated across all bolus-associated segments, giving validation metric D1. An ideal index of nociception should correlate negatively with the effect-site concentration of fentanyl.

4.4.5 Explainability Analysis (Category E)

Most of the models used to generate indices of nociception throughout this study were black-box neural network models. These models are not inherently interpretable, which can be problematic for the deployment of these models in a healthcare setting. Poor explainability leads to reduced trust and uptake of models by clinicians [161]. There are increasing regulatory frameworks internationally encouraging the use of explainable AI, including the European Union Artificial Intelligence Act, which requires AI models used for medical decision support to be either inherently interpretable or use post-hoc explainability techniques [162].

One class of post-hoc explainability techniques that are frequently used are surrogate models, which involve an inherently interpretable model being trained on the inputs and predicted outputs of a black-box model [163]. In this case, a linear regression model was used as a surrogate model. For each input in the test set (i.e. each time point across all test set patients), the average delta, alpha, and beta power were calculated by using a short-time Fourier transform (STFT). The natural log of these powers was taken, and they then underwent standard z-score normalisation. The normalised delta, alpha, and beta power were used as predictors for the nociception index. The coefficients of the linear regression can then be inspected to determine the relative effects of each frequency band on the index.

Explainability analyses can be used as a validation technique. An ideal nociception index would have findings consistent with the literature, which suggests an increase in nociception is associated with an increase in beta power, a decrease in alpha power, and a paradoxical increase in delta power. It is noted that these findings are not consistent across all studies, so findings that are incongruent with this pattern do not necessarily invalidate a prospective nociception index [12]. The coefficients for delta power, alpha power, and beta power give metrics E1, E2, and E3 respectively.

4.5 Baseline Indices

Some baseline indices were employed to demonstrate the unified validation framework, and to provide some baseline comparisons against which more sophisticated models could be benchmarked.

4.5.1 Constant Index

This is a trivial index which outputs zero at all time points. Table 4.3 gives the results for this index.

4.5.2 OpenIbis

The BIS has been shown to have some value as a nociception index [124, 125, 126]. Ideally, an existing commercial index specifically for nociception, such as qNOX or PTI would be analysed here. However, these indices are proprietary and no open-source variants exist. OpenIbis was therefore used as a comparison. The results of the analysis of OpenIbis are shown in Table 4.4.

OpenIbis was a poor predictor of pain score category, with all AUROCs being below 0.5. This indicates that OpenIbis serves better as an inverse predictor of postoperative pain. At the start of surgery, OpenIbis decreased on average rather than increased. It did appear to be responsive to fentanyl, with weak negative correlations in response to fentanyl boluses. Increasing delta and alpha were associated with lower OpenIbis scores, whereas increasing beta was

Table 4.3: Summary of test set results for constant index.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.500 (0.500–0.500)
AUROC > 3	0.500 (0.500–0.500)
AUROC > 6	0.500 (0.500–0.500)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.000 ($p = \text{NaN}$)
Δ AUROC > 3	0.000 ($p = \text{NaN}$)
Δ AUROC > 6	0.000 ($p = \text{NaN}$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = \text{NaN}$, Spearman $\rho = \text{NaN}$
B2. Intra-patient corr.	Mean Pearson $r = \text{NaN}$, Mean Spearman $\rho = \text{NaN}$
C. Start of surgery	
C1. Mean comparison	Pre: 0.000, Surg: 0.000, Diff: 0.000
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = \text{NaN}$, Mean Spearman $\rho = \text{NaN}$
E. Explainability	
E1. Delta coeff.	0.000 ($p = \text{NaN}$)
E2. Alpha coeff.	0.000 ($p = \text{NaN}$)
E3. Beta coeff.	0.000 ($p = \text{NaN}$)

Table 4.4: Summary of test set results for OpenIbis.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.381 (0.283–0.493)
AUROC > 3	0.444 (0.337–0.557)
AUROC > 6	0.526 (0.337–0.700)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.015 ($p = 0.210$)
Δ AUROC > 3	0.015 ($p = 0.096$)
Δ AUROC > 6	−0.012 ($p = 0.327$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = 1.000$ ($p < 0.001$), Spearman $\rho = 1.000$ ($p < 0.001$)
B2. Intra-patient corr.	Mean Pearson $r = 1.000$, Mean Spearman $\rho = 1.000$
C. Start of surgery	
C1. Mean comparison	Pre: 40.366, Surg: 38.825, Diff: −1.541
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.258$, Mean Spearman $\rho = -0.248$
E. Explainability	
E1. Delta coeff.	−1.786 ($p < 0.001$)
E2. Alpha coeff.	−3.875 ($p < 0.001$)
E3. Beta coeff.	5.962 ($p < 0.001$)

associated with higher OpenIbis scores. These findings are consistent with OpenIbis being primarily an index of hypnotic depth rather than an index of nociception, and they suggest that the value of OpenIbis for predicting nociception is very limited.

4.6 Pain Score Category Prediction

4.6.1 Methods

Direct prediction of the pain score category was used to generate an initial index of nociception. This is a comparatively simple approach for pain score prediction. The final fifteen minutes of EEG data from each surgery were taken, and divided into ten-second segments. Each segment was labelled with the corresponding postoperative pain score category for that patient (none, mild, moderate, severe).

Given that postoperative pain scores are an ordinal category, Consistent Rank Logits (CORAL) was used. This approach changes ordinal classification into a series of binary classification tasks, with each task representing whether the threshold for each category has been exceeded [164]. In this case, the ordinal classification was changed into a series of three binary classification tasks representing whether the patient exceeded the threshold for mild pain, moderate pain, or severe pain (i.e. no pain was represented as $[0, 0, 0]$, mild pain as $[1, 0, 0]$, moderate pain as $[1, 1, 0]$, and severe pain as $[1, 1, 1]$).

In CORAL, a model $f(x)$ is defined, which gives a single scalar index as an output for each input x . For each pain score threshold k , a bias term β_k is trained. Given the sigmoid function σ , pain score p , and EEG input x , the probability that a pain score exceeds a threshold is given by:

$$P(p > k \mid x) = \sigma(f(x) + \beta_k)$$

The model $f(x)$ and bias term β_k are trained end-to-end. Due to the formulation of the labels, the bias terms will be monotonically decreasing (i.e. $\beta_0 \geq \beta_3 \geq \beta_6$). The result of this is that the probability predictions of each class will also be monotonically decreasing (i.e. $P(p > 0 \mid x) \geq P(p > 3 \mid x) \geq P(p > 6 \mid x)$). $f(x)$ can then be used as a single scalar index of nociception.

Several types of model were trained for $f(x)$, each with a different input formulation:

- A one-dimensional CNN was trained using time-domain EEG data. This was a basic CNN with four convolutional layers ([32, 64, 128, 256]), followed by a two-layer perceptron ([128, 64]) with a dropout layer (rate 0.3) between the perceptron layers.
- Data were transformed into the frequency domain using the STFT (with a maximum frequency of 40 Hz). The magnitudes of the STFT were

averaged across the time segment, and squared to give a power spectrum with 103 frequencies. The log of the power spectrum was used as the input to a three-layer multilayer perceptron ([256, 128, 64]).

- The STFT was used to generate a two-dimensional spectrogram of each input signal with a time resolution of one second. This was used as the input to a two-dimensional CNN ([32, 64]), and then a transformer layer.
- Similarly, the STFT was used to generate a two-dimensional spectrogram, but phase data were retained. Phase data were represented as a sine-cosine embedding for each point on the spectrogram, giving a three-dimensional representation of the STFT output. This was used as the input to a similar CNN-transformer.
- The SingLEM embeddings were calculated for each second of each segment. They were then passed through a two-layer perceptron ([128, 64]), before being mean-pooled across the time dimension.

Each model was trained with both the standard clean dataset, and the EEGdenoiseNet clean dataset. They were also trained with varying segment sizes: 10 seconds, 30 seconds, 60 seconds, and 120 seconds. The temporal resolution of the dataset remained 10 seconds for each variant in order to ensure all variants would be comparable on a label level (therefore each variant from 30 seconds upwards involved overlapping segments in the input data).

All models were trained using binary cross-entropy across the three binary CORAL targets, and an AdamW optimiser with a learning rate of 1×10^{-3} , except for the transformer models which were trained with a learning rate of 1×10^{-4} . Models were validated on the validation set and early stopping using the validation loss (with a patience of 10 epochs) was used to recover the best model.

4.6.2 Rank-Based Transform

The best model $f(x)$ outputs an arbitrarily scaled quantity which corresponds to the level of nociception. In order to generate an index between 0 and 100, $f(x)$ undergoes a rank-based cumulative distribution function (CDF) transform, using the distribution of output indices from the combined train and validation set. This is equivalent to determining which percentile the index is at, where 0 represents the minimum index level in the train/validation dataset, and 100 represents the maximum index level in the train/validation dataset.

Given samples $(x_i)_{i=1}^n$ from the combined train and validation set, and an arbitrary sample x , the transformed score $\phi(x)$ can be calculated as follows:

$$\phi(x) = 100 \cdot \frac{1}{n} \sum_{i=1}^n \mathbf{1}(f(x_i) \leq f(x)) \quad (4.1)$$

This transform is monotonic and preserves the rank-order of the predicted indices while ensuring that it ranges between 0 and 100.

4.6.3 Results

The two validation metrics used were the binary cross-entropy loss function, and mean absolute error (MAE), where the four classes were treated as 0, 1, 2, and 3. The results on the validation set are shown in Table 4.5. According to the MAE, the best performing model was the one-dimensional CNN using the standard EEG data. Note that, if the modal class was predicted for all patients (i.e. no pain), the MAE would be 1.023. Most of the models tested demonstrate only a very modest improvement from this, if any. Figure 4.1 shows the training and validation curves for the best model. Overfitting appears to occur almost immediately, with rising validation loss and falling training loss by the second epoch.

Table 4.5: Best validation loss (MAE in brackets) across segment sizes and preprocessing pipelines. The pipeline with the lowest MAE is in **bold**.

Model	10s	30s	60s	120s
Standard cleaning				
1D CNN	0.612 (0.945)	0.634 (0.993)	0.629 (0.986)	0.731 (0.953)
PSD	0.638 (1.071)	0.649 (1.065)	0.771 (1.092)	0.768 (1.171)
STFT (mag)	0.641 (1.016)	0.654 (1.022)	0.654 (0.966)	0.648 (1.015)
STFT (all)	0.616 (0.976)	0.642 (1.048)	0.669 (1.024)	0.653 (1.023)
SingLEM MLP	0.596 (0.984)	0.597 (0.998)	0.605 (0.996)	0.622 (1.026)
EEGdenoiseNet cleaning				
1D CNN	0.590 (0.950)	0.601 (0.972)	0.607 (0.993)	0.608 (1.016)
PSD	0.633 (1.052)	0.656 (1.091)	0.669 (1.009)	0.706 (1.131)
STFT (mag)	0.625 (0.990)	0.682 (1.034)	0.657 (0.988)	0.682 (0.983)
STFT (all)	0.624 (0.998)	0.663 (1.027)	0.657 (1.017)	0.679 (1.004)
SingLEM MLP	0.595 (0.991)	0.598 (0.982)	0.603 (1.002)	0.607 (1.036)

After rank-based transform of the results from this index, the index was validated using the established validation pipeline. Table 4.6 gives the results of this validation. The ability of the model to predict postoperative pain was mixed: it was only able to predict postoperative pain score at the > 0 threshold on its own, and when combined in a model with demographic factors, the improvement in AUROC was modest and only significant for the > 0 and > 3 thresholds.

The primary issue with this index is the moderately strong negative correlation that OpenIbis appears to have with the index, particularly on an

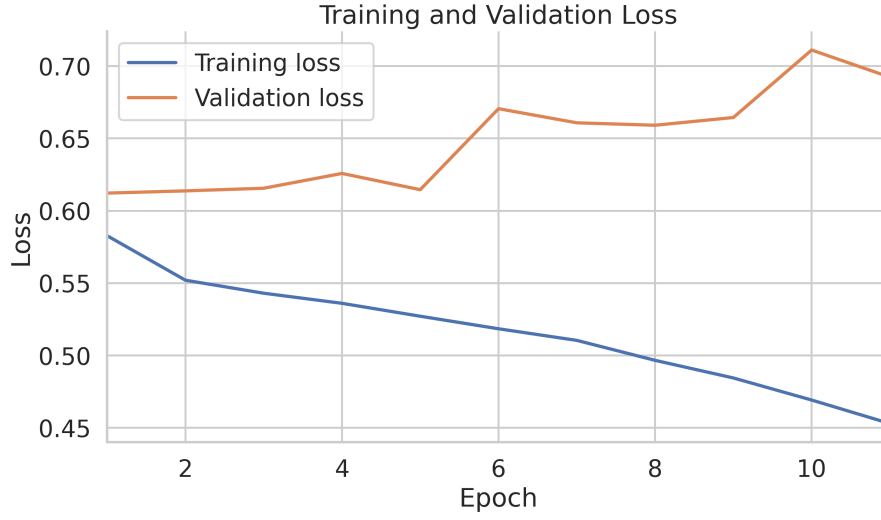


Figure 4.1: Training and validation curves for the best-performing pain score category prediction model.

Table 4.6: Summary of test set results for the best-performing pain score category prediction model.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.621 (0.516–0.721)
AUROC > 3	0.571 (0.454–0.678)
AUROC > 6	0.572 (0.391–0.730)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.031 ($p < 0.001$)
Δ AUROC > 3	0.026 ($p = 0.002$)
Δ AUROC > 6	0.048 ($p = 0.075$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.703$ ($p < 0.001$), Spearman $\rho = -0.687$ ($p < 0.001$)
B2. Intra-patient corr.	Mean Pearson $r = -0.364$, Mean Spearman $\rho = -0.234$
C. Start of surgery	
C1. Mean comparison	Pre: 52.510, Surg: 58.044, Diff: 5.535
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = 0.105$, Mean Spearman $\rho = 0.090$
E. Explainability	
E1. Delta coeff.	12.639 ($p < 0.001$)
E2. Alpha coeff.	19.650 ($p < 0.001$)
E3. Beta coeff.	0.536 ($p < 0.001$)

inter-patient level ($r = -0.703$). This is opposite to the expected correlation, and suggests that this index has been contaminated with the inverse of the hypnotic state.

This index demonstrated a small rise at the start of surgery (5.535), and had a positive correlation with effect-site concentration of fentanyl rather than the expected negative one. Increases in this index were associated with increases in delta, alpha, and beta powers, with the rise in delta power being most significant. The large rises in delta and alpha power are consistent with hypnotic state contamination. This may also explain why the correlation with fentanyl is the opposite of what would be anticipated from an index of nociception.

Figure 4.2 shows the value of this index over time for three example EEG recordings from the test dataset, with one from each constituent study, along with a five-minute moving average.

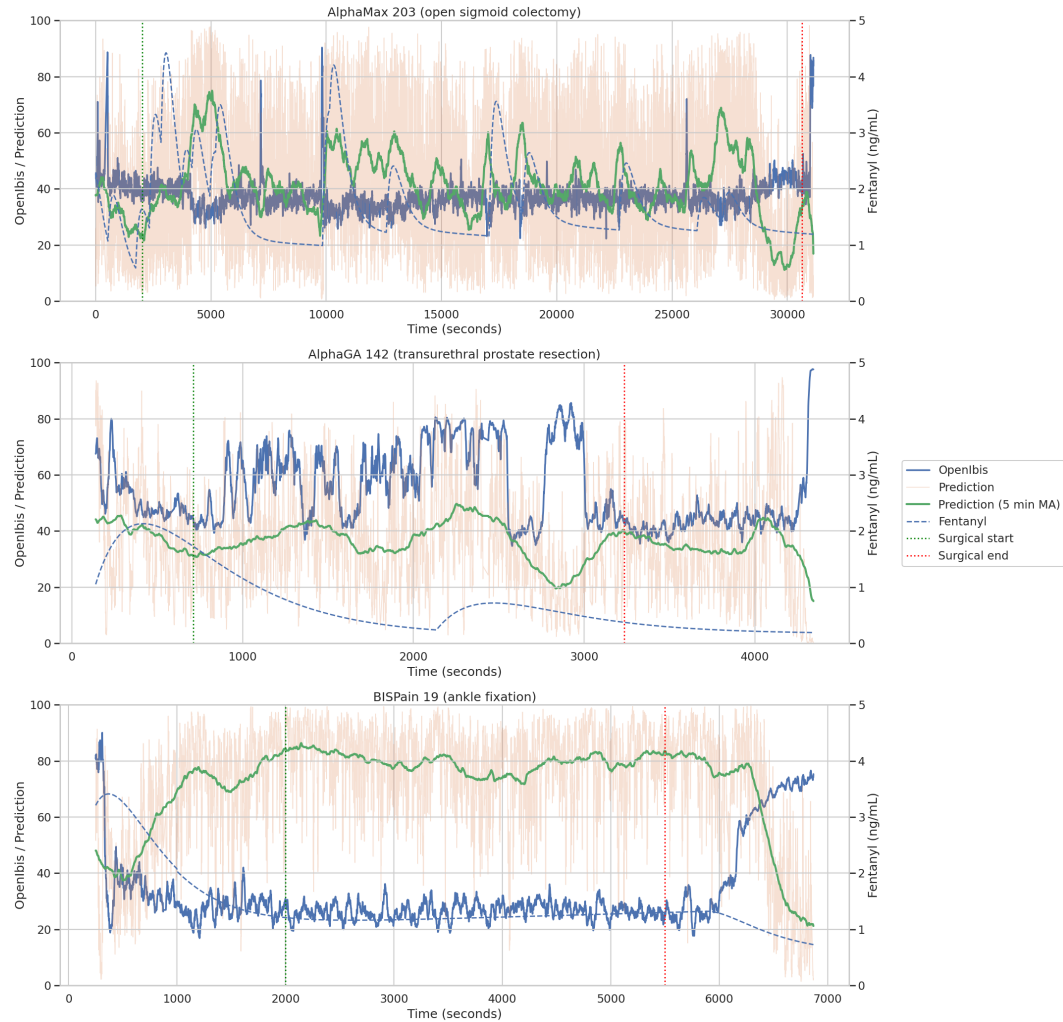


Figure 4.2: The best-performing model applied to three example EEG recordings from the test dataset.

4.6.4 Discussion

The index generated by directly predicting pain score category did not perform well on most metrics, and only had a modest improvement over prediction of the modal class. There was only one pain label per patient, but each patient had many segments of EEG data. This leads most models to very quickly overfit on the train set. Due to the small number of patients, it becomes easier for a model to identify individual patients rather than generalisable factors for prediction of postoperative pain. The overfitting becomes especially severe for higher capacity networks with large amounts of overlap between segments in the train set. This potentially explains why the most simple network with only ten-second segments was able to perform best on the validation set.

The other significant issue with this pipeline was the apparent contamination of the index with the inverse of hypnotic state, leading to relatively strong negative correlations between OpenIbis and the index. This is possibly a consequence of the drug titration paradox. As established earlier, individuals with lower OpenIbis scores (deeper anaesthesia) tend to have greater postoperative pain scores. Individuals who have signs of nociception (which would manifest as changes to EEG indices, movement, or changes in haemodynamic status) would be more likely to receive more opioid and volatile anaesthetic, leading to an increase in hypnotic depth. This effect may also be mediated by the direct effect of opioids on EEG, whereby they cause an increase in delta waves (which would result in an increase in the apparent depth of hypnotic state) [116]. Hypnotic state is substantially easier to infer from EEG than nociceptive state [165]; hence it is easier for a model to learn features associated with hypnotic state (and therefore postoperative pain score via the titration paradox association) than to learn features associated with nociception.

It is also plausible that the association between greater postoperative pain scores and deeper anaesthesia is partly a reflection of true underlying physiology. Increased delta waves have been seen in association with nociception [12]. In isolation, this would manifest as a lower OpenIbis score, although it is unclear whether this effect would be offset by the other changes (such as increased beta waves) that are associated with nociception.

This model also extrapolates outside the context of its training data. The model is specifically trained only on EEG data from the final 15 minutes of surgery. However, the use case is for EEG data from all points during surgery, which it may not be able to generalise to.

There is only one postoperative pain score per patient. Using postoperative pain category as the only label highly limits the ability of a model to demonstrate how nociception changes over the course of a surgery, and it effectively causes the sample size of the study population to become very small. In order to better utilise the full dataset, it will be necessary to either use unsupervised learning techniques, or use labels that change over time (such as fentanyl or hypnotic effect-site concentrations). These techniques will be addressed in the following chapters.

Chapter 5

Unsupervised Learning Approaches

5.1 Motivation

It is possible that unsupervised machine learning techniques, such as principal component analysis (PCA) or exploratory factor analysis (EFA), may be able to generate low dimensional representations of intraoperative EEG data which contain useful information about nociceptive state. Changes attributable to hypnotic agents are highly dominant in intraoperative EEG [82]. In order to generate an index of nociception, the hypnotic effect on EEG will have to be separated from the nociceptive effect. Although there are associations between the nociceptive and hypnotic dimensions of anaesthesia, they are not always coupled [3]. It may therefore be possible to separate them using PCA or EFA. Inter-patient factors also have a significant effect on EEG [100], and also need to be separated from the nociceptive effect.

Microstate analysis, generally achieved by clustering of multiple-channel EEG data, has shown promise in analysing pain in conscious patients [114], as well as distinguishing between consciousness and unconsciousness [98]. It is therefore plausible that microstate analysis, including the use of associated features (such as state occupancy and transition dynamics) may yield value when attempting to generate an index to quantify nociception. It has been suggested that single-channel EEG data generated from low density electrode arrays could potentially identify microstates using a hidden Markov model (HMM) [166].

5.2 Dimensionality Reduction

5.2.1 Background

PCA

Given a data matrix $X \in \mathbb{R}^{n \times d}$ with n samples and d features (which are centred to have zero mean), the eigendecomposition of the covariance matrix $\Sigma = \frac{1}{n-1}X^\top X$ gives eigenvectors v_i and eigenvalues λ_i , ordered such that $\lambda_1 \geq \lambda_2 \geq \dots$. Given the top k eigenvectors and eigenvalues (where k indicates the desired dimensionality), a projection matrix W_k can be constructed where $W_k = [v_1, v_2, \dots, v_k]$. This matrix can then be used to calculate the final projection: $Z = XW_k$. This is the optimal rank- k approximation of X in the least-squares sense.

The result of this is that the first eigenvector v_1 is selected to have maximum possible projected variance (concretely, $v_1 = \arg \max_{\|v\|=1} \text{Var}(Xv)$), and subsequent components are selected to have maximum possible projected variance conditional to being orthogonal to previous dimensions [167].

Due to their large effect on PCA, it would be anticipated that the early components of PCA would capture variance attributable to the hypnotic dimension. Nociceptive state is correlated with hypnotic state, but not perfectly so [3]. There therefore may exist a component of the nociceptive effect on EEG that lies orthogonal to the hypnotic effect. This could potentially be recovered by PCA and used as an index of nociception.

EFA

EFA treats observed features as being generated by a smaller number of latent factors, mapping the data matrix $X \in \mathbb{R}^{n \times d}$ to a lower dimensional latent factor matrix $F \in \mathbb{R}^{n \times k}$. This mapping is defined by a factor loading matrix $\Lambda \in \mathbb{R}^{d \times k}$, and a residual error matrix $E \in \mathbb{R}^{n \times d}$ (where the columns have a diagonal covariance matrix Ψ). The mapping is then calculated as follows:

$$X = F\Lambda^\top + E$$

Multiple algorithms exist for EFA, such as maximum likelihood estimation, principal axis factoring, and least-squares off-diagonal covariance minimisation. These algorithms use Σ , the covariance matrix of X , and attempt to define Λ and E such that $\Sigma \approx \Lambda\Lambda^\top + \Psi$.

The key difference from PCA is that EFA does not necessarily generate orthogonal components. Oblique rotation methods such as oblimin and promax allow for non-orthogonal factors to be generated [168]. As the nociceptive and hypnotic dimensions of anaesthesia are correlated, this may be a desirable characteristic. EFA may recover a more complete index of nociception, as opposed to PCA which will only recover the component orthogonal to the hypnotic dimension.

5.2.2 Visualisation

The key dataset used in this analysis was the output of the STFT calculated previously using the standard clean EEG, giving a time-frequency domain representation of each EEG recording. This is calculated with a hop length of one second, and a maximum frequency of 40 Hz, giving 103 frequency bins per second. Raw EEG has an unstable covariance structure and is highly non-stationary, making it unsuitable for PCA [169]. A time-frequency representation (i.e. a 2D spectrogram) was therefore chosen as the input to the PCA.

Five variants of the STFT dataset $D_{\text{train/val}}^{(w)}$ were generated from the combined train and validation dataset, where $w \in \{1, 10, 30, 60, 120\}$ is the number of seconds of EEG data used to generate each input. $D_{\text{train/val}}^{(1)}$, for instance, uses only one second of data. Because the time resolution is one second, this effectively becomes a pure frequency-domain dataset. $D_{\text{train/val}}^{(60)}$ and $D_{\text{train/val}}^{(120)}$ were down-sampled to take one sample every two seconds, and every four seconds respectively. The output of SingLEM $D_{\text{train/val}}^{\text{SingLEM}}$ was used as a secondary dataset, giving a 16-dimensional embedding for each second of EEG data.

Due to the large size of the total dataset, a sample of time points (along with their associated spectrograms and embeddings) was used for the PCA analysis. Broadly, larger samples have been shown to be better, but a frequent ‘rule-of-thumb’ that is used for determining appropriate population sizes for PCA is ten times the number of features [170]. 275 samples were taken per patient, giving a total sample number of 124,575 (which is over 40 times the dimensionality of the input dataset). The samples were taken evenly across all time points.

Data from all patients in the combined train/validation set were used to generate visualisations to determine if any obvious clusters or microstates formed. All datasets were standardised using standard Z-score scaling to achieve a mean of 0 and a standard deviation of 1 for each feature. The STFT datasets then underwent PCA to reduce dimensionality to 50, while the SingLEM dataset kept a 16-dimensional embedding. All datasets were then reduced into two dimensions using Uniform Manifold Approximation and Projection (UMAP), using the default parameters of a neighbour count of 15, and a minimum Euclidean distance of 0.1. These plots were visually inspected for any obvious clusters or associations with OpenIbis or specific patients.

Figure 5.1 shows the plots for each variant of the dataset, coloured by density of points. Visually, there do not appear to be obvious clusters in the one second data, nor in the SingLEM embeddings, which both appear to be homogeneous. More variation in density appears with longer input windows.

Figure 5.2 shows the plots coloured by OpenIbis index. In all STFT plots, OpenIbis appears to be dominant along one or both UMAP axes. The SingLEM embeddings appear to be more homogeneously distributed in terms of OpenIbis scores.

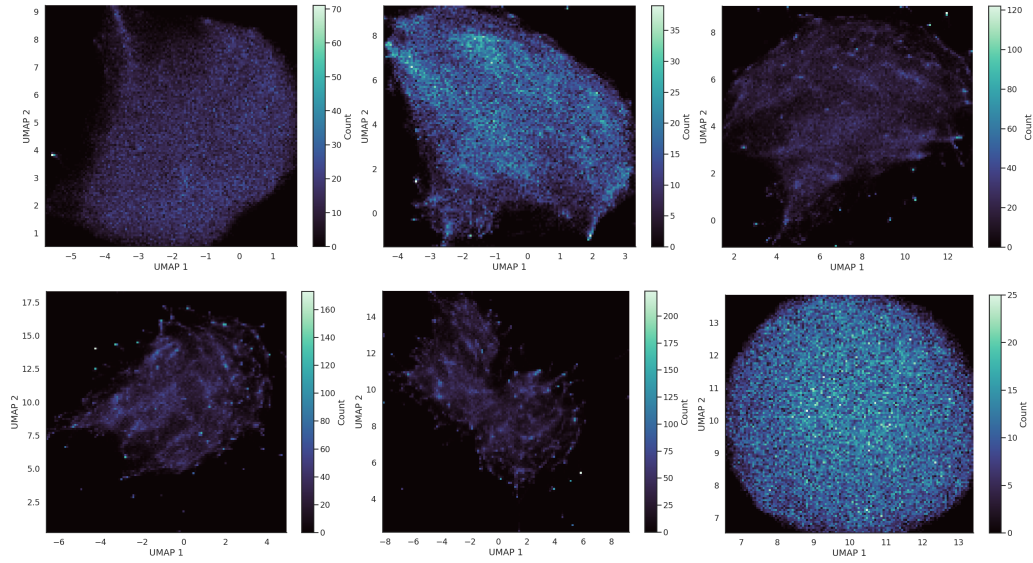


Figure 5.1: UMAP output for:

Top row: 1 s STFT dataset, 10 s STFT dataset, 30 s STFT dataset.

Bottom row: 60 s STFT dataset, 120 s STFT dataset, SingLEM dataset.

Coloured by density.

Figure 5.3 shows plots coloured by patient ID. Due to the large number of patients, this plot was reduced to show only five patients. In smaller time windows, although there is still some clustering attributable to patient identity, most patients appear to have data points distributed widely across the manifold. When longer time windows are used, patient data points appear to form denser clusters with less overlap. The SingLEM embeddings again demonstrate an apparent homogeneous distribution of data points.

Obvious clusters that may be attributable to microstates are not apparent from these visualisations. Due to the short duration of microstates [95, 96], it would be anticipated that these would be most apparent in the datasets with shorter inputs. However, this does not appear to be the case. More density variation appears in datasets with longer inputs, however these clusters are unlikely to represent microstates. The datasets with longer inputs appear to cluster data points from the same patient closer together, as seen in Figure 5.3, which explains the observed density variation.

5.2.3 Methods

Two analyses were conducted: one using PCA, and one using EFA. The broad aim of this section was to decompose the EEG data into lower dimensional components/factors. These components/factors were then analysed to determine if they fulfilled the previously defined criteria for an ideal index of nociception.

The same STFT dataset as previously was used in this analysis, with

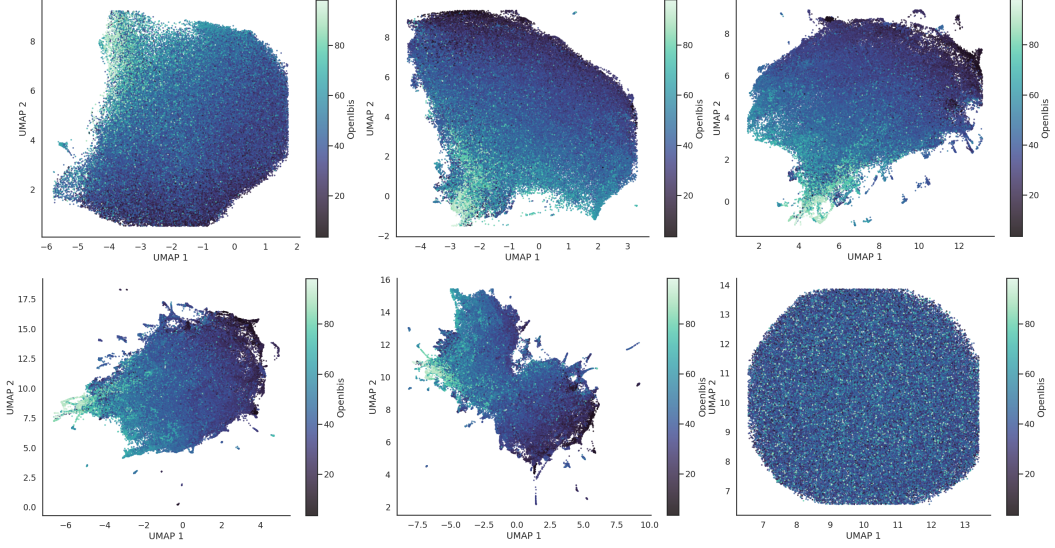


Figure 5.2: UMAP output for:

Top row: 1 s STFT dataset, 10 s STFT dataset, 30 s STFT dataset.

Bottom row: 60 s STFT dataset, 120 s STFT dataset, SingLEM dataset.

Coloured by OpenIbis.

the same five variants $D_{\text{train/val}}^{(w)}$, with $w \in \{1, 10, 30, 60, 120\}$. For the PCA analysis, down-sampling was performed as previously, with every second sample taken for $D_{\text{train/val}}^{(60)}$ and every fourth sample taken for $D_{\text{train/val}}^{(120)}$. For the EFA analysis, the down-sampling was more aggressive to account for the increased computational requirements of EFA: $D_{\text{train/val}}^{(30)}$ was down-sampled to every fourth sample, $D_{\text{train/val}}^{(60)}$ was down-sampled to every eighth sample, and $D_{\text{train/val}}^{(120)}$ was down-sampled to every sixteenth sample.

Each dataset was assembled as a matrix $X_{\text{sample}}^{(w)}$, combining the previously defined 275 samples from each patient. Any rows that included NaN indicators were dropped, and all rows were scaled to have a mean of 0 and a variance of 1. Both PCA and EFA were conducted with a component/factor count of $k = 10$, after scree plots showed eigenvalues reducing substantially after this point.

PCA was conducted using default Scikit-learn parameters. EFA was conducted using the Factor Analyzer library using the minimum residual method, with an oblimin rotation method (therefore permitting substantial correlation between factors). The fit PCA and EFA models were then used on each full dataset $X_{\text{train/val}}^{(w)}$, giving a PCA component analysis matrix $Z_{\text{train/val}}^{(w)}$ and an EFA latent factor matrix $F_{\text{train/val}}^{(w)}$ for each of the five variants.

Each component or factor $j \in \{1, \dots, k\}$ derived from the PCA/EFA was used to generate a candidate index for patient p : $z_{p,j}^{(w)}(t)$ for PCA, and $f_{p,j}^{(w)}(t)$ for EFA. The index values for component j across all patients in the

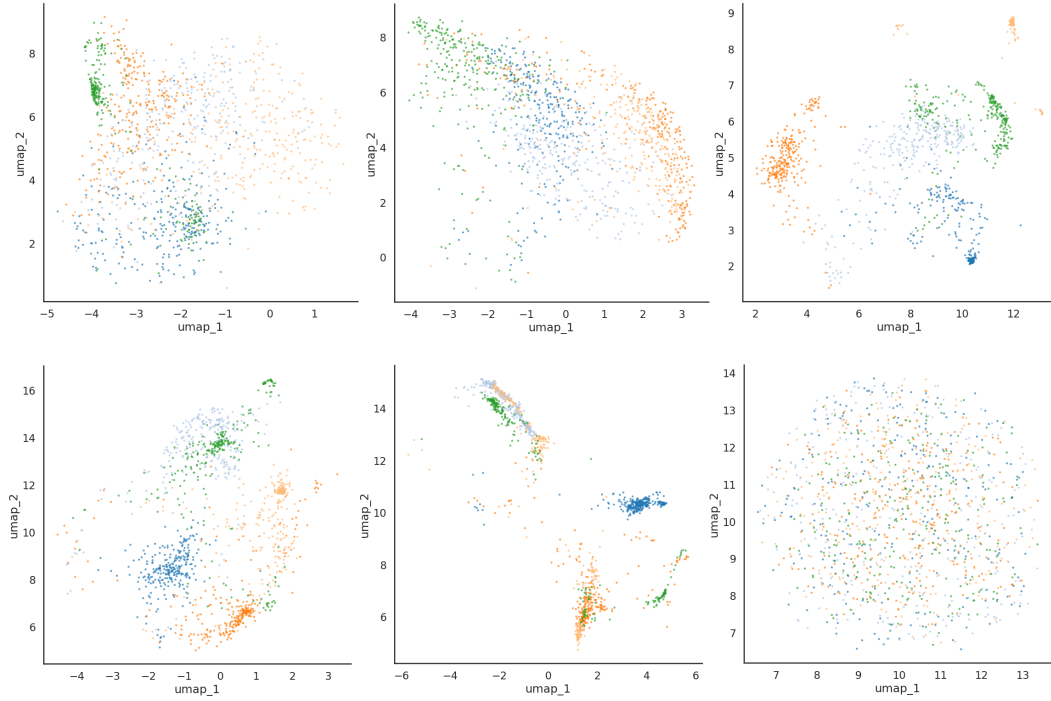


Figure 5.3: UMAP output for:

Top row: 1 s STFT dataset, 10 s STFT dataset, 30 s STFT dataset.

Bottom row: 60 s STFT dataset, 120 s STFT dataset, SingLEM dataset.

Coloured by patient ID (for a sample of five patients).

train/validation dataset were used as a reference distribution for the monotonic CDF transform described in Equation 4.1, giving $\phi_{p,j}^{(w)}(t)$, a 0–100 index for each component j . There were ten candidate indices with five dataset variants, across PCA and EFA, giving a total of 100 candidate indices.

Each of these indices were evaluated using the previously defined evaluation factors: ability to predict postoperative pain, weak positive correlation with OpenIbis, rise associated with the start of surgery, responsiveness to fentanyl, and congruence of surrogate model coefficients with known spectral predictors of nociception. This evaluation was performed using the corresponding train/validation set $D_{\text{train/val}}^{(w)}$ from which that index was derived.

Each candidate index was evaluated across each of the criteria, with Category A metrics (prediction of postoperative pain), Category C metrics (rise at the start of surgery), and Category D metrics (fentanyl responsiveness) being the most strongly considered factors. It should be noted that PCA components and EFA factors are sign indeterminate. Any eigenvector v could equally be replaced with $-v$ without substantially changing PCA, and any factor f could be equally replaced with $-f$ without substantially changing EFA. The result is that the evaluation of every candidate index needs to be considered in both directions. An index that has an AUROC significantly below 0.5, a fall at the start of surgery, and a positive correlation with fentanyl levels, is

equally valuable to an index that fulfils these criteria in the other direction. Each candidate index was therefore evaluated to determine if it fulfilled the criteria in either direction.

In order to account for selection bias, the best candidate index underwent final evaluation on the test set. The test set underwent the same pre-processing as the train/validation set in order to give $D_{\text{test}}^{(w)}$. Standard scaler parameters and the CDF reference distribution were taken from the train/validation set to avoid data leakage. For explainability purposes, the loadings mapping the input STFT representation to the best index were extracted and plotted.

5.2.4 Results

The full train/validation set results for each candidate index are given in Appendix A. Generally, it appeared as though the hypnotic dimension of anaesthesia was captured by the first or second principal component for PCA, and subsequent components had low levels of correlation with OpenIbis. Factors that performed well on the validation metrics tended to appear later, between the sixth and tenth component/factor. For EFA, there was weak to moderate correlation with OpenIbis for most factors.

After considering all metrics in the unified validation framework, the candidate index which most strongly fulfilled the most criteria was PCA component 6 from $D_{\text{train/val}}^{(60)}$. These metrics were initially inverted, so the initial discovered index was subtracted from 100 to give the final index. The evaluation results of the final index on the train/validation set are given in Table 5.1.

This index appears to fulfil most of the criteria. It appears to be significantly predictive of pain category, although this association is weak and it becomes non-significant when added to a demographic model. There is essentially no correlation between OpenIbis and this index on an inter-patient basis. However, on an intra-patient basis, there is, on average, a weak positive correlation. The index has a modest rise of 5.474 points between the pre-surgical period and the surgical period. On average, it has a weak negative correlation with fentanyl effect-site concentration. The surrogate model finds that an increased index is associated with increased delta power, increased beta power, and reduced alpha power.

The majority of findings were stable when this index was evaluated on the test set, as is shown in Table 5.2. However, the weak ability for the index to predict postoperative pain score was no longer exhibited in this dataset. Figure 5.4 shows how this index behaves with three example EEG recordings from the test set.

The loadings that determined how this index is calculated are shown in Figure 5.5. For the purposes of readability, these loadings are inverted such that the red portions indicate frequencies associated with a higher nociception index, and the blue portions indicate frequencies associated with a lower nociception index. The anticipated pattern of delta arousal, alpha dropout, and

Table 5.1: Summary of train/validation set results for PCA component 6 from 60-second dataset.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.577 (0.524–0.629)
AUROC > 3	0.568 (0.513–0.625)
AUROC > 6	0.581 (0.505–0.656)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.004 ($p = 0.150$)
Δ AUROC > 3	−0.000 ($p = 0.311$)
Δ AUROC > 6	0.003 ($p = 0.281$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.017$ ($p = 0.715$), Spearman $\rho = -0.001$ ($p = 0.979$)
B2. Intra-patient corr.	Mean Pearson $r = 0.127$, Mean Spearman $\rho = 0.095$
C. Start of surgery	
C1. Mean comparison	Pre: 43.965, Surg: 49.439, Diff: 5.474
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.115$, Mean Spearman $\rho = -0.109$
E. Explainability	
E1. Delta coeff.	4.184 ($p < 0.001$)
E2. Alpha coeff.	−7.330 ($p < 0.001$)
E3. Beta coeff.	7.856 ($p < 0.001$)

Table 5.2: Summary of test set results for PCA component 6 from 60-second dataset.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.511 (0.405–0.618)
AUROC > 3	0.590 (0.479–0.700)
AUROC > 6	0.455 (0.318–0.598)
A2. Pain prediction (nested)	
Δ AUROC > 0	−0.012 ($p = 0.150$)
Δ AUROC > 3	0.004 ($p = 0.311$)
Δ AUROC > 6	−0.022 ($p = 0.281$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.083$ ($p = 0.381$), Spearman $\rho = -0.072$ ($p = 0.450$)
B2. Intra-patient corr.	Mean Pearson $r = 0.120$, Mean Spearman $\rho = 0.094$
C. Start of surgery	
C1. Mean comparison	Pre: 41.788, Surg: 50.928, Diff: 9.140
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.155$, Mean Spearman $\rho = -0.149$
E. Explainability	
E1. Delta coeff.	8.888 ($p < 0.001$)
E2. Alpha coeff.	−4.910 ($p < 0.001$)
E3. Beta coeff.	12.928 ($p < 0.001$)

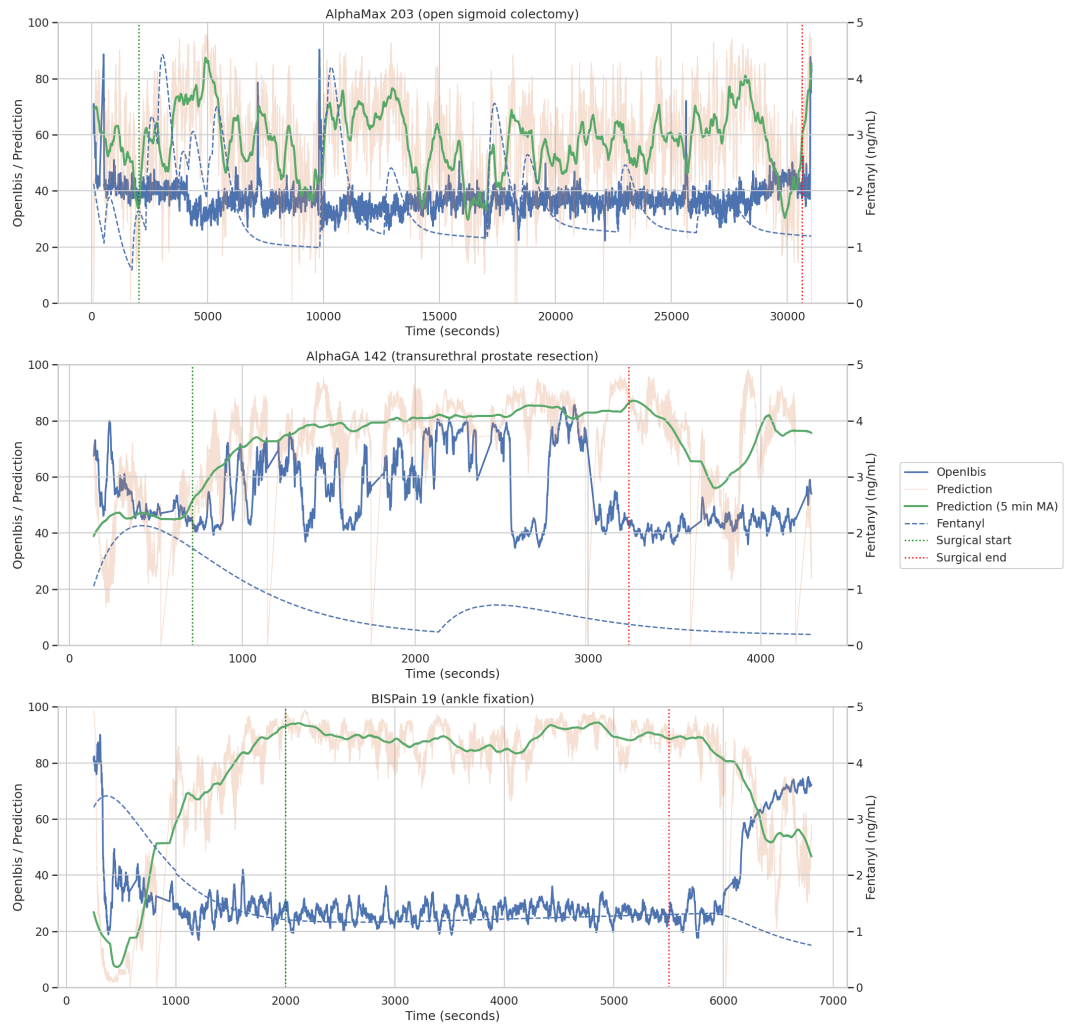


Figure 5.4: PCA component 6 applied to three example EEG recordings from the test dataset.

beta arousal is visible in the PCA loadings.

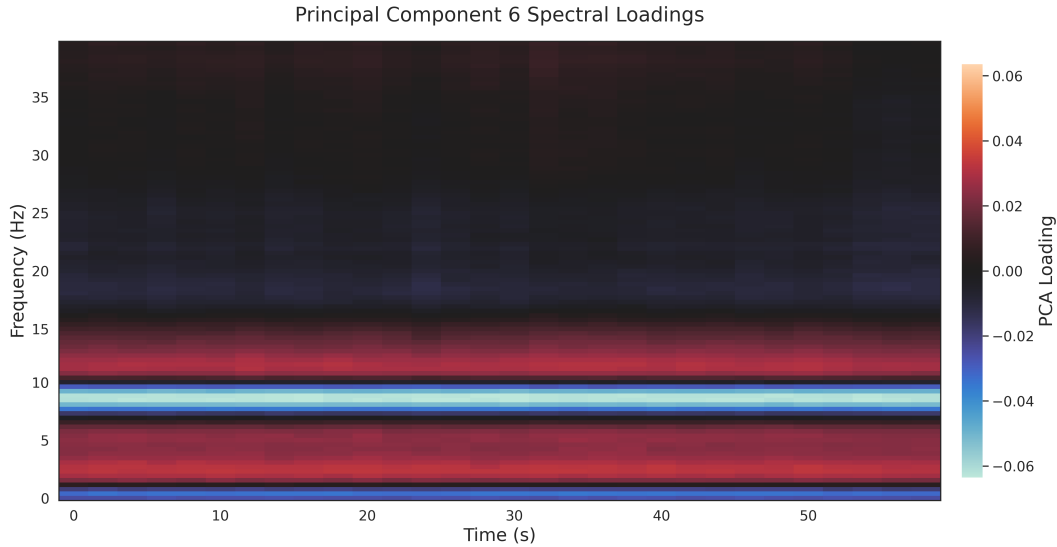


Figure 5.5: PCA component 6 loadings (inverted).

5.2.5 Discussion

This approach generated an index which may plausibly represent the component of nociception that is orthogonal to the hypnotic dimension. However, it did not consistently have useful predictive value for postoperative pain score. PCA is a highly interpretable and comparatively simple ML approach, although it is limited by the fact that it can only identify dimensions of variation that are orthogonal to each other. EFA can identify correlated factors, although in this case EFA was unable to identify any factor that outperformed any of the PCA components. It is possible that orthogonality of components was useful in this case. Hypnotic depth is a large confounding factor, and by extracting a factor that is presumably orthogonal to this, a large portion of the confounding may have been removed from the index. This is supported by the OpenIbis correlations: the first component typically has a strong OpenIbis correlation, and subsequent components have much weaker correlations. EFA, however, appears to have substantial hypnotic contamination throughout the majority of factors.

However, as was previously established, it is anticipated that some correlations between the hypnotic dimension and the nociceptive dimension do exist (i.e. the dimensions need not be entirely orthogonal). Some component of nociceptive response may be embedded in the hypnotic dimensions captured in early components of the PCA, and it is difficult to extricate this from the pure hypnotic EEG response.

The spectral loadings were consistent across the 60-second input period, with no changes in the relevant frequencies over time. This is approximately

equivalent to the input being a pure frequency domain representation, with frequencies averaged over a minute (it is possible that the 60-second input performed better than shorter input because the longer input was more robust to noise). More complex time-varying signatures of nociception were not observed when using this approach.

The frequency loading analysis in this case was broadly consistent with the literature. An increase in delta power (particularly high delta to low theta) was associated with an increase in the index. This is consistent with the ‘paradoxical’ delta arousal that has been observed in previous studies [12]. A clear band corresponding to a reduction in alpha power (particularly low alpha power) was observed as being associated with an increase in the nociception index. Finally, an increase in low beta power was consistent with an increase in the nociception index, with the relevance of frequencies above 16 Hz being low.

5.3 Hidden Markov Models and Microstates

5.3.1 Background

Formulation

A hidden Markov model (HMM) is an unsupervised learning method which infers a state sequence $Q = q_1, \dots, q_N$ from a time series of observations $X = x_1, \dots, x_N$. The states are chosen from a predefined discrete state space $S = \{s_1, \dots, s_K\}$. A transition matrix A is defined to describe the probability of a state s_i being succeeded by a state s_j , concretely $a_{ij} = P(q_n = s_j | q_{n-1} = s_i)$. These transitions are assumed to only depend on the previous state (i.e. satisfy the Markov property): $P(q_n | q_1, \dots, q_{n-1}) = P(q_n | q_{n-1})$. B represents the set of probability distributions mapping the current state $q_n = s_j$ to the current observation x_n . Where there is a discrete set of observations, B becomes an emission matrix. Where observations are continuous, $B = \{b_j(\cdot)\}_{j=1}^K$ is a set of continuous probability density functions describing the probability of observing an outcome x_n given a state s_j : $P(x_n | q_n = s_j)$. Observations are assumed to be conditionally independent given the state sequence. Finally, an initial distribution $\pi_i = P(q_1 = s_i)$ is defined for the initial state of the HMM [171].

Due to the Markov and independence assumptions, the full joint likelihood of observing X and Q given the key parameters $\lambda = (A, B, \pi)$ is:

$$P(X, Q | \lambda) = \pi_{q_1} \prod_{n=1}^N b_{q_n}(x_n) \prod_{n=2}^N a_{q_{n-1}, q_n}$$

There are several algorithms for estimating the state sequence Q and model parameters λ . Classically, the Baum–Welch expectation-maximisation algorithm is used, which utilises the Forward-Backward algorithm for computing

expected counts of state transitions [172]. The Gaussian-Linear HMM package in Python, used in this analysis, uses a stochastic variational inference training algorithm [173].

These algorithms output two discrete probability distributions at each time step, $\gamma_n(j)$ and $\xi_n(i, j)$. $\gamma_n(j)$ is the marginal posterior probability of s_j being occupied at time t , and is the basis for determining state occupancy. $\xi_n(i, j)$ is the joint posterior probability of a transition from s_i to s_j occurring at time n , and is the basis for determining a state transition matrix [172].

Single-Channel EEG

Fabuš (2023) validated a method for identifying microstates from low density electrode arrays using HMMs, demonstrating that spatiotemporal microstates can be recovered with reasonable accuracy even without full multiple-channel EEG. This work draws an equivalence between microstates and the state sequence Q from an HMM.

A dataset of 16 healthy volunteers undergoing propofol anaesthesia was used, and full 32-channel EEG recordings were taken. In order to estimate microstates, a time-delay embedded (TDE) HMM was used with 11 lags. Concretely, given an EEG sampled at 100 Hz, with measurements y_n at each time point, a feature vector $x_n = [y_{n-5}, y_{n-4}, \dots, y_n, \dots, y_{n+5}]$ is constructed encompassing a 100 ms window. In the 32-channel EEG case, this was reduced to a 24-dimensional embedding using PCA. The feature vectors were concatenated into a single input for HMM state estimation.

HMM state estimation was carried out using $K = 6$ states, which was empirically shown to provide the most reproducible states. Fractional occupancies and state transition matrices were then estimated. The HMM states were then estimated using single-channel EEG data using different pairs of electrodes, and the Spearman correlation between the fractional occupancies of the full 32-channel EEG data HMM states and the single-channel EEG states was calculated. High concordance was found, with the Spearman correlation ranging between 0.7 and 0.85 depending on the montage used. This suggests that spatiotemporal microstates can be reconstructed with relatively high accuracy from single-channel EEG data alone [166].

5.3.2 Methods

The methods employed by Fabuš (2023) were replicated in order to train an HMM. The train/validation dataset was used for training. Due to the large amount of EEG data in this study, a truncated version of the train/validation dataset was used, where a random contiguous 120-second sample of EEG was taken from each patient. These samples were time-embedded, concatenated, and used to train an HMM using the stochastic variational inference algorithm [173]. The trained HMM was then used to perform inference on the full

train/validation dataset, giving $\gamma_{p,n}(j)$ and $\xi_{p,n}(i, j)$ for each time point n , for each patient p in the train/validation dataset.

Initial statistics were calculated by summing $\gamma_{p,n}(j)$ and $\xi_{p,n}(i, j)$ across the entire dataset and then normalising by total time and outgoing transition counts respectively, generating estimates of overall fractional occupancy and the transition probability matrix:

$$\text{FO}_j = \frac{\sum_p \sum_n \gamma_{p,n}(j)}{\sum_p N_p}$$

and

$$\hat{P}_{ij} = \frac{\sum_p \sum_n \xi_{p,n}(i, j)}{\sum_{j'} \sum_p \sum_n \xi_{p,n}(i, j')}$$

where FO_j denotes the global fractional occupancy of state j , $\hat{P}_{ij}$ denotes the estimated transition probability from state i to state j , and N_p denotes the number of time points for participant p .

Statistics were then calculated on a per-second basis. We define a one-second window $\mathcal{W}_t = \{n : t \cdot f_s \leq n < (t+1) \cdot f_s\}$ at the t -th second, with the number of samples in the window being the sampling rate $f_s = 100$ Hz. The fractional occupancy $\text{FO}_{p,t,j}$ of state j over t for patient p is then given by:

$$\text{FO}_{p,t,j} = \frac{1}{f_s} \sum_{n \in \mathcal{W}_t} \gamma_{p,n}(j)$$

This gives six fractional occupancies for $K = 6$ states. The per-second transition matrix $\hat{P}_{p,t,ij}$ is similarly calculated as follows, giving 36 transition rates (including six self-transitions):

$$\hat{P}_{p,t,ij} = \frac{\sum_{n \in \mathcal{W}_t} \xi_{p,n}(i, j)}{\sum_{j'=1}^K \sum_{n \in \mathcal{W}_t} \xi_{p,n}(i, j')}$$

Another useful statistic calculated every second is the state switching rate $\text{SR}_{p,t}$. If $\Xi_{p,n} \in [0, 1]^{K \times K}$ is the full matrix of joint posterior probabilities for state transitions, the probability at a given time point that there would be a self-transition (and therefore no state switching) is given by $\text{tr}(\Xi_{p,n})$. The average state switching rate over the course of a second is therefore given by:

$$\text{SR}_{p,t} = \frac{1}{f_s} \sum_{n \in \mathcal{W}_t} (1 - \text{tr}(\Xi_{p,n}))$$

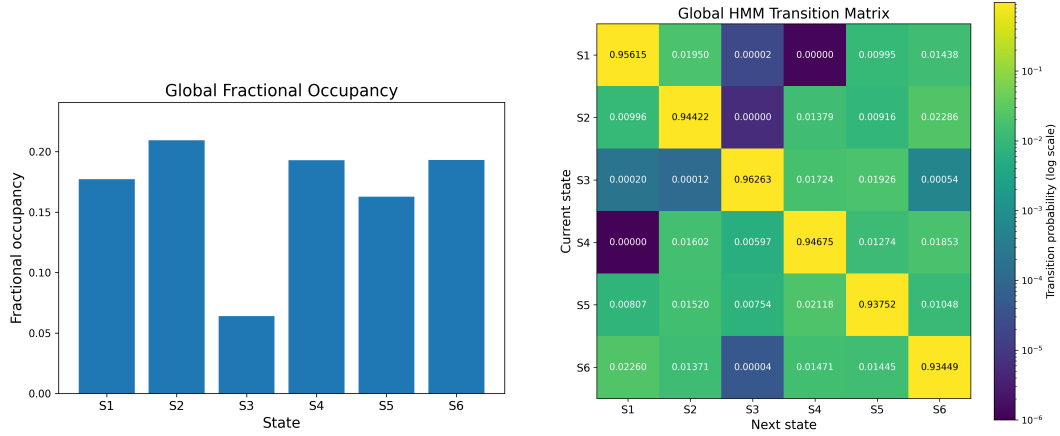
Finally, the state entropy $H_{p,t}$ is calculated, representing uncertainty in the transition distribution $\Xi_{p,n}$:

$$H_{p,t} = -\frac{1}{f_s} \sum_{n \in \mathcal{W}_t} \sum_{i=1}^K \sum_{j=1}^K \xi_{p,n}(i, j) \log_2 \xi_{p,n}(i, j)$$

This gives a total of 44 statistics that can be calculated from the HMM. These were considered as prospective indices of nociception, and each was transformed using the same monotonic CDF, with the train/validation set HMM output as a reference distribution. Each statistic was then validated on the train/validation set, using the previously defined validation criteria. The best model, based on overall performance across the validation criteria, was then chosen for evaluation on the test dataset. As with PCA/EFA, there is no prior calibration of sign, therefore the evaluation of each statistic was considered in both directions.

The HMM was used to infer $\gamma_{p,n}(j)$ and $\xi_{p,n}(i, j)$ values for the test dataset, and used for validation of the best statistic (including the monotonic CDF transform using the train/validation results as a reference distribution).

5.3.3 Results



(a) Global fractional occupancy per HMM state. (b) Global HMM state transition probabilities.

Figure 5.6: HMM state characterisation across all patients: fractional occupancy (left) and transition probability matrix (right).

Figure 5.6 gives the fractional occupancy and state transition results across the entire dataset. The most frequently occupied state was state 2, and the least frequently occupied state was state 3. There were very few transitions directly between state 1 and state 4.

The full train/validation set results are again given in Appendix A. There was no index which clearly fulfilled all evaluation criteria. Correlation with OpenIbis was weak to moderate throughout all HMM statistics. The best index was that which was derived from the fractional occupancy of state 5. This statistic was inverted and then transformed into a 0–100 metric, and the performance of this index on the train/validation set is shown in Table 5.3.

This index, at least on the train/validation set, begins to approach moderate predictive ability for postoperative pain score category. It adds significant

Table 5.3: Summary of train/validation set results for inverted fractional occupancy of state 5.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.656 (0.607–0.703)
AUROC > 3	0.643 (0.590–0.692)
AUROC > 6	0.588 (0.507–0.663)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.008 ($p = 0.022$)
Δ AUROC > 3	0.003 ($p = 0.198$)
Δ AUROC > 6	0.001 ($p = 0.891$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.698$ ($p < 0.001$), Spearman $\rho = -0.781$ ($p < 0.001$)
B2. Intra-patient corr.	Mean Pearson $r = -0.132$, Mean Spearman $\rho = -0.161$
C. Start of surgery	
C1. Mean comparison	Pre: 50.175, Surg: 53.595, Diff: 3.419
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = 0.029$, Mean Spearman $\rho = 0.036$
E. Explainability	
E1. Delta coeff.	1.921 ($p < 0.001$)
E2. Alpha coeff.	6.994 ($p < 0.001$)
E3. Beta coeff.	-0.043 ($p = 0.002$)

predictive ability for prediction of pain score > 0 ($p = 0.022$). However, it does not perform as expected on the other criteria. The rise between the pre-surgical period and the surgical period is modest (3.4 points), and the correlation with fentanyl effect-site concentration is very weakly positive.

The correlation with OpenIbis is reasonably weak on an intra-patient level ($r = -0.132$), but is strong on an inter-patient level ($r = -0.698$). The surrogate model suggests that higher scores with this index occur with delta arousal, alpha arousal, and beta dropout.

The results of evaluation on the test set are shown in Table 5.4. The results are broadly comparable to those observed on the train/validation set, aside from a reduction in the ability of the model to predict pain categories > 3 and > 6 . The performance of this index on the same three example patients is shown in Figure 5.7.

5.3.4 Discussion

The best index from the HMM approach did not fulfil the criteria well, besides being able to weakly predict postoperative pain score. The association with fentanyl effect-site concentration was not in the expected direction, the rise in the index at the start of surgery was low, and the correlation with OpenIbis was strong. Some of the poor performance may be attributed to the high variance of the signal, as is evident from Figure 5.7. As it is calculated from one-second windows only, it would be highly influenced by small amounts of variation.

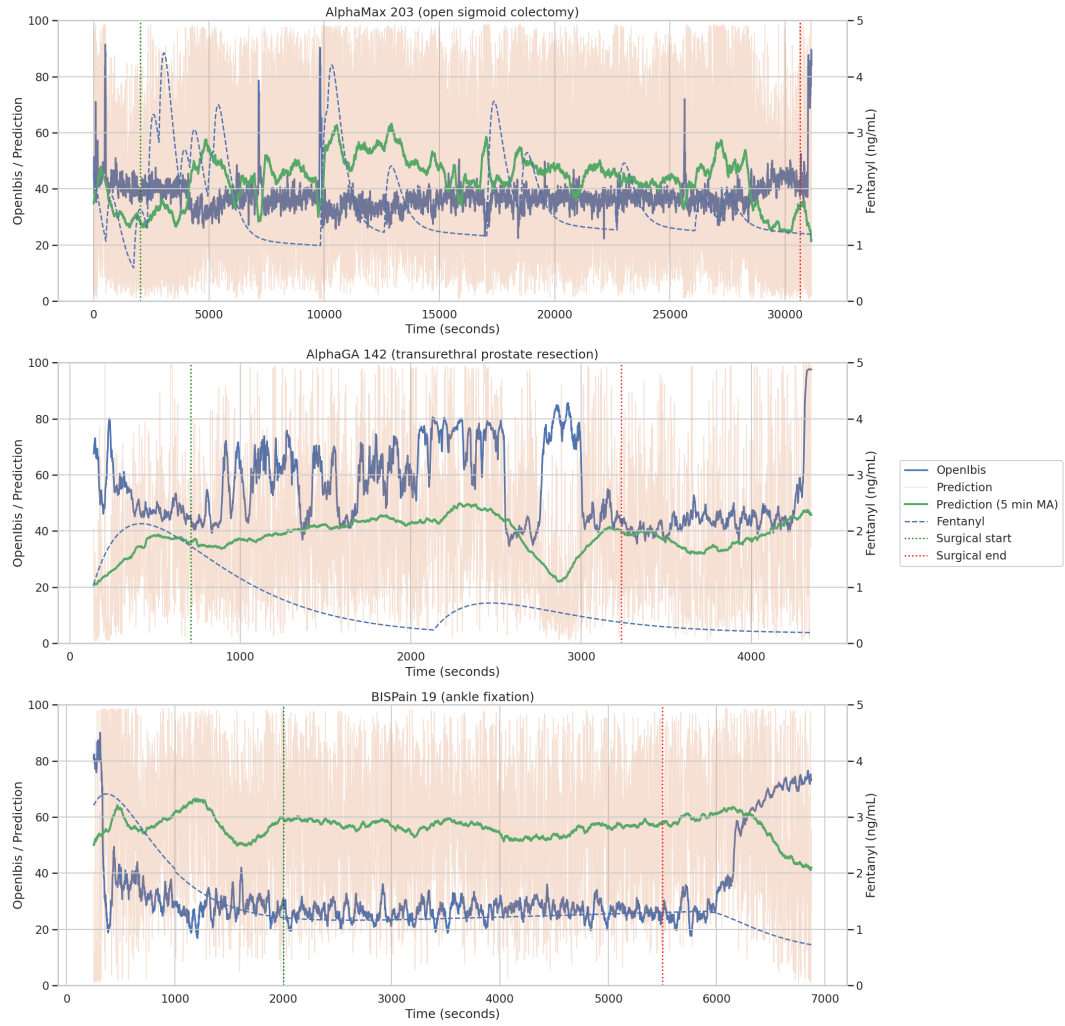


Figure 5.7: Inverted fractional occupancy of state 5 applied to three example EEG recordings from the test dataset.

Table 5.4: Summary of test set results for inverted fractional occupancy of state 5.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.635 (0.528–0.736)
AUROC > 3	0.523 (0.408–0.634)
AUROC > 6	0.548 (0.372–0.720)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.023 ($p = 0.022$)
Δ AUROC > 3	0.002 ($p = 0.198$)
Δ AUROC > 6	−0.004 ($p = 0.891$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.786$ ($p < 0.001$), Spearman $\rho = -0.785$ ($p < 0.001$)
B2. Intra-patient corr.	Mean Pearson $r = -0.182$, Mean Spearman $\rho = -0.187$
C. Start of surgery	
C1. Mean comparison	Pre: 54.761, Surg: 57.720, Diff: 2.959
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = 0.049$, Mean Spearman $\rho = 0.055$
E. Explainability	
E1. Delta coeff.	1.414 ($p < 0.001$)
E2. Alpha coeff.	6.134 ($p < 0.001$)
E3. Beta coeff.	−0.367 ($p < 0.001$)

Nevertheless, it cannot be concluded with confidence that this measure is a true measure of nociception.

Most of the indices evaluated as part of the HMM approach (in Table A.11 and Table A.12) had moderate to strong correlations with OpenIbis on an inter-patient level. Unlike PCA, which looks specifically for features orthogonal to the early components, this approach had no method of accounting for confounding from the hypnotic component. Much like the initial model generated from postoperative pain score prediction, this model therefore appeared to have substantial contamination from the hypnotic component, as it correlates negatively with OpenIbis.

Despite this, the model was able to achieve significant predictive ability for pain score at the > 0 threshold. This may be in part due to the hypnotic correlation itself: it has been demonstrated in previous chapters that deeper hypnotic levels of anaesthesia associates with higher postoperative pain scores in this dataset.

Achieving appropriate validation or explainability of the microstates identified as part of this analysis is not possible without full multiple-channel data. It is not possible to definitively determine how well the hidden states estimated by the HMM correlate with genuine spatiotemporal microstates. It is therefore unclear whether the lack of success of this method is due to poor ability to identify microstates, or if microstate fractional occupancies and transition dynamics are poor correlates for nociception. Future research utilising microstate analyses for nociception prediction should include multiple-channel data.

The lack of success of any single statistic to provide a useful index for nociception prediction does not mean that the HMM statistics do not provide value for nociception detection in general. They can reasonably be used as features for supervised ML methods. This is a direction that will be elaborated upon in subsequent chapters.

Chapter 6

EEG-Informed Pharmacodynamic Non-Linear Mixed Effects Model

6.1 Motivation

As an analgesic agent, it is anticipated that fentanyl will have a large impact on intraoperative nociception, either attenuating the signal or blocking it completely [35]. The effect of fentanyl can be estimated from the effect-site concentration using the sigmoid Hill equation, giving an indication of the level of antinociception [50]. This could plausibly be used as a label to train a supervised ML regression model from EEG, which could then be used to generate an index of nociception.

The first difficulty with this is that anti-nociception levels alone are insufficient to estimate a patient’s nociception status. Levels of surgical stimulus vary over the course of surgery, and the level of nociception that a patient experiences would be a function of present and past stimuli, as well as levels of analgesia. The second difficulty is that PD parameters for fentanyl are not constant across all patients. Some patients are more sensitive to fentanyl than others, even at identical effect-site concentrations [174]. A mixed-effect modelling approach is therefore taken, modelling both the level of surgical stimulus and the patient-level PD parameters as random effects. The non-linear sigmoid Hill equation is used to link modelled PD parameters and effect-site fentanyl concentration to antinociceptive effect. This is the basis of the EEG-informed pharmacodynamic non-linear mixed effects (EEG-PD-NLME) model, the novel model introduced in this chapter.

The other difficulty with the use of fentanyl to derive labels is the drug titration paradox, which can occur whenever studying a quantity that is actively titrated [58]. Fentanyl is titrated throughout surgery in response to anticipated or current surgical stimuli, changes in heart rate, blood pressure, respiratory rate [11], or EEG parameters (such as BIS or alpha power changes,

as in the AlphaMax study [149]). The decision was therefore made to restrict training to portions of EEG where effect-site concentration was rising following a bolus of fentanyl. These portions of EEG feature variation in the effect-site concentration of fentanyl without any deliberate titration by the anaesthetist, therefore avoiding the drug titration paradox.

6.2 Mixed-Effects Models in Pharmacology

PK and PD non-linear mixed effects (PK/PD-NLME) models are frequently used in pharmacological research in order to investigate the relationship between drug dosages and drug effects. They allow for multiple data points from multiple patients (i.e. hierarchical data) to be used to estimate PK and PD parameters. These parameters are estimated on both a population level, and on an individual level [175]. PK/PD-NLME models have applications across the drug development pipeline, including dose optimisation in clinical trials, exposure-response analysis, and in some cases, individualisation of dosing [176].

Concretely, given a non-linear pharmacological model $g(\cdot)$, a set of inputs of the system u_p (such as drug dosages and times for those dosages), and a set of pharmacological parameters θ_p , the drug response $y_{p,t}$ at time t for patient p is given by:

$$y_{p,t} = g(t, u_p, \theta_p) + \epsilon_{p,t}$$

The patient-level pharmacological parameters θ_p are generally composed of fixed effect parameters ϕ , which serve as a population average. Each patient has a set of random effects η_p , and θ_p is calculated as a combination of the fixed and random effects, using some function $h(\cdot)$ to combine them:

$$\theta_p = h(\phi, \eta_p)$$

The goal of PK/PD-NLME models is to take a model structure (defined by the functions g and h), and a dataset consisting of clinical response observations $y_{p,t}$, and drug dosage data u_p . From this, the population fixed effects ϕ are calculated, as well as individual random effects η_p .

Maximum likelihood estimation (MLE) would ideally be used to find ϕ , but this process is complicated by the non-linearity of the models. The population likelihood $P(y \mid \phi)$ is calculated by marginalising over the patient-level random effects:

$$P(y \mid \phi) = \prod_p \int P(y_p \mid \phi, \eta_p) P(\eta_p) d\eta_p$$

However, due to the non-linearity of g , this integral becomes intractable, hence no exact MLE solution can be found. Various approximations are typically employed in order to gain estimates of these parameters. There are several methods for this implemented in the NONMEM software package:

- First Order (FO) approximations linearise the model at $\eta = 0$, allowing for an estimate of population likelihood to be calculated.
- First Order Conditional Estimation (FOCE) linearises the model at individual-specific values of η .
- The Laplace method is similar to FOCE but uses a second-order (quadratic) approximation.
- Monte Carlo Expectation-Maximisation methods alternate between Monte Carlo simulations of η followed by updates to θ via MLE.
- Full Markov Chain Monte-Carlo Bayesian methods have been used, although they are computationally expensive [177].

6.3 EEG-PD-NLME Model

The EEG-PD-NLME model is inspired by PK/PD-NLME approaches, and combines them and conventional ML in a single pipeline. Figure 6.1 summarises the stages of the EEG-PD-NLME model.

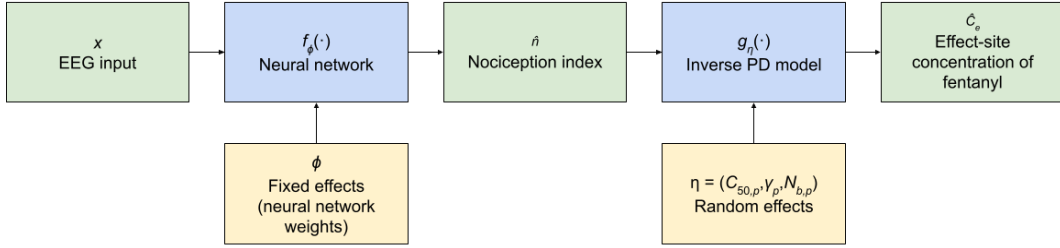


Figure 6.1: Overview of the EEG-PD-NLME model.

The model isolates the periods of time during surgery when the effect-site concentration of fentanyl is rising. It is assumed that any change in the nociceptive stimulus during these periods is small compared to the sharply rising change in fentanyl effect-site concentration, therefore nociceptive input is treated as a constant over the fentanyl bolus period. Rises in fentanyl effect-site concentration were chosen rather than falls as they tend to be steeper, and therefore any fluctuations in nociceptive stimulus are more likely to be negligible in comparison.

The following are defined as part of the EEG-PD-NLME:

- p identifies patients.
- b identifies the individual boluses of fentanyl. There may be multiple boluses per patient.

- i identifies individual EEG segments and fentanyl concentrations, sampled at one segment per second.
- x , the EEG input. The length and representation of the EEG input can vary. There are multiple EEG inputs per bolus b .
- $f_\phi(\cdot)$, the neural network. The goal of this network is to take the input x , and output a patient- and bolus-agnostic index for the level of nociception.
- ϕ , the fixed effects. In this case, the fixed effects are the parameters of the neural network f_ϕ (i.e. they are shared across all patients).
- $\hat{n} \in (0, 1)$ or $\in (-1, 1)$, the index of nociception. This is the output of f_ϕ , and is the ultimate quantity of interest for this model. It is the level of nociception that a patient is estimated to ‘perceive’ given both the input nociceptive stimulus and anti-nociception provided by any analgesia. It is an estimate of the true n .
- $g_\eta(\cdot)$, the non-linear inverse PD model. In this case, the model that is used is the Hill equation.
- η , the random effects that define the inverse PD model. There are three factors that make up the random effects:
 - $C_{50,p} \in \mathbb{R}^+$, the effect-site concentration of fentanyl required to generate an effect half as large as the maximal effect. Every patient has their own value for $C_{50,p}$ which is the same for every bolus b .
 - $\gamma_p \in \mathbb{R}^+$, which determines the steepness of the sigmoid Hill curve. Again, it is unique for each patient and is the same across boluses.
 - $N_b \in (0, 1)$, which measures the level of nociceptive stimulus. It represents the level of nociception that a patient would experience in the absence of fentanyl. It is assumed to be constant over the course of a bolus b . A patient with multiple boluses of fentanyl will have different values of N_b for each bolus to represent the changing levels of nociception over the course of a surgery.
- $C_e \in \mathbb{R}_{\geq 0}$, the effect-site concentration of fentanyl. This is the label which is used for regression. The output of g_η is $\hat{C}_e$, the predicted effect-site concentration.

The end goal of this model is to train f_ϕ , such that it can take an EEG input and output $\hat{n}$, an estimate of the level of nociception. This is achieved by training the entire EEG-PD-NLME pipeline end-to-end. EEGs x are inputs into the composite model $g_\eta(f_\phi(\cdot))$, which outputs a prediction of the concentration $\hat{C}_e$. This is used to calculate a loss function $\mathcal{L}(\eta, \phi) = \frac{1}{N} \sum_{i=1}^N \ell(C_{e,i}, \hat{C}_{e,i})$, which is then used to optimise ϕ and η via gradient descent.

For each bolus b , multiple EEG inputs are provided such that the model has data that shows how the EEG x changes as C_e rises, whilst the other factors in η , including the nociceptive stimulus N_b , are assumed to remain constant. In order to fit a changing C_e , the fixed effect network f_ϕ will need to output a useful index. This allows $\hat{n}$ to be identifiable and separable from the random effects in η .

6.3.1 Mathematical Formulation

The basis of g_η is the sigmoid Hill equation. A drug effect E is determined by:

$$E = \frac{E_{max}C_e^\gamma}{C_{50}^\gamma + C_e^\gamma}$$

This drug effect acts to reduce the baseline nociceptive stimulus N_b , giving n . There are two ways this reduction can be mathematically modelled: as a proportional reduction in nociception, or as an absolute reduction in nociception. There is no known definitively correct way to model opioid effect, so both models will be explored.

Proportional Model

Drug effect is scaled such that $E \in [0, 1]$, giving $E_{max} = 1$. This drug effect is considered as the proportion by which N_b is reduced to give n :

$$n = N_b(1 - E) = N_b \left(1 - \frac{C_e^{\gamma_p}}{C_{50,p}^{\gamma_p} + C_e^{\gamma_p}} \right) = \frac{N_b C_{50,p}^{\gamma_p}}{C_{50,p}^{\gamma_p} + C_e^{\gamma_p}} \quad (6.1)$$

According to this model, n gets asymptotically close to 0 with increasing C_e , but never reaches it. When $C_e = C_{50,p}$, the level of nociception is reduced by half. When $C_e = 0$, $n = N_b$. Because N_b ranges between 0 and 1, $n \in (0, 1)$ in this model.

This equation provides a mapping from C_e to n . The EEG-PD-NLME generates $\hat{n}$, and calculates $\hat{C}_e$ using g_η . In this case, g_η can be determined by inverting the previous equation.

$$\hat{C}_e = g_\eta(\hat{n}) = C_{50,p} \left(\frac{N_b - \hat{n}}{\hat{n}} \right)^{\frac{1}{\gamma_p}}$$

Given the neural network f_ϕ parameterised by weights ϕ , and an input EEG segment x , the complete EEG-PD-NLME model is therefore given by:

$$\hat{C}_e = g_\eta(\hat{n}) = g_\eta(f_\phi(x)) = C_{50,p} \left(\frac{N_b - f_\phi(x)}{f_\phi(x)} \right)^{\frac{1}{\gamma_p}} \quad (6.2)$$

This model is fully differentiable and therefore the parameters are learnable via gradient descent. Note that $\hat{C}_e$ is only defined when $N_b \geq f_\phi(x)$. This is a constraint that will need to be enforced during training.

It should also be noted that, in this model, there is a singularity when $\hat{n} = 0$. If the output of $f_\phi(x)$ is calculated by a sigmoid function, this singularity is avoided. However, there is a risk of numeric instability as $\hat{n}$ approaches 0.

Absolute Model

Again, drug effect is scaled such that $E \in [0, 1)$ and $E_{max} = 1$. However, in this model, this drug effect represents an absolute reduction in the level of nociception.

$$n = N_b - E = N_b - \frac{C_e^{\gamma_p}}{C_{50,p}^{\gamma_p} + C_e^{\gamma_p}} \quad (6.3)$$

In this case, if $E > N_b$, n becomes negative, with $n \in (-1, 1)$. One could attempt to place a ReLU around the function, causing it to be positive only. However, this causes difficulties when the function is inverted. g_η loses injectivity, as all C_e large enough to cause $E \geq N_b$ will map to $\hat{n} = 0$.

It is noted that fentanyl has a direct effect on EEG, even in the absence of nociceptive stimulus [116]. In this model, the direct effect of fentanyl on EEG, after generating a complete blockade of any nociceptive stimulus, could be interpreted as negative n .

Inverting this as previously gives:

$$\hat{C}_e = g_\eta(\hat{n}) = C_{50,p} \left(\frac{N_b - \hat{n}}{1 - N_b + \hat{n}} \right)^{\frac{1}{\gamma_p}}$$

When the neural network f_ϕ is substituted in, this gives a final complete EEG-PD-NLME model:

$$\hat{C}_e = g_\eta(\hat{n}) = g_\eta(f_\phi(x)) = C_{50,p} \left(\frac{N_b - f_\phi(x)}{1 - N_b + f_\phi(x)} \right)^{\frac{1}{\gamma_p}} \quad (6.4)$$

Again, this is fully differentiable. There are two domain constraints for the model to remain well-defined. Firstly, as with the proportional model, $N_b \geq f_\phi(x)$. Secondly, $f_\phi(x) \in (-1, 1)$ creates the possibility that $f_\phi(x)$ falls below $N_b - 1$. Another constraint must therefore be imposed: $N_b - 1 < f_\phi(x)$.

Again, there is a risk of a singularity and numeric instability, this time when $f_\phi(x)$ approaches $N_b - 1$.

Random Effects

In a standard PK/PD-NLME, the PK/PD parameters are modelled as a combination of population averages ϕ (the fixed effects) and patient-specific factors η_p (the random effect). The patient-specific factors are assumed to be normally distributed: $\eta_p \sim \mathcal{N}(0, \Omega)$, where Ω is the covariance matrix of the random effects.

However, in this model, the fixed effects ϕ have already been defined as the parameters of f_ϕ . Given that the aim of this model is to determine these parameters rather than population PK/PD parameters, the fixed-effect component of the PD model was replaced with literature-derived constants. These values were $C_{50,\text{pop}} = 1.0$ and $\gamma_{\text{pop}} = 2.5$, based on estimates from the literature [178, 179]. $N_{b,\text{pop}} = 0.75$ was selected to allow for substantial variation in n below N_b , whilst retaining flexibility for bolus-specific variation in N_b .

The raw values $\tilde{C}_{50,p}$, $\tilde{\gamma}_p$, and $\tilde{N}_b$ are the unconstrained parameters learned as part of the EEG-PD-NLME model, and they are combined with the population averages to give the values used in the model. The softplus function is used to ensure positivity of $C_{50,p}$ and γ_p , and the sigmoid function $\sigma(\cdot)$ is used to constrain N_b between 0 and 1.

Given fixed constants $\tilde{C}_{50,\text{pop}}$, where $\text{softplus}(\tilde{C}_{50,\text{pop}}) = C_{50,\text{pop}}$, and $\tilde{\gamma}_{\text{pop}}$, where $\text{softplus}(\tilde{\gamma}_{\text{pop}}) = \gamma_{\text{pop}}$, the patient-level parameters are estimated as:

$$\begin{aligned} C_{50,p} &= \text{softplus}(\tilde{C}_{50,\text{pop}} + \tilde{C}_{50,p}) \\ \gamma_p &= \text{softplus}(\tilde{\gamma}_{\text{pop}} + \tilde{\gamma}_p) \end{aligned}$$

Likewise, given $\tilde{N}_{b,\text{pop}}$, where $\sigma(\tilde{N}_{b,\text{pop}}) = N_{b,\text{pop}}$, the constrained value of N_b is calculated as:

$$N_b = \sigma(\tilde{N}_{b,\text{pop}} + \tilde{N}_b)$$

As previously discussed, a full Bayesian hierarchical model is intractable due to the non-linear PD model. However, Gaussian priors (in the softplus and sigmoid space) can effectively still be placed over the model by applying an l_2 weight decay over the unconstrained parameters $\tilde{\eta} = (\tilde{C}_{50,p}, \tilde{\gamma}_p, \tilde{N}_b)$. This is not an exact Gaussian prior in a full Bayesian inference sense, but it can nevertheless be used to constrain the random effects.

Network Output Scaling

The output of the fixed effect network f_ϕ is scaled, with the scaling being different depending on whether the proportional or absolute model is used. For a given EEG segment x , the network initially outputs an unconstrained output $\tilde{f}_\phi(x)$.

In the proportional model, $n \in (0, 1)$. Consequently, $\tilde{f}_\phi(x)$ is scaled between 0 and 1 using the sigmoid function:

$$\hat{n} = f_\phi(x) = \sigma(\tilde{f}_\phi(x))$$

In the absolute model, $n \in (-1, 1)$. Scaling is therefore between -1 and 1 using the hyperbolic tangent function:

$$\hat{n} = f_\phi(x) = \tanh(\tilde{f}_\phi(x))$$

All the transforms of the network output, as well as those of the random effects, are differentiable, allowing for gradient descent to be used.

6.3.2 Loss Functions

Fentanyl Concentration MSE

Given that the labels are the effect-site fentanyl concentrations, the mean squared error (MSE) for prediction of C_e is a straightforward loss function to use. If $\hat{C}_{e,i}$ is the predicted effect-site concentration associated with a given EEG instance x_i (with associated random effects η), the per-instance loss is:

$$\ell(C_{e,i}, \hat{C}_{e,i}) = (C_{e,i} - \hat{C}_{e,i})^2$$

For the total dataset of size N (encompassing all data points i across all boluses and all patients), the loss function is:

$$\mathcal{L}(\eta, \phi) = \frac{1}{N} \sum_{i=1}^N (C_{e,i} - \hat{C}_{e,i})^2$$

An MSE loss function is equivalent to maximum likelihood estimation (MLE) with normally distributed residuals:

$$\hat{C}_{e,i} + \epsilon_i = C_{e,i}, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

The residuals are unlikely to be normally distributed in practice, given the PD transform that the output of f_ϕ undergoes.

Proportional Model MLE Loss

Rather than treating the residuals of $\hat{C}_{e,i}$ as normally distributed, these loss functions instead treat the residuals of $\hat{n}$ as normally distributed:

$$\hat{n}_i + \epsilon_i = n_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

A loss function is derived by calculating the negative log-likelihood (NLL) of this distribution and using a change of variables. The full derivation of this loss function and its variants is in Appendix B. The NLL calculation depends on how the parameter σ is set. In this study, three options were explored:

- Assume the error in $\hat{n}_i$ will be homoscedastic, and set a specific value of σ for this.
- Assume the error in $\hat{n}_i$ will be homoscedastic, but calculate an optimal value for it using the summed negative log-likelihood.
- Assume the error in $\hat{n}_i$ will be heteroscedastic but proportional to the value of $\hat{n}_i$.

If a specific value of σ is chosen, the loss function becomes:

$$\mathcal{L}(\eta, \phi) = \sum_{i=1}^N \left[\frac{(g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2}{2\sigma^2} - \log(N_{b(i)}) - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

σ in this case acts as a tunable parameter that determines the balance between fitting the fixed effects ϕ and the random effects η . When σ is high, fitting the random effects is more strongly weighted. When σ is small, fitting the fixed effect network f_{ϕ} is more strongly weighted.

If the optimal σ is calculated, the loss function becomes:

$$\mathcal{L}(\eta, \phi) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{i=1}^N (g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2 \right) + \sum_{i=1}^N \left[-\log(N_{b(i)}) - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

If the loss is heteroscedastic but proportional to $\hat{n}_i$, the loss function becomes:

$$\mathcal{L}(\eta, \phi) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{i=1}^N \frac{(n_i - f_{\phi}(x_i))^2}{n_i^2} \right) + \sum_{i=1}^N \left[\log(n_i) - \log N_{b(i)} - \gamma_{p(i)} \log C_{50,p(i)} \right. \\ \left. - \log \gamma_{p(i)} - (\gamma_{p(i)} - 1) \log C_{e,i} + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

Absolute Model MLE Loss

A similar process can be followed using the absolute model, with derivation again in Appendix B. As n can become negative, the NLL cannot be calculated for the heteroscedastic case. However, for the other two cases, the loss functions can be calculated.

If σ is arbitrarily chosen, the loss function is:

$$\mathcal{L}(\eta, \phi) = \sum_{i=1}^N \left[\frac{(g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2}{2\sigma^2} - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

If σ is chosen optimally, the loss function is:

$$\mathcal{L}(\eta, \phi) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{i=1}^N (g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2 \right) + \sum_{i=1}^N \left[-\gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

6.3.3 Practical Considerations

For ease of implementation, the inverse PD model g_{η} was calculated in the log domain. Mathematically, the calculation is equivalent, but the log domain allows for the EEG-PD-NLME formula to be calculated primarily using additions, subtractions, and multiplications rather than repeated, computationally expensive power operations. The proportional model, for instance, has the form:

$$\log \hat{C}_e = \log C_{50,p} + \frac{1}{\gamma_p} \left[\log(N_b - f_{\phi}(x)) - \log(f_{\phi}(x)) \right]$$

The output of this is then exponentiated to get the final prediction for $\hat{C}_e$. Similarly, the absolute model has the following form in the log domain:

$$\log \hat{C}_e = \log C_{50,p} + \frac{1}{\gamma_p} \left[\log(N_b - f_{\phi}(x)) - \log(1 - N_b + f_{\phi}(x)) \right]$$

As was previously mentioned, there is a domain constraint where $N_b \geq f_{\phi}(x)$, which is slightly strengthened in the log domain form to $N_b > f_{\phi}(x)$ (as $N_b = f_{\phi}(x)$ would cause a logarithm to be taken of 0). In order to avoid this, $N_b - f_{\phi}(x)$ is clamped to a minimum value of $\varepsilon = 1 \times 10^{-6}$, and a barrier term B_1 is added to the loss function. Given an EEG segment x_i and incoming nociceptive stimulus N_b , the barrier term is:

$$B_1(x_i, N_b) = \frac{1}{100} \text{softplus}(100(f_{\phi}(x_i) - N_b))$$

The absolute model has the additional domain constraint of $N_b - 1 < f_{\phi}(x)$. This is similarly avoided by clamping the denominator term $1 - N_b + f_{\phi}(x)$ to a minimum value of $\varepsilon = 1 \times 10^{-6}$. A second barrier term B_2 is added to the loss function:

$$B_2(x_i, N_b) = \frac{1}{100} \text{softplus}(100(N_b - f_{\phi}(x_i) - 1))$$

Additionally, both models have a risk of instability due to denominator singularities. For the purposes of maintaining model stability, the maximum effect-site concentration of fentanyl that could be predicted by the model was

clamped to 15 ng/mL. Typical intraoperative fentanyl concentrations are substantially below this, and allowing prediction above this will propagate instability.

6.4 Methods

6.4.1 Dataset

The EEG-PD-NLME model used a supervised learning regression paradigm, requiring associated pairs of EEG segments x_i and fentanyl effect-site concentrations $C_{e,i}$. The dataset was limited only to time periods where the effect-site concentration of fentanyl was rising in association with a bolus. These segments required a rise of at least 0.01 ng/mL per second to be included, and they had to be at least 30 seconds long. They had a maximum length of 1000 seconds (so that any segments of continuously rising fentanyl that were longer than this had to be divided into separate segments). Each contiguous segment was given a unique ID which could then be associated with a unique N_b value in the model.

Any EEG segments that began before the first incision were excluded, as many of these segments would lack nociceptive stimulus and therefore would not contribute anything meaningful to the model. Any segments which contained missing data were also excluded.

The data points $(x_i, C_{e,i})$ were sampled at a rate of one sample per second of EEG data. When the length of the EEG segments w was greater than one second, this led to substantial overlap in EEG segments between data points. This served as a data augmentation technique that increased the effective size of the dataset.

The EEG segments x_i were sampled in the time period immediately following the $C_{e,i}$ measurement, satisfying the correct direction of causality for this scenario (one would anticipate any EEG manifestations of a given fentanyl effect-site concentration to occur after that concentration was achieved, rather than before). Concretely, given a C_e sampled at time t , the corresponding EEG segment would be $x_{[t:t+w]}$.

6.4.2 Per-Epoch Validation

After being trained on the train set, the model was validated on the validation set each epoch. However, because the original split was at the patient level (in order to avoid data leakage), the random effects η_{train} were not relevant to the validation set. Test-time adaptation was therefore performed in order to learn an optimal set of random effects for the validation set η_{val} . These random effects were used in the full pipeline with the previously trained f_ϕ in order to calculate estimated effect-site concentrations C_e . This was used to calculate

validation metrics, with the goal of evaluating whether the fixed effect network f_ϕ was capable of generalising outside of its training data.

Given the validation dataset D_{val} , the relevant n_i values were calculated using $f_\phi(x_i)$ for each x_i in the dataset. The relevant η_{val} values were then obtained by taking each $(n_i, C_{e,i})$ pair and fitting g_η to this data, with $\eta_{\text{val}} = \{C_{50,p}, \gamma_p, N_b\}$ as learnable parameters. A limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) optimiser was used with 20 iterations in order to quickly fit these parameters at the end of each epoch of training.

Due to this test-time adaptation, the validation metrics in early epochs would usually appear much better than the metrics from the training data. While the training was jointly optimising η and ϕ (which would not necessarily have converged at early epochs), the test-time adaptation process began with a fixed ϕ and quickly found an optimal η . This allowed for the model to perform substantially better on the validation set than the train set. Validation metrics would, however, be internally consistent, allowing for comparison of the fit of f_ϕ as training progressed.

In addition to the loss function, key metrics that were tracked over the course of training included mean squared error (MSE) and mean absolute error (MAE). The mean Spearman ρ between the f_ϕ output and C_e values across all boluses B was also calculated (where $\mathcal{I}_b$ was the set of data points associated with bolus b):

$$\bar{\rho} = \frac{1}{B} \sum_{b=1}^B \rho(\{f_\phi(x_i)\}_{i \in \mathcal{I}_b}, \{C_{e,i}\}_{i \in \mathcal{I}_b})$$

The Spearman ρ was chosen because it can serve as a measure of the fit of f_ϕ which functions independently of test-time adaptation and the random effects. Spearman ρ is a rank-order correlation, which was ideal for the non-linear relationship between $f_\phi(x)$ and C_e . It should be noted that a more negative correlation is indicative of better model fit: we anticipate that higher C_e would be associated with a lower nociception index.

6.4.3 Training

Model training was performed with a batch size of 64, with shuffled batches and an Adam optimiser with a default learning rate of 1×10^{-3} . The previously mentioned metrics (MSE, MAE, and $\bar{\rho}$) were calculated for the train set as well as the validation set. Training proceeded for a maximum of 200 epochs, with early stopping and restoration of the best state. As it was the metric that best represented the fit of the fixed parameters ϕ without being influenced by the random effects η , the $\bar{\rho}$ of the validation set was chosen as the metric for early stopping. It was also chosen as the primary metric for model selection and hyperparameter tuning.

As this metric can be quite noisy, a patience of 20 epochs was specified for early stopping. Additionally, an exponential moving average was used to

smooth the tracked $\bar{\rho}$. Given a validation $\bar{\rho}_e$ at epoch e , the smoothed metric is given by:

$$\begin{aligned}\bar{\rho}_{\text{smoothed},e} &= \alpha \bar{\rho}_e + (1 - \alpha) \bar{\rho}_{\text{smoothed},e-1}, \\ \bar{\rho}_{\text{smoothed},1} &= \bar{\rho}_1\end{aligned}$$

α was a tunable parameter that determined the degree of smoothing. For the purposes of these experiments, α was set to 0.2.

6.5 Experiments

Several experiments were carried out using the EEG-PD-NLME model to determine how altering inputs, model structures, and hyperparameters would influence the outcomes. The aim of this section is to develop an intuition for how these factors influence the outcomes, rather than finding an optimal combination of hyperparameters (which would require a more thorough exploration of the hyperparameter space).

6.5.1 Experiment 1: Initial Model

Description

An initial experiment was carried out in order to provide a baseline for future experiments. For this experiment, the input was ten seconds of EEG data from the standard cleaning dataset. A time-frequency representation of the EEG was generated from this EEG data, using a STFT. The hop length was one second, and frequencies were limited to 40 Hz. Both magnitude information and phase information (represented as a sine-cosine embedding) were included as part of the input. This gave an input of size ten seconds, with 103 frequency bins, and three channels.

The f_ϕ model used was a lightweight spectrogram attention network featuring:

- Normalisation of the magnitude channel over time and frequency.
- A small MLP of shape $[64, 64, 64]$ for each time-frequency bin.
- Trainable positional embeddings for time and frequency.
- An attention logit for each time-frequency bin.
- Pooling of the attention weights.
- A final regression head of shape $[64, 64, 1]$.

The proportional model was used for g_η , and the loss function was MSE for fentanyl effect-site concentration. A learning rate of 1×10^{-3} was used to train this model.

The training and validation curves for the loss function (MSE) and mean Spearman $\bar{\rho}$ were saved and plotted. All boluses in the validation set were ranked by their Pearson r value, comparing $\hat{C}_e$ to C_e . The best, median, and worst boluses were identified and the predicted and actual fentanyl curves were plotted. Plots of the random effects were also generated.

Results

There were 287 boluses in the train set from 136 different patients (of the total 396 patients from the train set). This model trained for 12 epochs. The train loss / MSE was 0.446, and the Spearman $\bar{\rho}$ was -0.030 . The validation loss / MSE was 0.386, and the Spearman $\bar{\rho}$ was -0.081 . The $\bar{\rho}$ in both these cases indicates only a very weak negative average correlation between $f_\phi(x)$ and C_e . The mean $C_{50,p}$ was 0.821, with a standard deviation of 0.266. The mean γ_p was 2.769, with a standard deviation of 0.424. The mean N_b was 0.650, with a standard deviation of 0.149. Training curves and prediction examples for this model are shown in Figure 6.2, and the random effect distribution is shown in Figure C.1.

6.5.2 Experiment 2: PD Model and Loss Function

Description

The hyperparameters as described in the previous section were kept the same, but the model used for g_η was varied between the proportional and absolute models. The loss function was also varied between each of the previously discussed loss functions:

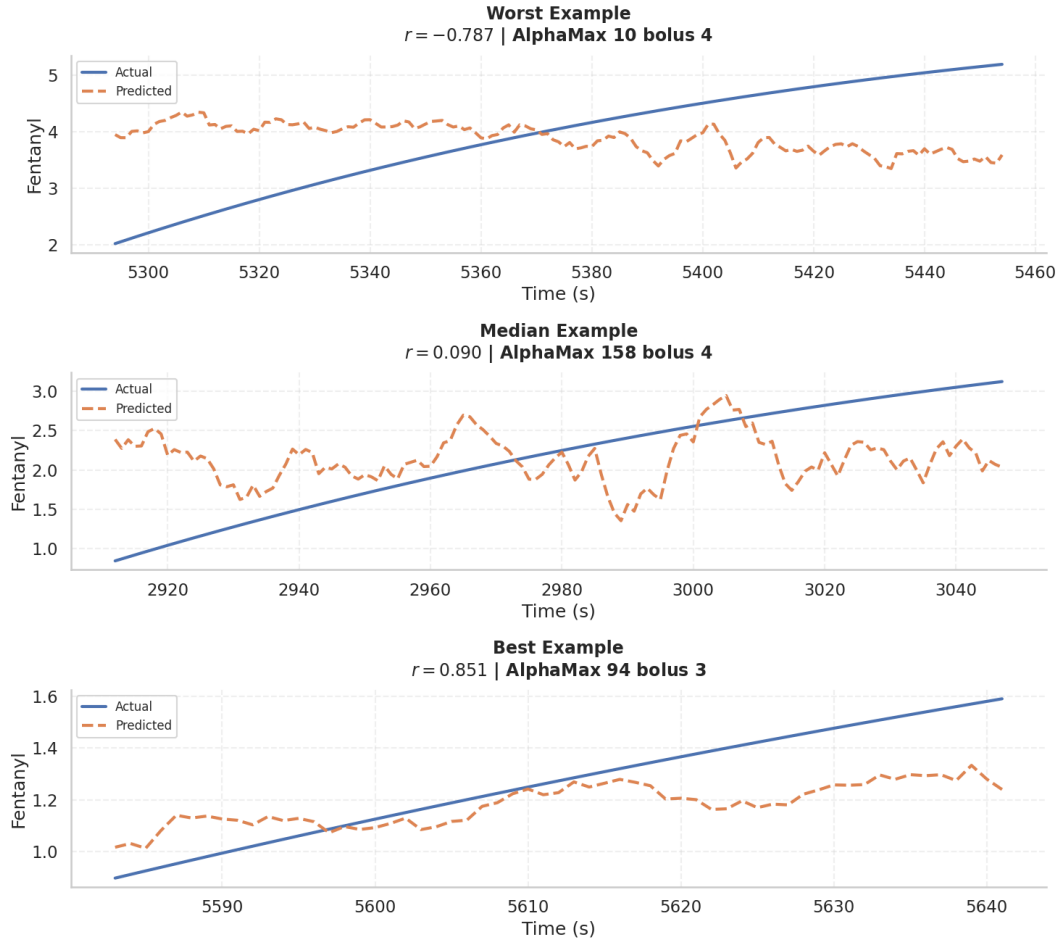
- MSE for fentanyl effect-site concentration
- The derived NLL loss with the calculated optimal σ .
- The derived NLL loss with the variable (heteroscedastic) σ (for the proportional model only).
- The derived NLL loss, with varying values of $\sigma \in \{0.001, 0.01, 0.1, 1\}$.

Results

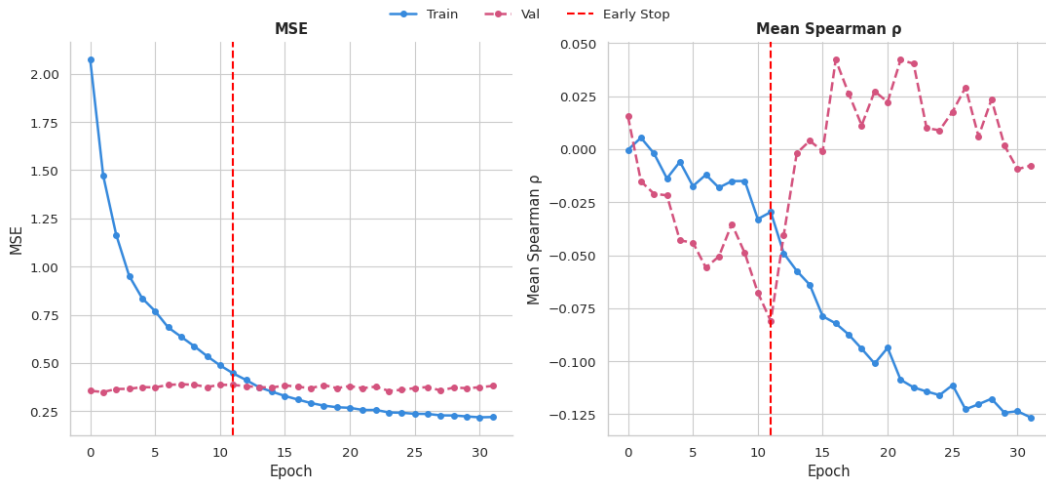
The results for this experiment are shown in Table 6.1. The loss values used throughout this table cannot be compared, as each row used a different loss function. $\bar{\rho}$ and MSE were metrics that can be compared across loss functions.

All loss functions appeared to produce similarly weak negative mean Spearman correlations $\bar{\rho}$, indicating a weak fit for f_ϕ . Very small values of σ were

Bolus-Level Pearson Correlation Examples



(a) Predicted vs. actual $\hat{C}_e$ and C_e fentanyl curves for the best, median, and worst boluses from validation set ranked by Pearson r .



(b) Training and validation curves for MSE loss and mean Spearman $\bar{\rho}$ across epochs.

Figure 6.2: Bolus prediction examples and training curves from Experiment 1.

Table 6.1: Experiment 2 results for the proportional and absolute models across loss functions.

Loss Function	Train			Validation			Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	Loss	MSE	$\bar{\rho}$	Loss	MSE	N_b	C_{50}	γ
<i>Proportional Model</i>									
Fentanyl MSE	-0.0296	0.4463	0.4463	-0.0811	0.3849	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
NLL, optimal σ	-0.0974	-0.4496	1.0503	-0.0245	-0.7537	0.4471	0.764 ± 0.092	1.113 ± 0.358	2.323 ± 0.454
NLL, heteroscedastic σ	-0.0036	0.0479	7.4769	-0.0232	-0.9050	1.9280	0.727 ± 0.012	0.913 ± 0.048	2.623 ± 0.080
NLL, $\sigma = 0.001$	-0.2256	529.1502	0.3188	-0.0643	183.7173	10.3698	0.644 ± 0.146	0.797 ± 0.329	1.266 ± 0.407
NLL, $\sigma = 0.01$	-0.2255	8.0814	0.3157	-0.0895	5.8111	3.7951	0.647 ± 0.147	0.809 ± 0.335	1.283 ± 0.411
NLL, $\sigma = 0.1$	-0.0745	2.4341	1.1625	-0.0437	2.1215	0.3923	0.756 ± 0.081	1.051 ± 0.314	2.207 ± 0.357
NLL, $\sigma = 1$	-0.1157	1.3257	1.5007	-0.0178	1.0034	0.8600	0.896 ± 0.016	1.393 ± 0.461	3.209 ± 0.625
<i>Absolute Model</i>									
Fentanyl MSE	-0.0770	0.2382	0.2381	-0.0873	0.3401	0.3401	0.682 ± 0.087	0.809 ± 0.225	2.753 ± 0.322
NLL, optimal σ	-0.0309	-0.4180	1.0645	-0.0575	-0.8259	0.3498	0.742 ± 0.074	1.083 ± 0.277	2.321 ± 0.326
NLL, $\sigma = 0.001$	-0.0299	10378.9297	1.4566	-0.0387	252.9782	21.0067	0.749 ± 0.053	0.941 ± 0.190	2.138 ± 0.157
NLL, $\sigma = 0.01$	-0.0358	105.5906	1.4569	-0.0494	6.2749	8.6694	0.749 ± 0.053	0.942 ± 0.191	2.139 ± 0.157
NLL, $\sigma = 0.1$	-0.0362	2.7806	1.4068	-0.0422	1.9979	0.3467	0.748 ± 0.053	1.000 ± 0.211	2.216 ± 0.199
NLL, $\sigma = 1$	-0.1998	1.1081	0.5448	-0.0678	0.9687	0.4624	0.708 ± 0.136	1.514 ± 0.619	3.666 ± 0.749

associated with very poor model fit overall according to MSE: these models strongly prioritised fitting the fixed effects ϕ whilst neglecting the random effects η , hence leading to overall poor model fit. The heteroscedastic variant of the loss performed relatively poorly. The fentanyl MSE was the loss function that appeared to perform best on both the proportional model and absolute model, both by MSE (overall fit) and by $\bar{\rho}$. Performance was comparable between the proportional model and the absolute model.

6.5.3 Experiment 3: Input Type

Description

In this experiment, the type of input was varied, with a different suitable f_ϕ architecture for each type of input. The remainder of the hyperparameters were kept the same as the baseline experiment. The types of input that were used were:

- A constant input. This meant the f_ϕ output was also constant, and the fit of the model was entirely determined by g_η . For the purposes of developing an index of nociception, this input did not have any value, but it served as a valuable baseline comparison. It was only used on the validation set.
- Time domain input (i.e. EEG signal). This input was passed through a one-dimensional CNN. The CNN consisted of four convolution blocks with 32, 64, 128, and 256 output channels respectively, with kernel sizes of 7, 5, 3, and 3. Each convolution layer was followed by a batch normalisation layer and a GELU activation, with max pooling between the blocks. This was followed by an adaptive average pooling layer, and an MLP of shape [128, 64] with GELU and dropout layers of 0.3 between each layer.

- Frequency domain input, using a power spectrum as input (calculated by aggregating the STFT across the time dimension, and taking the square of the magnitude). The network used here was two MLPs in series. The first had layer sizes [256, 128, 64], outputting a 64-dimensional representation. Each of the layers in this MLP had batch normalisation and a GELU activation following it. The second MLP [128, 64] used ReLU activations.
- STFT input, using magnitude only. This used a very similar spectral attention network to the baseline experiment, except there was only one input channel and hence no aggregation across channels.
- STFT input, using magnitude and phase. This was identical to the baseline experiment.
- SingLEM embeddings. An MLP was trained with hidden layers [128, 64], with dropout layers of 0.3 and ReLU activations. The 16-dimensional embeddings were combined across ten seconds to give 160 input features.
- HMM results. For each second of EEG data, the HMM fractional occupancies γ and transition probabilities ξ were concatenated into a single 42-dimensional representation. Across ten seconds, this gave 420 input dimensions, which were passed through the same [128, 64] MLP that was used for SingLEM.

Results

The results of this experiment are given in Table 6.2. The constant baseline outperformed all other models in terms of MSE, with an MSE of 0.336. This suggested that the signal for training f_ϕ was very weak.

Table 6.2: Experiment 3 results across input types.

Input Type	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
Constant	—	—	—	0.3362	—	—	—
Time domain (1D CNN)	-0.3805	0.1413	-0.0838	0.3780	0.638 ± 0.195	0.893 ± 0.320	2.733 ± 0.447
PSD (MLP)	-0.0232	1.7137	-0.0048	0.4987	0.743 ± 0.014	0.974 ± 0.057	2.548 ± 0.082
STFT (magnitude and phase)	-0.0296	0.4463	-0.0811	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
STFT (magnitude only)	-0.0928	0.2841	-0.0795	0.3776	0.598 ± 0.226	0.828 ± 0.296	2.736 ± 0.523
SingLEM (MLP)	-0.1137	0.6492	-0.0670	0.4265	0.707 ± 0.065	0.878 ± 0.186	2.695 ± 0.275
HMM (MLP)	-0.3031	0.1429	-0.0452	0.3421	0.595 ± 0.237	0.848 ± 0.307	2.758 ± 0.561

Results were otherwise broadly comparable across each of the input types, with similarly weak negative correlations for $\bar{\rho}$ across each of the categories. The pure power spectrum input did perform noticeably worse than the other inputs.

6.5.4 Experiment 4: Model Type

Description

In this experiment, all hyperparameters were fixed as in the initial experiment (including an STFT input with phase preserved), but the model used for f_ϕ was varied. The models that were used were:

- The lightweight spectral attention model that was used in the baseline experiment.
- An MLP that flattened the time-frequency input, and used it in a simple MLP with hidden layers of size [256, 128], with dropout layers of 0.2 and ReLU activations between layers.
- A small CNN with three layers of 32, 64, and 128 channels, with GELU activations, stride-2 downsampling, global average pooling, and a linear regression head.
- A ResNet-18 backbone using ImageNet weights, with a modified input layer for spectrogram inputs, and a linear regression head.
- A hybrid CNN-transformer architecture. Inputs were standardised over channels, and then features were extracted using a convolutional block with channel counts [32, 64], with ReLU activations and group normalisation. This output was projected into an embedding space of dimension 128, after which the features went through a frequency attention mechanism. Scalar attention scores were computed for each frequency bin, and then normalised. Temporal dependencies were then captured using a transformer encoder with multi-head self-attention. Mean pooling was then used over the time dimension, and the remaining features passed through a linear regression head.

Results

The results of this experiment are given in Table 6.3.

Table 6.3: Experiment 4 results across f_ϕ architectures.

Architecture	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
Spectral attention	-0.0296	0.4463	-0.0811	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
Flat MLP	-0.0011	2.3205	-0.0597	0.4397	0.742 ± 0.018	0.969 ± 0.069	2.550 ± 0.093
Small CNN	-0.0204	1.9884	-0.0688	0.3475	0.739 ± 0.020	0.959 ± 0.074	2.565 ± 0.107
ResNet-18	-0.4010	0.1145	-0.0742	0.4646	0.648 ± 0.175	0.839 ± 0.278	2.815 ± 0.437
CNN-transformer	-0.0124	0.3445	-0.3547	0.3534	0.589 ± 0.229	0.811 ± 0.296	2.700 ± 0.531

The CNN-transformer performed better on the validation set than the other networks, but the train set $\bar{\rho}$ indicated close to no mean correlation between

$f_\phi(x)$ and C_e . The networks otherwise all performed comparably, with $\bar{\rho}$ values indicating very weak mean correlations.

6.5.5 Experiment 5: Dataset Type

Description

This experiment used the same hyperparameter setup as the initial experiment, but it compared the baseline results (which used the standard cleaning dataset) against the results when the EEGdenoiseNet-cleaned dataset was used.

Results

Table 6.4: Experiment 5 results comparing standard and EEGdenoiseNet-cleaned datasets.

Dataset	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
Standard cleaning	-0.0296	0.4463	-0.0811	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
EEGdenoiseNet cleaning	-0.0215	0.7520	-0.0269	0.3830	0.693 ± 0.082	0.848 ± 0.212	2.746 ± 0.321

The results, shown in Table 6.4, indicated that the standard cleaning dataset appeared to do slightly better across most metrics, but the $\bar{\rho}$ values overall still indicated only a weak correlation.

6.5.6 Experiment 6: Input Length

Description

In this experiment, the length of the EEG segments that were being used as inputs to f_ϕ was varied. The ResNet-18 architecture, previously described above, was used here due to the ability of large CNNs to extract features from large images such as spectrograms.

Five different segment lengths were used: one second, ten seconds, 30 seconds, 60 seconds, and 120 seconds. The sampling rate remained at one sample per second, meaning that the longer EEG inputs had more overlap between consecutive segments.

Results

Table 6.5 shows the results of this experiment. Again, the $\bar{\rho}$ correlations were weak, with the 60-second validation model even yielding a positive validation $\bar{\rho}$ (i.e. in the opposite direction from what was expected). Despite clearly overfitting on the train data, the best performing model in terms of $\bar{\rho}$ was the 120-second model.

Table 6.5: Experiment 6 results across EEG segment lengths, using a ResNet-18 f_ϕ architecture.

Segment Length	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
1 second	-0.0282	2.0872	-0.0375	0.4885	0.740 ± 0.019	0.959 ± 0.070	2.564 ± 0.099
10 seconds	-0.4010	0.1145	-0.0742	0.4646	0.648 ± 0.175	0.839 ± 0.278	2.815 ± 0.437
30 seconds	-0.3772	0.2154	-0.0942	0.7678	0.735 ± 0.039	0.943 ± 0.121	2.599 ± 0.137
60 seconds	-0.7823	0.0406	0.0351	0.4988	0.733 ± 0.043	0.911 ± 0.135	2.673 ± 0.156
120 seconds	-0.9808	0.0009	-0.1567	0.5563	0.730 ± 0.062	0.881 ± 0.173	2.745 ± 0.212

6.5.7 Experiment 7: Learning Rate

Description

This experiment was conducted in order to determine what the effect would be of varying the relative learning rates between the fixed effects ϕ and the random effects η . The learning rate for the fixed effects was set to 1×10^{-3} , as in the baseline model. Then, the learning rate for the random effects was varied. The values that were tested were 1×10^{-2} , 1×10^{-3} (identical to the baseline model), and 1×10^{-4} .

The other hyperparameters in this experiment were identical to the baseline experiment.

Results

The results for this experiment, shown in Table 6.6, show the effect of varying the learning rate for the random effects.

Table 6.6: Experiment 7 results across learning rate configurations for fixed effects (ϕ) and random effects (η), denoted $\text{lr}_\phi/\text{lr}_\eta$.

$\text{lr}_\phi / \text{lr}_\eta$	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
$10^{-3} / 10^{-2}$	-0.0620	0.3016	0.0170	0.3502	0.594 ± 0.253	0.966 ± 0.405	2.942 ± 0.558
$10^{-3} / 10^{-3}$	-0.0296	0.4463	-0.0811	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
$10^{-3} / 10^{-4}$	-0.1043	1.0058	-0.0614	0.5796	0.734 ± 0.027	0.942 ± 0.098	2.594 ± 0.139

When the learning rate for the random effects was much larger than the learning rate for the fixed effects, the random effects failed to fit meaningfully, with $\bar{\rho}$ being positive. The overall model, however, achieved the best MSE on both train and validation sets, consistent with the previous finding that random effect fit was the most important factor for the model fit overall. When the learning rate for the random effects was much lower than the learning rate for the fixed effects, the overall model fit was the poorest, but the performance in terms of $\bar{\rho}$ was comparable to the baseline model.

6.5.8 Experiment 8: Weight Decay

Description

The baseline model was set with no weight decay, therefore allowing the random effects to vary without constraint. In this experiment, a weight decay of two different strengths (1×10^{-2} and 1×10^{-1}) was applied to the random effects to determine if any meaningful improvement resulted. An additional experiment was carried out to determine if adding a weight decay to the fixed effect network f_ϕ generated any improvement.

Results

Table 6.7: Experiment 8 results across weight decay configurations applied to the random effects η and fixed effects ϕ .

Weight Decay	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
None	-0.0296	0.4463	-0.0811	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
η : 10^{-2}	-0.0506	0.8665	-0.0530	0.4895	0.749 ± 0.015	0.965 ± 0.182	2.492 ± 0.191
η : 10^{-1}	-0.0587	1.6065	-0.0510	0.5352	0.750 ± 0.002	1.000 ± 0.047	2.500 ± 0.040
ϕ : 10^{-4}	-0.0192	0.4441	-0.0765	0.3757	0.640 ± 0.162	0.817 ± 0.272	2.766 ± 0.442

The results in Table 6.7 demonstrate that a weight decay on the random effects appeared to degrade the overall performance of the model, particularly in terms of MSE. This appeared to be because a strong weight decay highly constrains the random effects towards their baseline values, as can be seen from the listed random effects, particularly for the strongest weight decay of 10^{-1} . A strong weight decay on the random effects means the random effects are not able to move to the values which best fit the data.

The weight decay on ϕ did not appear to create any substantial difference to the performance of the overall network.

6.5.9 Experiment 9: Baseline EEG Comparison

Description

In this experiment, for each data point, an additional segment of EEG $\tilde{x}$ was selected from the pre-surgical period. It was randomly selected from between the point 180 seconds after intubation, and the first surgical incision, and it was used as an additional input into the fixed effects model (i.e. the fixed effect model becomes $f_\phi(x, \tilde{x})$). This period gave the model an example of the patient's EEG in the absence of nociception. There was substantial inter-patient variation in EEG [101], and features that were associated with nociception could have also been associated with baseline patient factors (for instance, low alpha power could be explained by both nociception and the patient being

older). Access to a baseline EEG with no nociception may have allowed the model to differentiate whether an EEG feature is patient-specific or associated with nociception.

The experiments were as follows:

- Baseline model $f_\phi(x)$ with the same architecture and hyperparameters as the initial model, and no addition of $\tilde{x}$.
- Concatenation model. The baseline model f_ϕ was modified to remove the final [64, 64, 1] regression head, giving $\tilde{f}_\phi$. This was used as an encoder, which was used to generate 128-dimensional embeddings of x and $\tilde{x}$. These embeddings were concatenated, and used as inputs to the final regression head h . Concretely:

$$f_\phi(x, \tilde{x}) = h([\tilde{f}_\phi(x), \tilde{f}_\phi(\tilde{x})])$$

- Cross-attention model. This model was similar to the concatenation model, but instead performed cross-attention between the embeddings:

$$f_\phi(x, \tilde{x}) = h(\text{crossattn}(\tilde{f}_\phi(x), \tilde{f}_\phi(\tilde{x})))$$

- Subtractive model. This model took the difference between the encoding of x and $\tilde{x}$:

$$f_\phi(x, \tilde{x}) = h(\tilde{f}_\phi(x) - \tilde{f}_\phi(\tilde{x}))$$

Results

The results for this experiment are shown in Table 6.8. In terms of $\bar{\rho}$, there was no model where $\tilde{x}$ appeared to improve the results beyond the baseline. The subtractive model seemed particularly poor, with a positive $\bar{\rho}$.

Table 6.8: Experiment 9 results across baseline EEG integration strategies.

Architecture	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
None (baseline)	-0.0296	0.4463	-0.0811	0.3849	0.650 ± 0.149	0.821 ± 0.266	2.769 ± 0.424
Concatenation	-0.1840	0.1775	-0.0679	0.3543	0.583 ± 0.245	0.836 ± 0.309	2.765 ± 0.555
Cross-attention	-0.0306	0.6137	-0.0287	0.3806	0.677 ± 0.108	0.834 ± 0.237	2.757 ± 0.365
Subtractive	-0.0078	2.6421	0.0137	0.3540	0.747 ± 0.017	0.990 ± 0.082	2.582 ± 0.109

6.5.10 Final Evaluation

Description

The strongest performing model on the validation set in terms of $\bar{\rho}$, the 120-second variant of the ResNet-18 model, was more closely analysed. As with the

baseline model, training and validation curves were recorded and plotted, the best, median, and worst boluses were plotted, and distributions of the random effects were plotted. Then, the model was evaluated on the test set using the previously established set of evaluation criteria. These criteria specifically evaluate f_ϕ as an index of nociception, and are separate from any requirement to fit the random effects. They evaluate the performance of the model on every data point in the test set, not just the data points associated with fentanyl boluses.

As with previous models, the output of f_ϕ underwent a monotonic CDF transform to give a percentile from 0–100. The distribution of data points used for this transform was the entire combined train and validation dataset.

An explainability analysis was then conducted using Gradient-weighted Class Activation Mapping (GradCAM) [180]. Activations were taken from the final block of ResNet-18.

Finally, the random effects η were analysed, with violin plots being made to plot each of N_b , $C_{50,p}$ and γ_p against pain score category. Analysis of variance (ANOVA) was conducted in order to determine if there was any difference in these random effects between patients with different pain score categories.

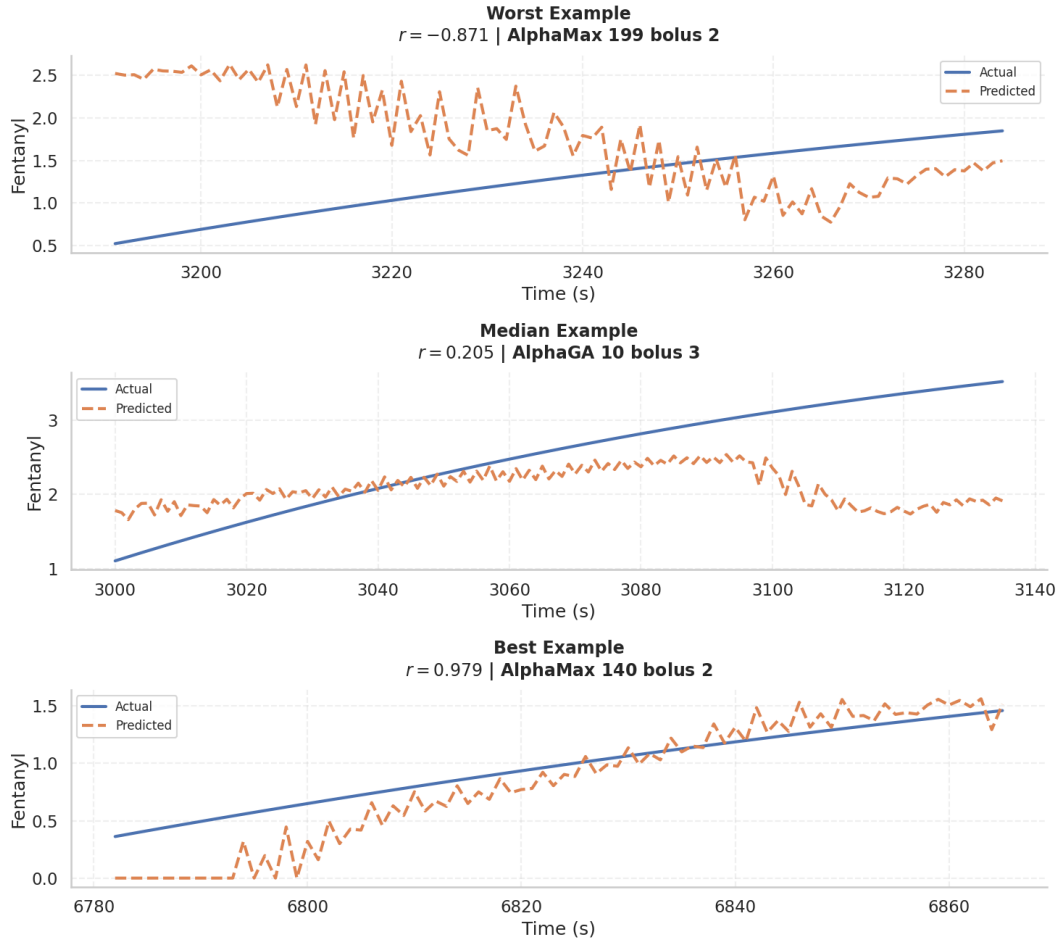
Results

The training and validation results for the 120-second ResNet-18 model are shown in Figure 6.3, with random effect distributions in Figure C.2. It appears as though a double-descent-like phenomenon may have occurred during training. The model has overfit on the training data, and this is initially associated with $\bar{\rho}$ becoming less negative. However, with further overfitting, this becomes more negative again. The MSE loss remains roughly constant during this period.

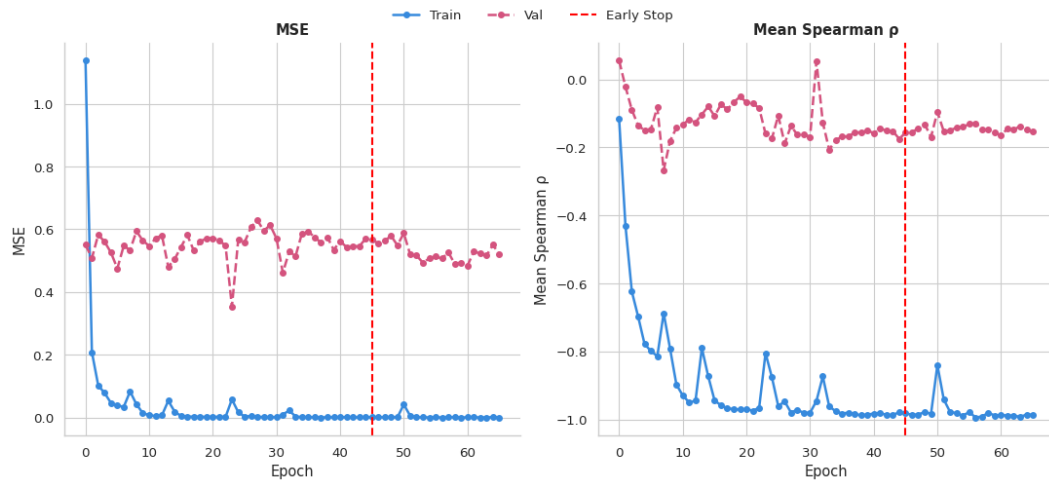
Table 6.9 summarises the performance of the model on the test set, and Figure 6.4 gives some example plots for this model on full test examples. The overall results were mixed. There appeared to be essentially no predictive ability of this model for postoperative pain score, violating criteria A1 and A2. OpenIbis correlation was weakly positive on an intra-patient level, but became moderately strong on an inter-patient level ($r = 0.519$). The index appeared to fall when surgery begins, rather than rise. The fentanyl bolus correlation was a weak negative correlation, as expected, although the correlation was weaker than the performance of this model on the validation data may have suggested ($r = -0.032$ and $\rho = -0.063$). This model showed the expected pattern with alpha and beta powers, as it demonstrated alpha depression and beta arousal with a higher pain score according to the surrogate model used in Category E. However, rather than showing the ‘paradoxical’ delta arousal that would be expected, this model appeared to associate increased nociception with delta depression.

Figure 6.5 gives three examples of 120-second inputs to this model, alongside the corresponding saliency maps generated by GradCAM. As would be

Bolus-Level Pearson Correlation Examples



(a) Predicted vs. actual $\hat{C}_e$ and C_e fentanyl curves for the best, median, and worst boluses from validation set ranked by Pearson r .



(b) Training and validation curves for MSE loss and mean Spearman $\bar{\rho}$ across epochs.

Figure 6.3: Bolus prediction examples and training curves from 120-second ResNet-18 model.

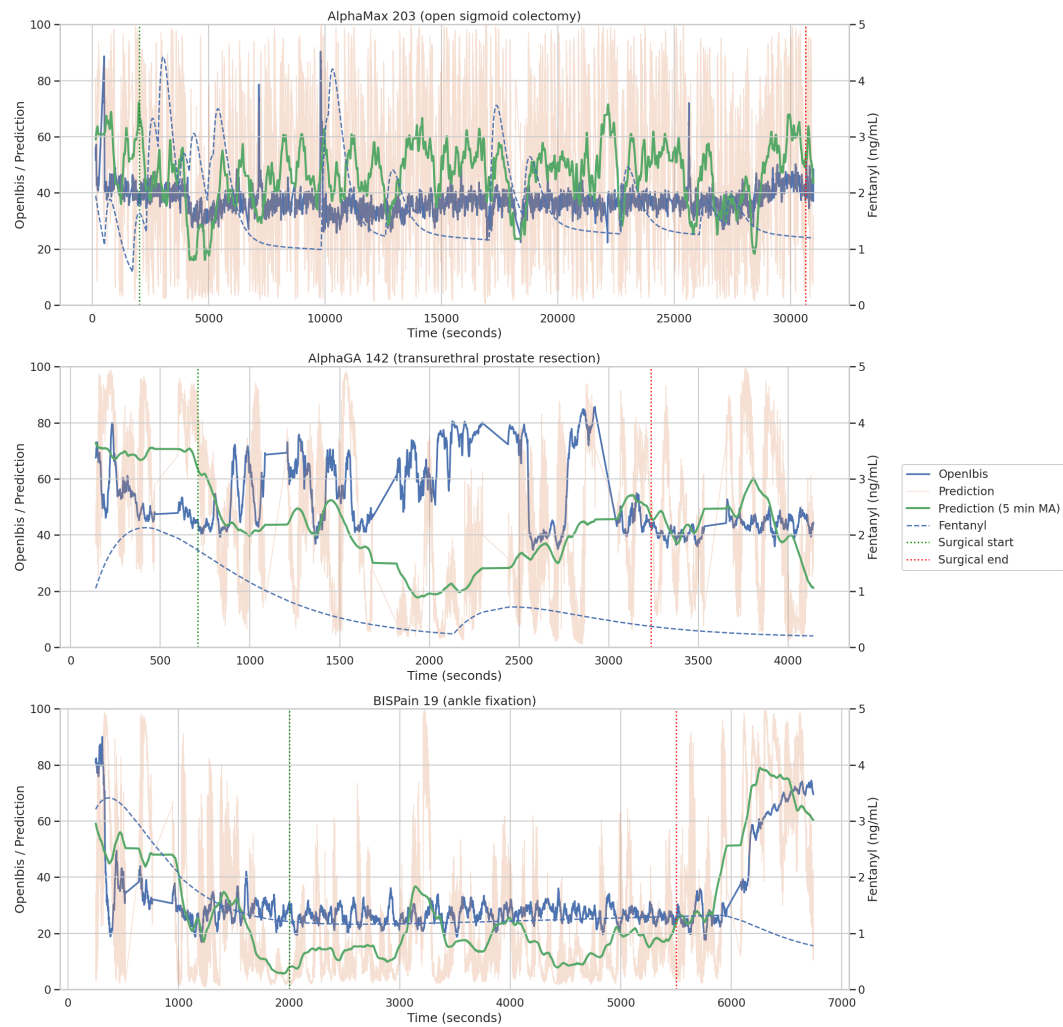


Figure 6.4: ResNet-18 model applied to three example EEG recordings from the test dataset.

Table 6.9: Summary of test set results for EEG-PD-NLME with ResNet-18 f_ϕ and 120-second input.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.462 (0.357–0.569)
AUROC > 3	0.452 (0.345–0.567)
AUROC > 6	0.525 (0.371–0.686)
A2. Pain prediction (nested)	
Δ AUROC > 0	−0.001 ($p = 0.728$)
Δ AUROC > 3	0.005 ($p = 0.505$)
Δ AUROC > 6	−0.008 ($p = 0.075$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = 0.519$ ($p < 0.001$), Spearman $\rho = 0.526$ ($p < 0.001$)
B2. Intra-patient corr.	Mean Pearson $r = 0.055$, Mean Spearman $\rho = 0.060$
C. Start of surgery	
C1. Mean comparison	Pre: 56.929, Surg: 50.284, Diff: −6.645
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.032$, Mean Spearman $\rho = -0.063$
E. Explainability	
E1. Delta coeff.	−14.430 ($p < 0.001$)
E2. Alpha coeff.	−8.401 ($p < 0.001$)
E3. Beta coeff.	0.853 ($p < 0.001$)

anticipated, the most salient portions of the spectrograms are the first 20 seconds, although regions of reasonable saliency extend all the way to 100 seconds in one of the plots. The most salient frequencies appeared to be those between 5–10 Hz, with all frequencies below 20 Hz showing reasonable saliency.

Finally, Figure 6.6 shows the distribution of random effects across each of the four pain score categories (none, mild, moderate, and severe). ANOVA found there was no significant difference in mean value between these four groups in any of the three random effect parameters.

6.6 Pre-Surgical Data

The previous EEG-PD-NLME model was entirely reliant on C_e of fentanyl, which does create a risk that it could be modelling direct fentanyl effect rather than nociception. The rise in the index at the start of surgery is inconsistent with a true index of nociception (criterion C1). A variant of the EEG-PD-NLME model was therefore developed which could use pre-surgical data to address these issues.

6.6.1 Model Adjustments

The model used was identical to the ResNet-18 model with an input length of 120 seconds, but with an additional training loop that ran in parallel to

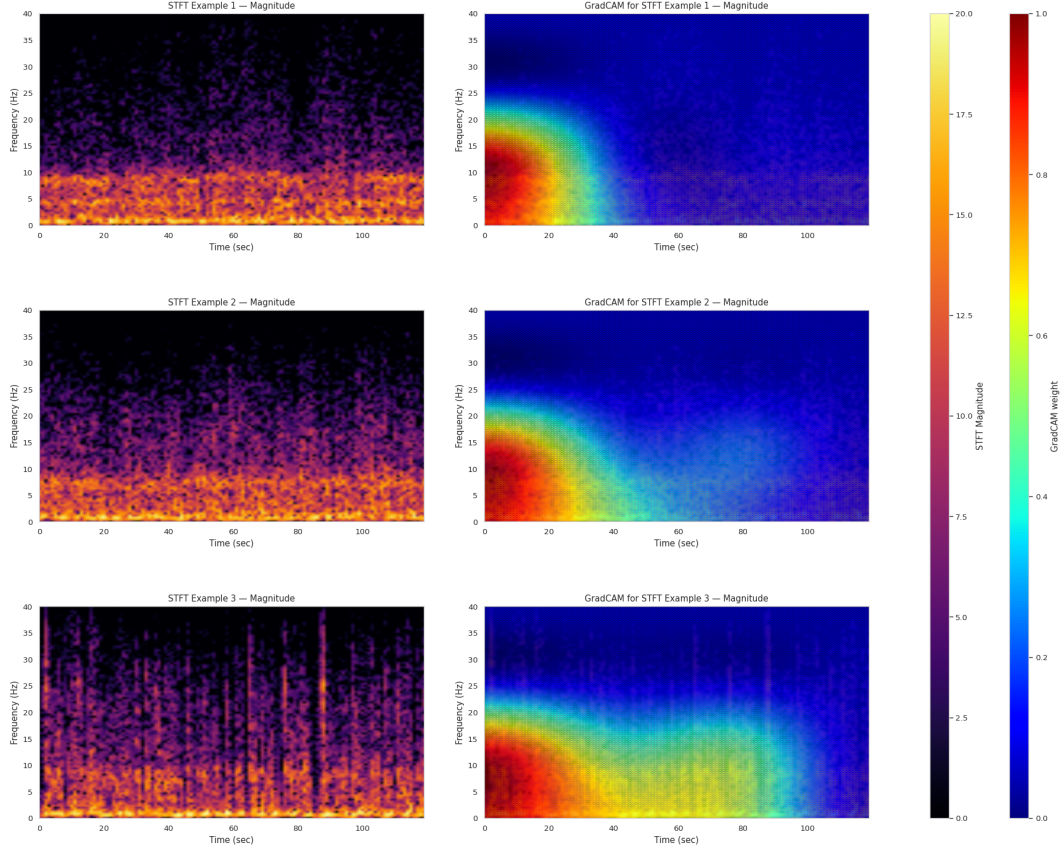


Figure 6.5: GradCAM visualisation of EEG spectrograms.

the standard EEG-PD-NLME model training. An additional dataset was constructed, containing only pre-surgical data points (i.e. data that is between the point 180 seconds after intubation, and the point of first surgical incision). Any data points that contained missing data or crossed over into the surgical period were removed. This data was assumed to have no nociceptive input.

The loss function $\mathcal{L}_B(\eta, \phi)$ (fentanyl MSE in this case) was calculated after passing a batch of bolus data B through the full EEG-PD-NLME model. Then, an equally sized batch of pre-surgical data $\tilde{B}$ was passed through f_ϕ , giving $\hat{n}$ values for each EEG segment in the pre-surgical batch. As these segments were from a time where nociceptive stimulus was assumed to be zero, the target was $n = 0$. The loss function for each batch $\tilde{B}$ was therefore calculated by taking the sum of squares of the output of f_ϕ , which was equivalent to an MSE loss with target zero for every data point:

$$\mathcal{L}_{\tilde{B}}(\phi) = \frac{1}{|\tilde{B}|} \sum_{i \in \tilde{B}} [f_\phi(x_i)]^2$$

This loss function was combined with the original loss function with a balancing hyperparameter λ_b to give the total loss function used for training:

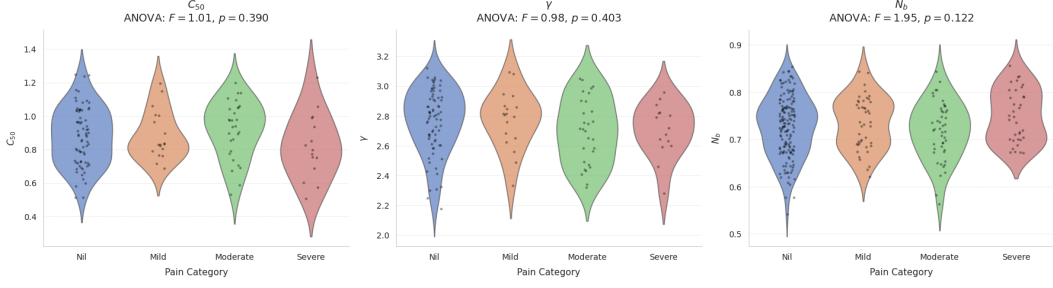


Figure 6.6: Violin plots for random effect distributions across each pain category.

$$\mathcal{L}_{B,\tilde{B}}(\eta, \phi) = \mathcal{L}_B(\eta, \phi) + \lambda_b \mathcal{L}_{\tilde{B}}(\phi)$$

The effect of this was that the fixed effects ϕ would be tuned such that pre-surgical EEG segments would be more likely to give a f_ϕ output of zero.

6.6.2 Experiment

Description

Three different values of λ_b were tested: 0.1, 0.5, and 1.0. The hyperparameters otherwise remained identical to the previously established ResNet-18 model, with 120-second STFT input, no weight decay, a learning rate of 1×10^{-3} , the EEG dataset with standard cleaning, and the proportional EEG-PD-NLME model variant. The training setup was identical, using an Adam optimiser with smoothed early stopping on $\bar{\rho}$. The base loss function $\mathcal{L}_B(\eta, \phi)$ was the fentanyl MSE.

The best model underwent the monotonic CDF transform and was evaluated on the test set, and the relevant prediction plots, random effects, and training curves were plotted.

Results

Table 6.10 gives the results for this experiment. The best performing model was when λ_b was 0.5. This model was further investigated, with relevant plots from the training and validation process shown in Figure 6.7 and Figure C.3.

The results of evaluating this model on the test set are shown in Table 6.11, with example plots given in Figure 6.8. Prediction of pain score category was again weak but better than the previous ResNet model, with prediction of pain score > 0 being significantly greater than chance (0.606; 95% CI: 0.504–0.706). The correlation with OpenIbis was non-significant on an inter-patient level, and the intra-patient correlation was also very low.

The effect of the additional training was only very weakly evident in criterion C1, which saw a small decline of 2.619 points from the pre-surgical period

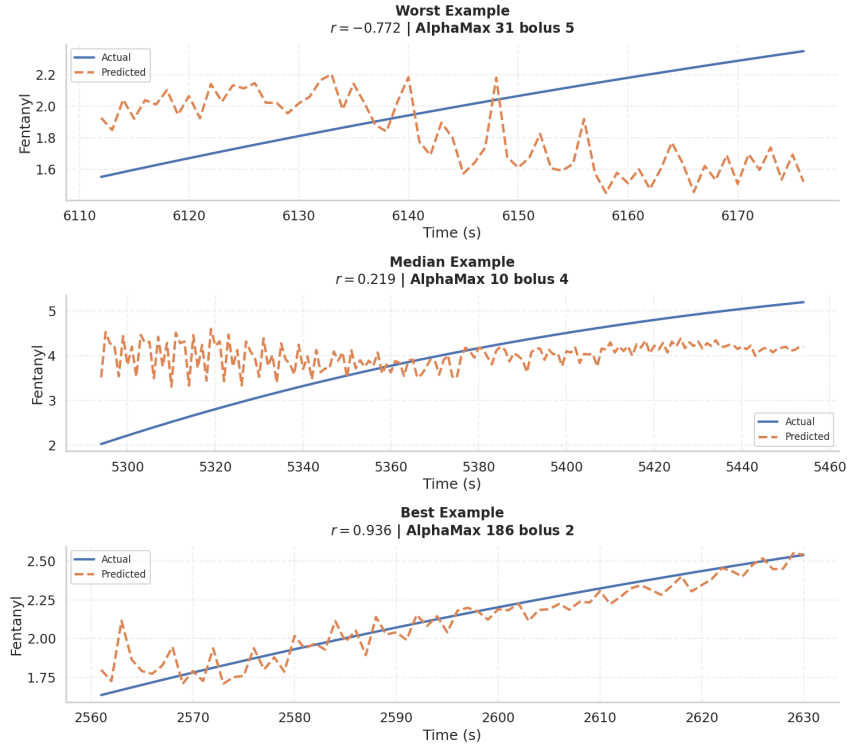
Table 6.10: Pre-surgical regularisation experiment results across values of the balancing hyperparameter λ_b , using the ResNet-18 f_ϕ architecture with 120-second STFT input.

λ_b	Train		Validation		Random Effects (Mean $\pm$ SD)		
	$\bar{\rho}$	MSE	$\bar{\rho}$	MSE	N_b	C_{50}	γ
0.1	-0.9823	0.0009	-0.1168	0.4572	0.726 ± 0.053	0.868 ± 0.159	2.748 ± 0.201
0.5	-0.9725	0.0029	-0.2244	0.3816	0.723 ± 0.048	0.865 ± 0.146	2.728 ± 0.174
1.0	-0.9884	0.0010	-0.1891	0.3843	0.728 ± 0.066	0.862 ± 0.167	2.787 ± 0.241

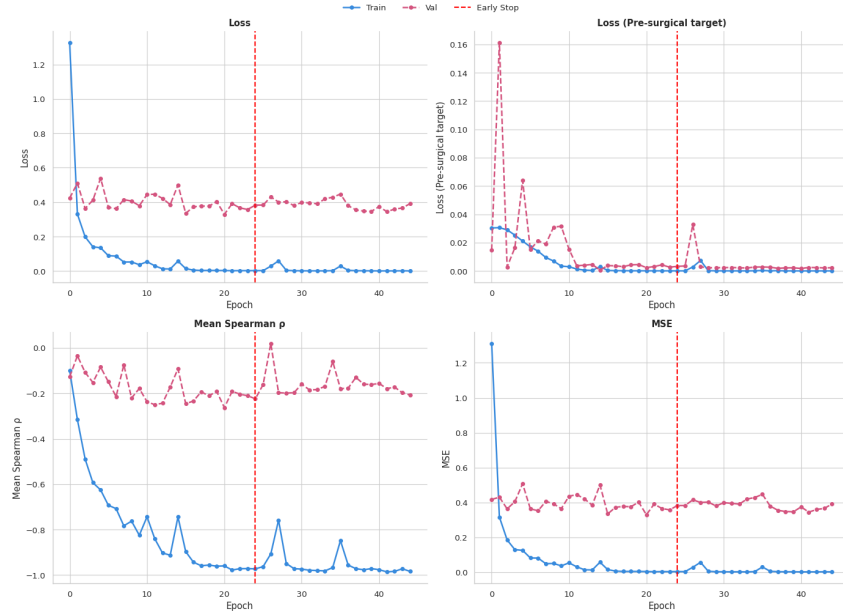
Table 6.11: Summary of test set results for EEG-PD-NLME with ResNet-18 f_ϕ and 120-second input, trained with additional pre-surgical data.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.606 (0.504–0.706)
AUROC > 3	0.539 (0.429–0.649)
AUROC > 6	0.547 (0.382–0.715)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.005 ($p = 0.587$)
Δ AUROC > 3	0.005 ($p = 0.665$)
Δ AUROC > 6	-0.020 ($p = 0.313$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = 0.040$ ($p = 0.676$), Spearman $\rho = 0.018$ ($p = 0.850$)
B2. Intra-patient corr.	Mean Pearson $r = 0.049$, Mean Spearman $\rho = 0.044$
C. Start of surgery	
C1. Mean comparison	Pre: 52.128, Surg: 49.509, Diff: -2.619
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.047$, Mean Spearman $\rho = -0.106$
E. Explainability	
E1. Delta coeff.	-2.584 ($p < 0.001$)
E2. Alpha coeff.	6.253 ($p < 0.001$)
E3. Beta coeff.	0.580 ($p < 0.001$)

Bolus-Level Pearson Correlation Examples



(a) Predicted vs. actual $\hat{C}_e$ and C_e fentanyl curves for the best, median, and worst boluses from validation set ranked by r .



(b) Training and validation curves for loss, MSE, pre-surgical loss term and mean Spearman $\bar{\rho}$ across epochs.

Figure 6.7: Bolus prediction examples and training curves from 120-second ResNet-18 model with pre-surgical regularisation ($\lambda_b = 0.5$).

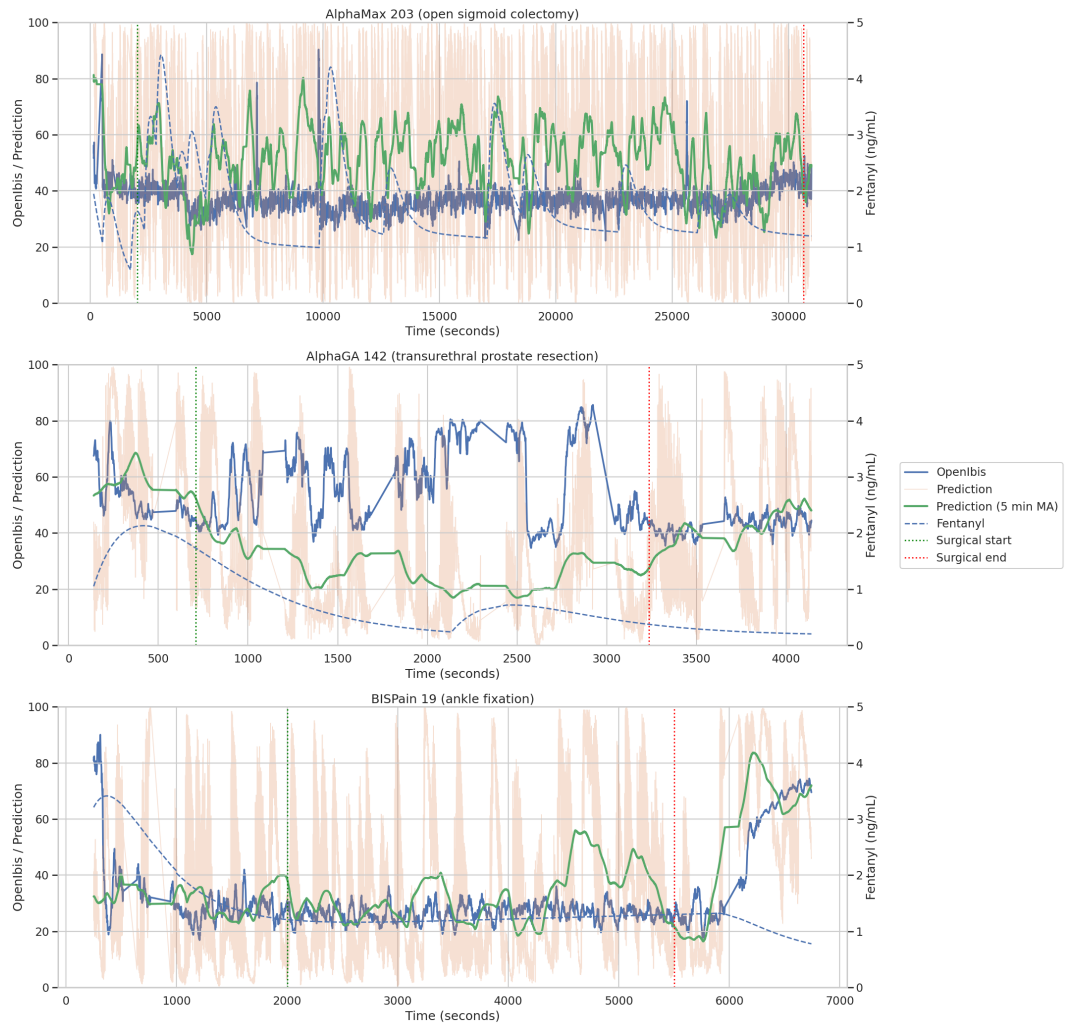


Figure 6.8: ResNet-18 model with additional training on pre-surgical data applied to three example EEG recordings from the test dataset.

to the surgical period. This is a smaller fall than that observed in the previous EEG-PD-NLME model, but it is still in the opposite direction to what was expected. The correlation with fentanyl concentration during boluses was slightly stronger but still weak, with a Spearman ρ of -0.106 . The surrogate model coefficients for the alpha and delta bands had opposite signs to what would be anticipated under criteria E1 and E3, as a higher index was associated with more alpha power and less delta power. The beta power association was only very weakly in the expected direction.

6.7 Discussion

The signal that was found by this model was weak. The mean Spearman correlation across all boluses, $\bar{\rho}$, indicated a weak negative correlation across most experiments. $\bar{\rho}$ was a noisy metric, and multiple testing on the validation set appeared to inflate it. When the same metric was calculated on the test set (as criterion D1), the mean correlation was substantially weaker than that which was found on the validation testing.

The fact that the constant f_ϕ model outperformed all other model variants in terms of MSE further reinforces the notion that the signal was very weak. A constant f_ϕ appears to allow for marginally better fitting of g_η and therefore improved MSE on the validation set.

The final results on the test set did not align well with the expected properties of an index of nociception. Postoperative pain prediction was poor. Despite the model being trained on fentanyl effect-site concentrations, the mean correlation with fentanyl effect-site concentration was also very weak, although in the correct direction. Criterion C1, the change in the index between the pre-surgical and surgical periods, was negative in the initial model. It was fulfilled somewhat better in the model that was trained on pre-surgical data, but the magnitude of this difference remained small.

It is reassuring that the random effects, particularly C_{50} and γ , appeared to be reasonably consistent across most of the well-performing models. In most experiments, the mean value of C_{50} ranged between 0.8 and 1.0, suggesting that the initial baseline value of 1.0 was an overestimate. The mean value of γ ranged between 2.6 and 2.8, suggesting that the initial baseline value of 2.5 was an underestimate. However, it was found that, when comparing the mean values of these PD parameters between patients who ultimately reported pain scores in different categories, there did not appear to be any statistically significant differences. This may be because pain score is a noisy target and is generally difficult to predict.

The saliency map generated by GradCAM on the 120-second ResNet-18 model predictions found that the highest saliency was at the start of the EEG segment (i.e. the time closest to when the fentanyl effect-site concentration label was measured). This aligns with clinical intuition and suggests that f_ϕ seems to have learned some signal, even if it was not a strong signal. The fre-

quencies with the highest saliency are those below 20 Hz with highest concentration in the alpha band region, which is logical given the known relationships between alpha power and nociception. The region of high saliency typically extends out to approximately 40 seconds after the label measurement, although in some cases this extends out to 100 seconds. This could be a consequence of extended EEG manifestations of nociception and analgesia, which have been observed in animal models [113]. It may also be due to autocorrelation: adjacent time windows are likely to have similar information, and therefore will receive similar saliency scores. The noisiness of EEG signal will mean that a model would likely benefit from using a large set of autocorrelated windows to generate a prediction.

The experiments conducted in this chapter aimed to investigate the effect of adjusting model hyperparameters, rather than to generate a model with optimal performance. Marginal performance gains may be achieved with a more thorough search of the hyperparameter space. In particular, the threshold for defining a bolus-associated EEG segment could be modified. In these experiments, a threshold of 0.01 ng/mL/s was used to define the qualifying segments. Lowering the threshold would increase the size of the dataset by allowing for more qualifying segments. However, the key assumption of the model (that any fluctuations in N_b over the course of a segment are negligible) would be more likely to be violated.

In terms of random effects, this model was limited to modelling PD parameters. However, it is known that PK parameters also differ between individuals. These parameters were scaled by patient weight in the initial calculation of fentanyl effect-site concentration, but patients of identical weight may have different PK profiles due to factors such as genetic variants, liver function, and other medications the patient may be on [174]. An extension of the EEG-PD-NLME model may also involve PK parameters being modelled as part of the random effects η . In the current model, much of this variation will be explained as part of the PD model, which is unlikely to have the correct inductive bias for modelling PK accurately.

It should be noted that, because this model is only trained on EEG data recorded while fentanyl concentration was rising, the use of this model as a general predictor of nociception is an extrapolation. The only labels for nociception used in this model are fentanyl effect-site concentration, and in the model trained on pre-surgical data, whether the surgery had started. It is unclear if this would introduce any systematic bias into the model's estimates.

Due to the use of fentanyl concentration as a surrogate label for nociception, this model cannot easily distinguish between any direct effect that fentanyl may have on the EEG, and the antinociceptive effect of fentanyl. It was hoped that the addition of pre-surgical data would cause f_ϕ to extract features that were genuinely associated with nociception rather than direct fentanyl effect, although it does not directly deconfound fentanyl effect from nociception.

This model has no mechanism to account for the hypnotic dimension of

anaesthesia. It is known that fentanyl has some effect on hypnosis [116]. The hypnotic dimension has a large effect on EEG, so it is possible that some of the ability of f_ϕ to detect changes in fentanyl may have been from the hypnotic effect. For the initial 120-second ResNet-18 model, the correlation with OpenIbis at the intra-patient level was weak ($r = 0.055$). However, the correlation on an inter-patient level was very strong ($r = 0.519$). This seems to suggest strong confounding that is tied to patient identity rather than intra-patient hypnotic levels.

When the pre-surgical data were added, the relationship remained similar, although the relationship appears to disappear on an inter-patient level. This, however, did not necessarily rule out confounding by hypnotic depth. Explicit mechanisms to adjust for the hypnotic depth of anaesthesia, either as measured by OpenIbis or hypnotic drug concentration, form the basis of the next chapter.

Chapter 7

Representation Learning

7.1 Motivation

The EEG-PD-NLME model used only data where fentanyl effect-site concentration was clearly rising following a bolus of fentanyl. This means that only a small portion of the total dataset is used for training. Ideally, the entire EEG dataset and the associated fentanyl effect-site concentrations would be used to construct a model. However, variable levels of surgical stimulus, as well as confounding effects such as the drug titration paradox, make using fentanyl-based methods more complicated than when using bolus data alone. To reduce the effect of these complications, this chapter simplifies the fentanyl effect-site concentration to a binary label, depending on whether fentanyl effect-site concentration is above or below a set threshold.

Another issue that has arisen in previous attempts to construct an index of nociception is confounding. EEG is highly confounded both by patient identity and hypnotic depth. An ideal index of nociception should function similarly regardless of patient identity or hypnotic depth (with the caveat that nociception and hypnotic depth are thought to have some correlation [3]). Adversarial techniques from representation learning are therefore utilised in this chapter in an attempt to either partially or fully remove patient identity or hypnotic depth information from representations of EEG data. This will allow for estimates of nociceptive state to be made with reduced confounding from nuisance variables.

7.2 Adversarial Learning

The prototypical adversarial architecture is the Generative Adversarial Network (GAN), which aims to learn a distribution p_{data} given data points $x \sim p_{\text{data}}$. There are two networks in a GAN: a generator network $G_{\theta}(z)$ and a discriminator network $D_{\phi}(x)$. The generator network takes an input noise variable $z \sim p(z)$, and outputs a sample in the data space $\tilde{x}$, drawn from the generator data distribution p_g . The discriminator takes this sample and out-

puts a probability that $\tilde{x}$ is drawn from p_{data} rather than p_g . This creates a minimax objective:

$$\min_{\theta} \max_{\phi} \mathbb{E}_{x \sim p_{\text{data}}(x)} [\log D_{\phi}(x)] + \mathbb{E}_{z \sim p(z)} [\log(1 - D_{\phi}(G_{\theta}(z)))]$$

Once the two networks reach a Nash equilibrium, the probability for both real and generated instances will be 0.5, at which point p_g can be considered to have approximated p_{data} [181].

The adversarial learning concept was extended to Domain-Adversarial Neural Networks (DANNs), which aim to train a network that creates representations invariant to domain shift. Given labelled data $(x_s, y_s) \sim p_s(x, y)$ from a source domain s and unlabelled data $x_t \sim p_t(x)$ from a target domain t , with $d \in \{0, 1\}$ denoting the domain label, the aim is to maximise performance on the target while maintaining representational invariance. The three parts of the model are:

- An encoder f_{θ} . This takes an input x from either domain, and outputs a representation z .
- A task head g_{ϕ} . This takes z as an input and outputs a prediction $\hat{y}$. This head has loss function $\mathcal{L}_y = \mathcal{L}(g_{\phi}(f_{\theta}(x_s)), y_s)$, taking labelled data points (x_s, y_s) as an input.
- A domain classifier h_{ψ} . This takes z as an input and classifies whether z is from the source ($d = 0$) or target domain ($d = 1$). This head has loss function $\mathcal{L}_d(h_{\psi}(f_{\theta}(x)), d)$, taking data points x_s and x_t as an input.

The encoder and the task head aim to minimise the task loss $\mathcal{L}_y$, whilst also aiming to maximise the domain classifier loss $\mathcal{L}_d$. Given λ , a balancing parameter that defines the strength of the adversarial loss relative to the task loss, the minimax objective is:

$$\min_{\theta, \phi} \max_{\psi} \mathcal{L}_y - \lambda \mathcal{L}_d$$

In practice, this objective is implemented using a gradient reversal layer (GRL). During forward propagation, the GRL R_{λ} is an identity transform:

$$R_{\lambda}(x) = x$$

However, during backpropagation, the GRL takes the gradient from the following layer and multiplies it by $-\lambda$:

$$\frac{dR_{\lambda}}{dx} = -\lambda I$$

This has the effect of reversing the optimisation direction of parameters upstream from the GRL. In DANNs, the GRL is placed between the encoder f_{θ} and the domain classification head h_{ψ} . While the weights ψ optimise for

domain classification, the weights θ optimise for domain invariance, whilst simultaneously optimising for task ability [182].

A similar adversarial approach was used by Özdenizci *et al.* (2020) to create subject-agnostic representations of EEG [183]. Consider a dataset $\{(x_i, y_i, s_i)\}_{i=1}^n$, where x_i is an EEG segment, y_i is a label, and $s_i \in \{1, \dots, S\}$ is the identity of a subject. Subject identity is considered to be a nuisance parameter, and it is assumed that the label y is marginally independent from subject identity. A similar model can be constructed with encoder f_θ , task head g_ϕ , and subject classifier h_ψ .

The task loss, subject classification loss, and overall minimax objective are effectively identical to the DANN objective:

$$\begin{aligned}\mathcal{L}_y &= \mathcal{L}(g_\phi(f_\theta(x_i)), y_i) \\ \mathcal{L}_s &= \mathcal{L}(h_\psi(f_\theta(x_i)), s_i) \\ \min_{\theta, \phi} \max_{\psi} \mathcal{L}_y - \lambda \mathcal{L}_s\end{aligned}$$

The implementation of this is similar to GRL, except it updates ψ separately from the other parameters at each batch update. Given learning rate κ , the parameter updates are:

$$\begin{aligned}\psi &\leftarrow \psi - \kappa \nabla_{\psi} \mathcal{L}_s \\ \theta, \phi &\leftarrow \theta, \phi - \kappa \nabla_{\theta, \phi} (\mathcal{L}_y - \lambda \mathcal{L}_s)\end{aligned}$$

7.3 Formulation

7.3.1 Baseline Binary Classification Model

In the baseline model without adversarial training, the following are defined:

- EEG segments x . The segment begins at time t seconds, and continues to time $t + w$.
- Binary outcome y . This is 1 when t is after the surgical start time t_{start} and the effect-site concentration of fentanyl C_e at time t is below a threshold C_f . Concretely:

$$y(t) = \mathbf{1}[t \geq t_{\text{start}} \wedge C_e(t) < C_f]$$

This is a substantial simplification of the EEG-PD-NLME model. When $y = 0$, nociception is unlikely, as the surgery has either not started, or the effect-site concentration of fentanyl is sufficiently high to block surgical stimuli. When $y = 1$, the surgery has begun and the concentration of fentanyl is low. It is noted that these categories are not necessarily perfectly accurate. A low concentration of fentanyl after the surgical start time does not necessarily mean

that surgical stimulus is occurring. Similarly, a strong surgical stimulus that occurs when fentanyl is high (particularly in a patient with a high C_{50}) may cause nociception.

The baseline model is composed of an encoder $f_\theta(x)$, which outputs a representation z . This representation is then used as an input to the task head $g_\phi(z)$, which outputs a prediction $\hat{y}$ (corresponding to the probability of $y = 1$). $g_\phi \circ f_\theta$ is trained as a single model, using binary cross-entropy loss $\mathcal{L}_y(y, \hat{y})$.

When used for inference, the logit output of $g_\phi(z)$ is used as the nociception index. This undergoes the previously mentioned monotonic CDF transform (using the combined train and validation set results as the reference distribution), giving an index ranging from 0 to 100.

7.3.2 Adversarial Heads

A similar implementation to Özdenizci *et al.* (2020) was employed to attempt to make z , the output of the encoder $f_\theta(x)$, invariant to patient identity p and some measure of hypnotic depth d (e.g. OpenIbis). This gives two possible adversarial heads:

- The first adversarial head, $h_{\psi_p}(z)$, is a classifier, trained with cross-entropy loss $\mathcal{L}_p(p, \hat{p})$.
- The second adversarial head, $h_{\psi_d}(z)$, is a regressor, trained with mean squared error $\mathcal{L}_d(d, \hat{d})$.

As in Özdenizci *et al.* (2020), each batch update alternates between updating ψ and updating θ and ϕ . If both adversarial heads were to be used simultaneously, given balancing constants λ_p and λ_d , the minimax objective would be:

$$\min_{\theta, \phi} \max_{\psi_p, \psi_d} \mathcal{L}_y - \lambda_p \mathcal{L}_p - \lambda_d \mathcal{L}_d$$

This gives the following alternating gradient updates:

$$\begin{aligned} \psi_p &\leftarrow \psi_p - \kappa \nabla_{\psi_p} \mathcal{L}_p \\ \psi_d &\leftarrow \psi_d - \kappa \nabla_{\psi_d} \mathcal{L}_d \\ \theta, \phi &\leftarrow \theta, \phi - \kappa \nabla_{\theta, \phi} (\mathcal{L}_y - \lambda_p \mathcal{L}_p - \lambda_d \mathcal{L}_d) \end{aligned}$$

The scenario where there are two adversarial heads is summarised by Figure 7.1.

7.3.3 Causal Analysis

Figure 7.2 is a directed acyclic graph (DAG) that represents the causal model that is used in this chapter. The following relationships are assumed to exist:

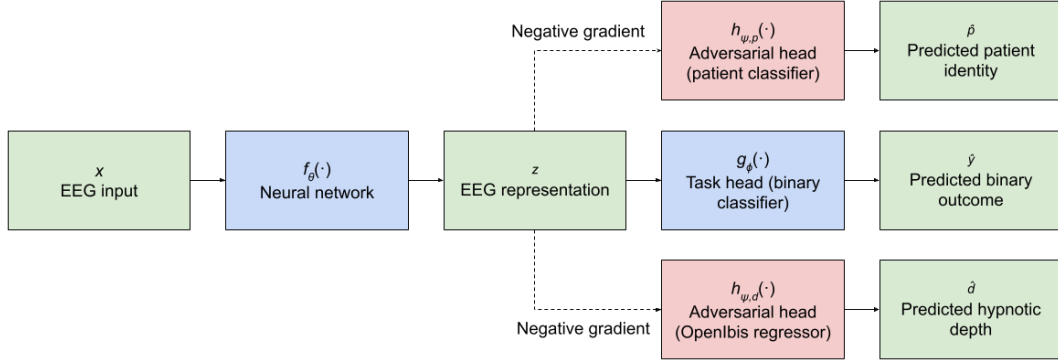


Figure 7.1: Overview of model with two adversarial heads.

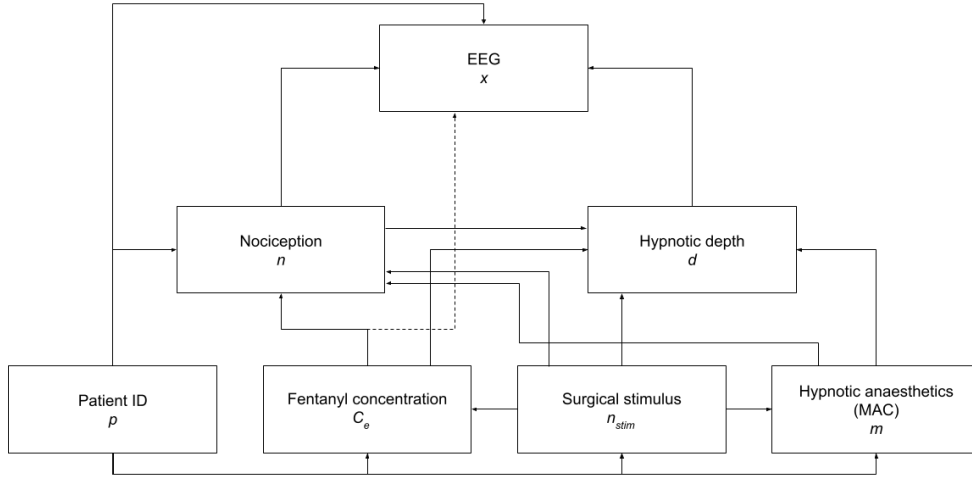


Figure 7.2: Directed acyclic graph illustrating the hypothesised causal relationships during general anaesthesia.

- Patient identity p includes information about the patient's prior brain state, electrode placement, surgical information, and anaesthetist behaviour. It therefore directly affects the EEG x , influences the dosing and PK of fentanyl C_e and hypnotic anaesthetics m , and affects the surgical stimulus n_{stim} .
- Fentanyl concentration C_e primarily affects nociception n , but also has some effect on hypnotic depth d . It also has a small direct effect on the EEG [117].
- Surgical stimulus n_{stim} affects nociception n and hypnotic depth d . The levels of fentanyl C_e and hypnotic anaesthetic m are also titrated in response to surgical stimuli.
- Hypnotic anaesthetic m primarily affects hypnotic depth d , but also has

some effect on nociception n .

- Nociception n affects EEG x . Given the correlations between n and hypnotic depth d , it is assumed that n has an effect on d .
- Hypnotic depth d has a very substantial effect on the EEG x .

The aim of this study is to produce a ML model that describes the relationship between x and n . However, as shown by Figure 7.2, there are multiple backdoor pathways that could confound estimation of this relationship.

Firstly, there are the pathways that are mediated by patient identity p . This directly affects levels of nociception and the EEG, as well as affecting fentanyl concentration, surgical stimulus, and levels of hypnotic anaesthetic. If this pathway is exploited, it could lead to spurious results, including effects similar to the drug titration paradox. For instance, a model that learns to exploit this backdoor pathway may learn patient-level features that, in reality, tend to be associated with increased nociception. However, as these patients receive more fentanyl, the model will associate these features with decreased nociception. It is therefore desirable to learn patient-agnostic representations of the EEG z , where $z \perp p$.

There are also backdoor pathways that are mediated by hypnotic depth d . Surgical stimulus, fentanyl concentration, and hypnotic anaesthetics affect both hypnotic depth and nociception. However, hypnotic depth mediates some of the effect of nociception, so it is anticipated that, in an ideal nociception index, there is some correlation between n and d . This means that the representation z should not necessarily be entirely invariant to d , as full invariance would eliminate some of the relationship that we are aiming to model. The goal is therefore to learn partially hypnotic-invariant representations of the EEG (where ideally the backdoor confounding pathways are eliminated but the direct effects remain).

The final backdoor pathway is the direct fentanyl effect. As fentanyl effect-site concentration is used to derive the label of nociception, there is no clear way to eliminate this confounder without eliminating the signal, and it remains a limitation of this model.

7.4 Methods

The dataset used for representation learning was the same STFT dataset that was used with the EEG-PD-NLME model. The standard cleaning variant was used, with 120 seconds of data for each x_i . These data points were sampled once every 60 seconds, meaning that there were 60 seconds of overlap between adjacent data points. Any data points that contained NaN values or were within 180 seconds of induction were excluded. y_i was determined for each EEG segment as previously described.

Due to its performance in the previous chapter, the model that was chosen for the encoder f_θ was the ResNet-18 model, pretrained with ImageNet weights. This model gives a 512-dimensional embedding z . The classification head g_ϕ and the adversarial heads h_ψ take z as an input.

All models were trained on the train set using binary cross-entropy loss, and evaluated on the validation set. The validation metrics that were used for the primary binary target were accuracy and AUROC. An Adam optimiser with learning rate 1×10^{-3} was used, and the batch size was 64. Training was conducted for a maximum of 50 epochs, with early stopping using validation AUROC as the stopping metric, and with a patience of 10 epochs.

7.5 Experiments

7.5.1 Experiment 1: Baseline Model

Description

The first experiment was conducted with the aim of establishing a baseline model for classifying the discrete label y , and selecting an appropriate fentanyl cut-off value C_f . For each $C_f \in \{1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$, the model $g_\phi \circ f_\theta$ was trained (i.e. with no adversarial head). The task head g_ϕ consisted of a single linear layer, taking the 512-dimensional input z and giving an output $\hat{y}$.

In addition to the hard cut-offs, an experiment with soft labels was also conducted. Using the values of $C_{50} = 1.0$ ng/mL and $\gamma = 2.5$, numeric values of $y \in [0, 1]$ were calculated from the effect-site concentration of fentanyl using the Hill equation. $y = 0$ when the EEG was before the recorded surgical start time. Otherwise, y was calculated as follows:

$$y = 1 - \frac{C_e^\gamma}{C_{50}^\gamma + C_e^\gamma}$$

The soft label experiment was conducted and evaluated similarly to the experiments with hard labels. Rather than validation AUROC being the key validation metric, due to the continuous nature of the labels, the concordance index was used instead, and accuracy was omitted.

After each model was trained, the output of the model on the combined train and validation set was calculated. These outputs were then transformed using the previously defined monotonic CDF transform, and evaluated. The best performing model on the unified validation framework was used as the basis for the following experiments.

The representation $z = f_\theta(x)$ from the best model was inspected. Ten of the patients from the train set were randomly selected, and the representations z were calculated for each EEG segment for each patient. The representations underwent UMAP dimensionality reduction, and were plotted on 2D axes to determine if there was any visual clustering of the representations by patient p .

Probes were then trained to analyse the representation z quantitatively, with the goal of determining how much information the baseline representations had about hypnotic state and patient identity. These representations were calculated across all EEG segments, across all patients in the train and validation sets. Probes were trained on the train set, and tested on the validation set.

The probes were trained to predict the following end points:

- OpenIbis score.
- Adjusted concentration of volatile anaesthetic m_{adj} , measured in MAC.

As volatile anaesthetic has an additive effect with propofol, the raw volatile anaesthetic concentration m was adjusted to take into account the effect-site concentration of propofol. A conversion factor of 3.6 was used for this [184]:

$$m_{\text{adj}} = m + \frac{C_{e,\text{prop}}}{3.6}$$

- Patient identity.
- Patient age.
- Patient weight.
- Patient sex.

Direct prediction of patient identity required a different train-validation split, as a patient-level split meant that this probe could not be evaluated. The train and validation sets were aggregated, and split on a per-segment basis into an 80% train set and a 20% validation set. The remaining probes used the pre-existing train and validation sets.

For the continuous outcome probes (OpenIbis, adjusted volatile anaesthetic concentration, age, weight), MSE and MAE were the metrics used. For the binary outcome probe (patient sex), accuracy and AUROC were used. For the categorical outcome probe (patient identity), accuracy, class-balanced accuracy, and F1 score were used.

There were three types of probes that were used:

- Baseline probe. For the continuous outcomes, this probe predicted the mean value. For the binary and categorical outcomes, this probe predicted the modal value. This provides a baseline to contextualise the performance of the other probes.
- Linear probe. For the continuous outcomes, this was Ridge regression with regularisation parameter $C = 1.0$. For the binary outcome, this was logistic regression. For the categorical outcome, this was categorical Ridge classification with $C = 1.0$.

- Non-linear probe. A Histogram-based Gradient Boosting (HGB) classifier or regressor was used for this, with a maximum depth of 6, a learning rate of 0.05, and a maximum iteration count of 300.

Results

Table 7.1 gives the validation results of this experiment.

Table 7.1: Experimental results across different values of C_f .

Model	$C_f = 1.0$	$C_f = 1.5$	$C_f = 2.0$	$C_f = 2.5$	$C_f = 3.0$	$C_f = 3.5$	$C_f = 4.0$	Smooth y
AUROC	0.5421	0.5833	0.6174	0.6710	0.7399	0.7731	0.7749	0.5280*
Accuracy	0.6133	0.6071	0.6570	0.8105	0.8396	0.8608	0.8797	–

* Concordance statistic rather than AUROC.

As C_f increased, y became easier to predict, as AUROC and accuracy both rose almost monotonically. This reflected the proportion of $y = 0$ data points after the surgical start time getting smaller. The classification problem therefore approached the much more confounded task of classifying whether a segment was taken from before surgery, or during surgery. These metrics alone are therefore insufficient to select C_f .

Table D.1 in Appendix D shows the full evaluation results of these models when the full validation framework is applied to the combined train-validation dataset. The strongest model was found when $C_f = 1.5$, which gave a model that was significantly predictive of postoperative pain score being above 0 (AUROC: 0.608; 95% CI: 0.557–0.662) and above 3 (AUROC: 0.605; 95% CI: 0.552–0.660). It also contributed to a small significant improvement in AUROC in the combined demographic model for both the pain score being above 0 (Δ AUROC: 0.003; $p = 0.022$) and above 3 (Δ AUROC: 0.003; $p = 0.040$). The mean index rose by 20.963 points from the pre-surgical to the surgical period, and it had a Pearson correlation of $r = -0.102$ with fentanyl during bolus periods.

This model was selected for further analysis. Figure 7.3 shows the UMAP for ten randomly selected patients. Although some patients appear to have some reasonable overlap with other patients, there are some patients who have many of their points clustered in a single part of the manifold (such as the cluster for AlphaGA 223).

Table 7.2 and Table 7.3 give the performance of the probes on the representation z generated by this model. The most strongly encoded variables appeared to be OpenIbis and patient identity, and they could be recovered most strongly with a linear probe. Mean prediction of OpenIbis gave an MSE of 167.9, but the linear probe had an MSE of 76.7. Modal prediction of patient identity gave a baseline prediction accuracy of 0.009, but the linear probe was able to recover patient identity correctly with an accuracy of 0.33. Age, sex, and volatile anaesthetic concentration were only weakly recoverable from the embedding. Weight did not appear to be recoverable from these probe results.

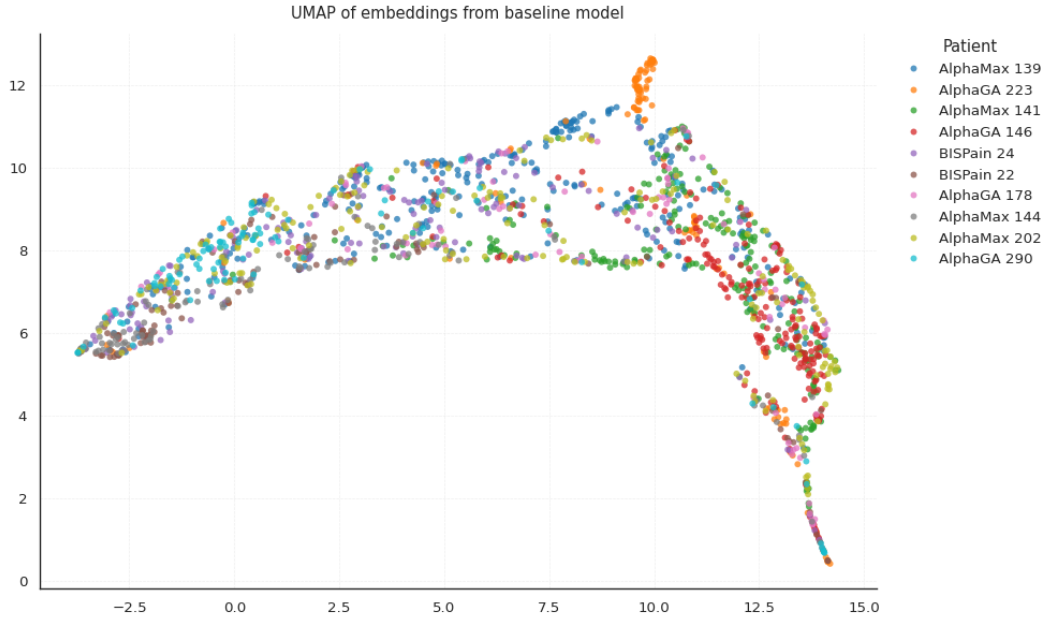


Figure 7.3: UMAP projection of the learned latent representations for ten randomly selected patients. Each point corresponds to a single sample, with colours indicating patient identity.

Table 7.2: Validation probe performance for continuous variables. Lower MAE and MSE indicate better predictive performance.

Variable	Metric	Baseline (Mean)	Linear Probe (Ridge)	Non-linear Probe (HGB)
OpenIbis	MAE	9.4616	6.0684	6.3466
	MSE	167.8557	76.6686	80.1432
m_{adj}	MAE	0.1667	0.1567	0.1593
	MSE	0.0585	0.0484	0.0498
Age	MAE	9.7387	8.7034	8.8106
	MSE	182.0475	130.1526	137.7963
Weight	MAE	15.2367	15.6511	15.3768
	MSE	384.1183	406.0251	389.0378

Table 7.3: Validation probe performance for categorical variables.

Variable	Metric	Baseline	Linear Probe	HGB Probe
Sex	Accuracy	0.6225	0.6219	0.6326
	AUROC	0.5000	0.6212	0.5959
Patient ID	Accuracy	0.0090	0.3306	0.0676
	Balanced acc.	0.0022	0.2157	0.0496
	Macro F1	0.0000	0.1890	0.0463

The strong performance of the linear probes on OpenIbis and patient identity suggested that these were encoded in z in a linearly separable manner. These two linear probes were the key evaluation metrics for subsequent experiments in this chapter.

7.5.2 Experiment 2: Patient Identity Invariance

Description

For this experiment, an additional validation set was constructed at the segment level, using 20% of the data from the existing train set. The original, patient-level validation set was used for evaluation of accuracy and AUROC. The segment-level validation set was used for evaluation of the adversarial mechanism.

This experiment ideally would introduce a linear adversarial head h_ψ , which takes z as an input, and outputs which patient p that the EEG segment belongs to. This head would be trained by cross-entropy loss, and could be trained using the previously described alternating method with each batch.

However, there is a difficulty with this method. Although h_ψ would be a very simple linear adversarial head, the high number of patients in the dataset turns it into a very large mapping between the 512-dimensional representation z and the 453 patients in the train set. This layer would take many hundreds of epochs to converge, whereas the task head g_ϕ typically converges within the first 20 epochs.

The Ridge classifier that was used for the probe was therefore used in lieu of the linear adversarial head. Whereas training this layer with stochastic gradient descent would be computationally expensive, training it analytically was relatively easy due to its linear nature. Given an encoder f_θ , the Ridge classifier was fit at the start of each epoch to predict patient identity:

$$\psi^* = \arg \min_{\psi} (\mathcal{L}_p(p, h_\psi(z)) + \alpha \|\psi\|_2^2)$$

The weights ψ^* were then copied into the adversarial head h_ψ , and were kept constant for the rest of the epoch. The encoder f_θ and task head g_ϕ were then trained to minimise task loss whilst maximising adversarial loss:

$$\min_{\theta, \phi}(\mathcal{L}_y - \lambda_p \mathcal{L}_p)$$

Gradient updates only affect the encoder and task head weights, with gradients flowing through the frozen adversarial head without updating its weights:

$$\theta, \phi \leftarrow \theta, \phi - \kappa \nabla_{\theta, \phi}(\mathcal{L}_y - \lambda_p \mathcal{L}_p)$$

Each epoch, the accuracy of the adversarial Ridge classifier on the train set was recorded. The segment-level validation set was used to monitor the accuracy of the adversarial task as the encoder was updated: this would be expected to be high at the start of each epoch, with a freshly optimised adversarial head, and would fall throughout the epoch as θ changes. The patient-level validation set was used to monitor the accuracy of the task head, with AUROC and accuracy calculated each epoch, and early stopping using the validation AUROC.

The remaining hyperparameter to tune was λ_p . In this experiment, several values were attempted: $\lambda_p \in \{0, 0.01, 0.1, 1, 10\}$. The outcomes of interest were AUROC and accuracy on the patient-level validation set (representing the ability of the model to perform the binary classification task), and the performance of a post-hoc probe to derive information from z . In this case, the specific probe that was used was the categorical Ridge classifier probe, and the performance of this probe on the segment-level validation set was reported as a measure of the amount of information present in z that was informative for determining p . Each experiment was repeated ten times, and a mean and standard deviation were reported. The previous UMAP plot from Experiment 1 was repeated with the EEG representations z from the model that demonstrated the greatest degree of patient-invariance, and the training and validation curves were provided for this model.

Results

Table 7.4: Effect of adversarial patient-invariance strength (λ_p) on task performance and patient-identification probe performance. Results are reported as mean $\pm$ standard deviation over ten random seeds. Lower probe metrics indicate greater patient invariance.

λ_p	Task AUROC	Probe acc.	Probe balanced acc.	Probe F1
0	0.598 $\pm$ 0.012	0.287 $\pm$ 0.028	0.191 $\pm$ 0.021	0.159 $\pm$ 0.021
0.01	0.605 $\pm$ 0.012	0.245 $\pm$ 0.036	0.160 $\pm$ 0.026	0.127 $\pm$ 0.026
0.1	0.581 $\pm$ 0.013	0.162 $\pm$ 0.029	0.101 $\pm$ 0.021	0.070 $\pm$ 0.021
1.0	0.574 $\pm$ 0.012	0.118 $\pm$ 0.026	0.069 $\pm$ 0.017	0.043 $\pm$ 0.017
10	0.567 $\pm$ 0.009	0.134 $\pm$ 0.030	0.083 $\pm$ 0.019	0.055 $\pm$ 0.019

Table 7.4 gives the results of this experiment. As the λ_p increased, there was a general tendency for the probe accuracy to fall, while the task AUROC also fell. The probe accuracy appeared to reach a minimum at around $\lambda_p = 1.0$. The minimum accuracy of 0.118 was substantially lower than the $\lambda_p = 0$ accuracy of 0.287, but it was not close to what would be achieved with a truly random probe (0.0090 with the baseline probe in Table 7.3). Increasing λ_p beyond this continued to degrade task performance whilst providing no additional benefit in terms of invariance. Partial patient invariance had been achieved, but not full invariance.

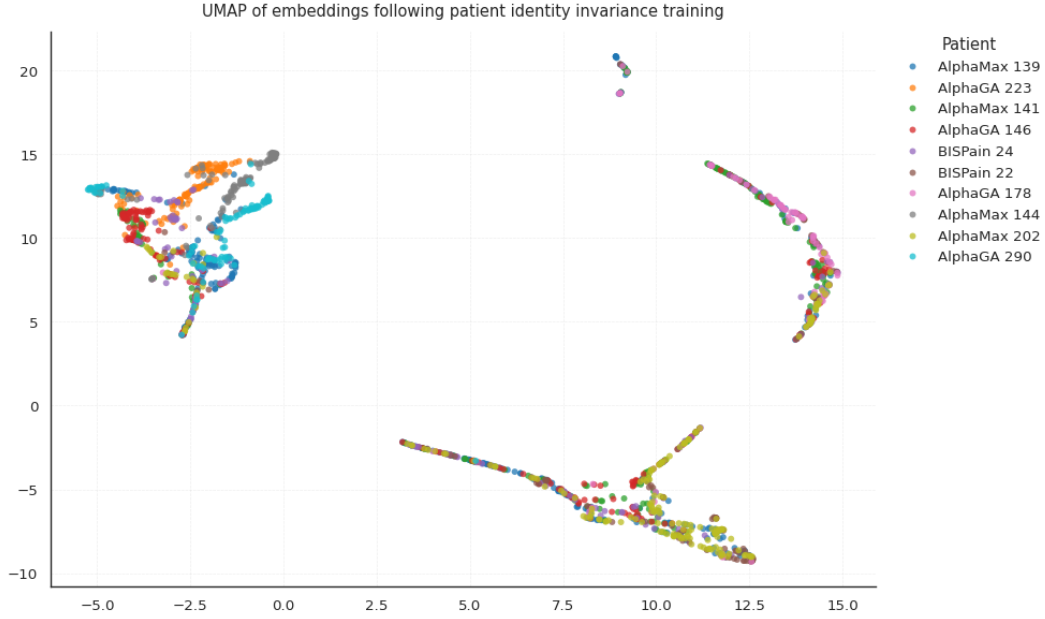


Figure 7.4: UMAP projection for partial patient-invariant model with $\lambda_p = 1$. Each point corresponds to a single sample, with colours indicating patient identity.

One of the trained models with $\lambda_p = 1.0$ was further investigated, with the two-dimensional UMAP plot for this model given in Figure 7.4. There are clear clusters in this plot, although it is noted that the clusters are heterogeneous by patient. There is still a small amount of patient-specific clustering, indicating that full patient identity invariance has not been achieved.

Figure 7.5 shows the training and validation curves for this model. The train loss appeared to reach a minimum in the first epoch, before the adversarial accuracy had converged. As adversarial training progressed and the adversarial classification accuracy fell, the task loss increased.

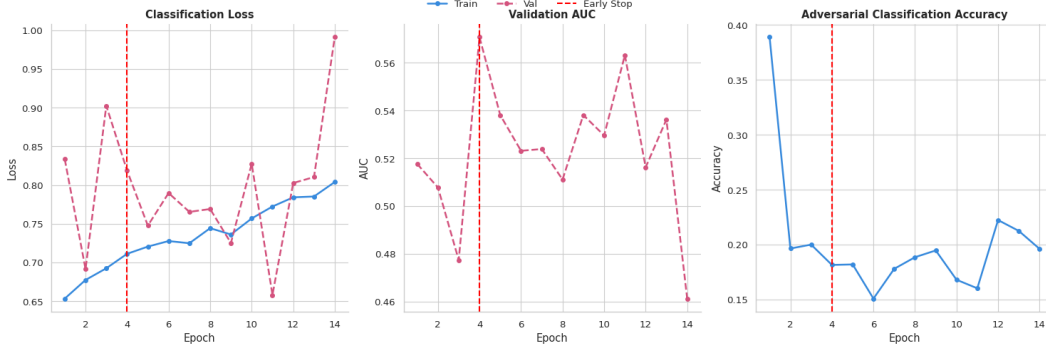


Figure 7.5: Train and validation loss, validation AUROC, and adversarial classification accuracy.

7.5.3 Experiment 3: Hypnotic Depth Invariance

Description

This experiment took a more standard approach towards adversarial training, directly using the alternating parameter updates of Özdenizci *et al.* (2020) [183]. The standard, previously defined patient-level validation set was used, with no requirement for a segment-level validation set.

The adversarial head h_ψ in this experiment was a linear regression that takes the 512-dimensional representation z as an input and outputs some scalar quantity d that corresponds to the depth of anaesthesia. This head was trained using mean squared error loss, using the alternating updates between ψ and θ, ϕ as previously defined.

There were three depth-of-anaesthesia measures that were used as the scalar quantity d in this experiment:

- OpenIbis. For the purpose of this experiment, OpenIbis was divided by 100 to give a value between 0 and 1 during training.
- Adjusted volatile anaesthetic concentration, measured in MAC.
- Smoothed OpenIbis. This was also scaled between 0 and 1, and took a five-minute moving average of OpenIbis. OpenIbis tended to be a noisy measure, so it is possible that smoothing it would give a better indication of overall hypnotic state.

For each of these possible d measures, multiple values of the balancing constant λ_d were used, with $\lambda_d \in \{0, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000\}$, giving 27 total experiments. For each experiment, a post-hoc Ridge regression probe was fit to the representation z and the raw OpenIbis value, using the train set. The performance of this probe on the validation set was recorded, with the aim for the key metrics, MSE and MAE, to be higher (indicating that the model had less information about the hypnotic state of the patient).

Note that, regardless of d used during training, the post-hoc probe always used the raw, unscaled 0–100 OpenIbis value to enable comparisons across different methods. Each of these experiments was repeated ten times to give a mean and standard deviation. The model with the most invariance in terms of probe MSE was investigated further, with training and validation plots created.

Results

The results of this experiment are given in Table 7.5. Generally, as λ_d increased, the probe MAE and MSE increased, indicating increasing invariance, although this effect seemed to slow or reverse at the very highest values of λ_d . The probe MAE and MSE did not reach the levels of the mean-prediction baseline probe from Experiment 1, which reached 9.46 and 167.86 for MAE and MSE respectively. This indicates that partial invariance was achieved rather than full invariance. Task AUROC degraded only slightly as λ_d was increased.

In terms of probe MSE, the most invariant experiment was the 5-minute moving-average OpenIbis with $\lambda_d = 10$, which had an average probe MSE of 130.45 (although the standard deviation of this was very high). The training and validation curve for one of these models is given in Figure 7.6.

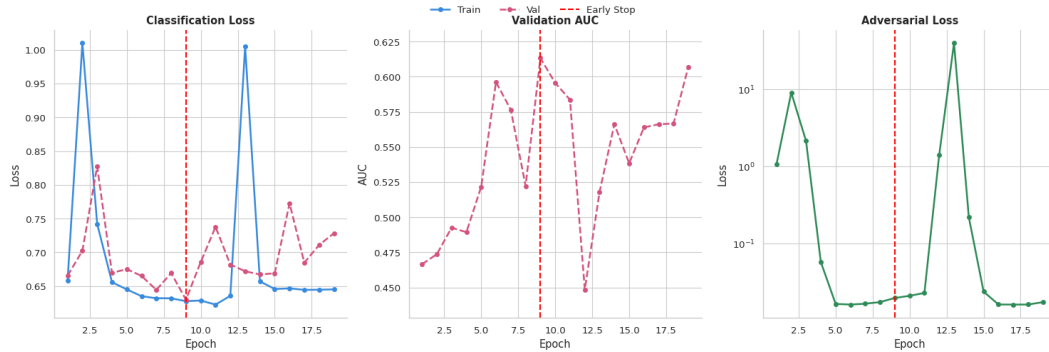


Figure 7.6: Train and validation loss, validation AUROC, and adversarial regression loss.

7.5.4 Experiment 4: Combined Adversarial Heads

Description

Experiment 2 and Experiment 3 investigated the effects of the adversarial heads h_{ψ_p} and h_{ψ_d} in isolation. In this experiment, both were combined into the same model. Rather than using just validation AUROC and probe performance, this experiment also investigated the performance of the model using the metrics from the unified validation framework, in order to identify an optimal set of hyperparameters for an index of nociception.

Like Experiment 2, this experiment had two validation sets: a patient-level validation set for evaluation of the task AUROC and accuracy, and hypnotic

Table 7.5: Effect of adversarial hypnotic-depth invariance strength (λ_d) on task performance and hypnotic-depth probe performance. Results are reported as mean $\pm$ standard deviation over 10 random seeds. Higher probe errors indicate greater invariance to hypnotic depth.

Depth measure	λ_d	Task AUROC	Probe MAE	Probe MSE
OpenIbis				
	0	0.606 ± 0.009	5.85 ± 0.23	72.79 ± 4.99
	0.01	0.603 ± 0.008	5.97 ± 0.46	74.89 ± 8.31
	0.1	0.610 ± 0.007	6.11 ± 0.45	77.66 ± 7.92
	1	0.620 ± 0.008	6.13 ± 0.35	81.79 ± 6.30
	10	0.606 ± 0.018	6.48 ± 0.93	91.61 ± 26.04
	100	0.587 ± 0.019	6.82 ± 0.66	97.93 ± 15.69
	1000	0.592 ± 0.014	7.02 ± 0.91	102.34 ± 20.68
	10000	0.586 ± 0.013	7.14 ± 1.00	108.47 ± 21.08
	100000	0.591 ± 0.014	7.18 ± 0.96	107.82 ± 27.43
Adjusted volatile anaesthetic concentration (m_{adj})				
	0	0.606 ± 0.009	5.85 ± 0.23	72.79 ± 4.99
	0.01	0.605 ± 0.007	6.05 ± 0.43	75.96 ± 8.40
	0.1	0.604 ± 0.006	6.00 ± 0.28	73.87 ± 4.65
	1	0.612 ± 0.018	6.59 ± 1.00	93.64 ± 23.85
	10	0.611 ± 0.011	6.32 ± 0.88	85.00 ± 20.07
	100	0.593 ± 0.009	6.65 ± 1.13	94.26 ± 26.62
	1000	0.582 ± 0.011	6.84 ± 0.88	101.53 ± 21.77
	10000	0.584 ± 0.015	7.23 ± 1.09	106.45 ± 25.54
	100000	0.582 ± 0.012	6.41 ± 0.68	86.35 ± 14.03
5-minute moving-average OpenIbis				
	0	0.606 ± 0.009	5.80 ± 0.29	71.34 ± 5.46
	0.01	0.608 ± 0.007	5.96 ± 0.32	75.03 ± 5.92
	0.1	0.605 ± 0.008	5.78 ± 0.27	72.65 ± 5.17
	1	0.621 ± 0.005	5.87 ± 0.18	76.65 ± 4.88
	10	0.608 ± 0.013	7.06 ± 1.82	130.45 ± 122.59
	100	0.580 ± 0.018	6.80 ± 0.77	112.02 ± 44.85
	1000	0.589 ± 0.013	7.49 ± 0.87	116.01 ± 21.53
	10000	0.591 ± 0.009	6.94 ± 0.89	100.98 ± 22.06
	100000	0.596 ± 0.013	6.91 ± 0.93	98.72 ± 20.68

depth invariance, and a segment-level validation set for evaluation of patient identity invariance. h_{ψ_p} was updated using the analytical method described in Experiment 2, and h_{ψ_d} was trained using the alternating updates described initially. The quantity d was the five-minute smoothed OpenIbis value, and the heads h_{ψ_p} and h_{ψ_d} were both linear.

Different values of λ_p and λ_d were used based on those that performed well in the previous two experiments, with a grid search being conducted over $\lambda_p \in \{0, 0.1, 1\}$ and $\lambda_d \in \{0, 10, 100, 1000, 10000\}$. For each experiment, two post-hoc probes were fit on the train set: one to predict patient identity, and one to predict the raw OpenIbis value. The patient identity probe was evaluated on the segment-level validation set, and the hypnotic depth probe was evaluated on the patient-level validation set. The same metrics as previously were used: accuracy, balanced accuracy, and F1 for patient identity, and MAE and MSE for depth. For each experiment, the resulting index underwent the monotonic CDF transform and was evaluated on the full combined train and validation dataset, using the unified validation framework criteria.

The best-performing models were then formally validated using the validation framework on the held-out test set. GradCAM was used on selected predictions in order to determine what parts of the input signal were most relevant for the predictions.

Results

Table 7.6 gives the probe results for this experiment. Results on the probes appeared to be noisy and inconsistent. There did appear to be a trend where the adversarial head for patient identity improved depth invariance, and the adversarial head for hypnotic depth appeared to improve patient identity invariance. The model with the lowest accuracy on the patient identity probe was when $\lambda_p = 0$ and $\lambda_d = 10$, giving an accuracy of 0.111 despite the model not having any adversarial training against patient identity. Similarly, the model with the highest MSE for OpenIbis was when $\lambda_p = 1$ and $\lambda_d = 0$, giving an MSE of 140.4. This may indicate a strong relationship between hypnotic depth and patient identity, and making a representation that is partially invariant to one of these makes it partially invariant to the other.

The evaluation of these models on the combined train and validation set is shown in Table D.2 in Appendix D. No model appears to outperform the baseline in terms of the criteria in Category A (postoperative pain prediction), C (change in index at the start of surgery), and D (correlation with fentanyl during bolus periods). The best model with an active adversarial head was when $\lambda_p = 0$ and $\lambda_d = 100$. Both the baseline model and this model were evaluated on the test set.

Table 7.7 gives the test set results for the baseline classifier. The ability of this model to predict postoperative pain was poor. It had a weak negative correlation with OpenIbis, particularly on an inter-patient level ($r = -0.284$), potentially indicating confounding with patient identity, hypnotic depth, or

Table 7.6: Patient identity and hypnotic depth probe performance for each adversarial configuration. Lower patient probe metrics indicate greater patient invariance; higher depth probe errors indicate greater depth invariance.

λ_p	λ_d	Patient identity probe			Depth probe	
		Acc.	Balanced acc.	F1	MAE	MSE
0	0	0.311	0.214	0.181	6.169	76.085
0.1	0	0.190	0.116	0.078	5.777	77.931
1	0	0.115	0.070	0.045	8.781	140.444
0	10	0.111	0.068	0.042	8.517	136.709
0.1	10	0.189	0.120	0.086	6.116	82.835
1	10	0.176	0.110	0.081	6.751	94.154
0	100	0.219	0.151	0.120	7.015	99.682
0.1	100	0.112	0.068	0.037	8.086	122.861
1	100	0.095	0.057	0.033	7.883	121.088
0	1000	0.269	0.182	0.148	5.626	72.333
0.1	1000	0.155	0.106	0.077	7.808	113.255
1	1000	0.129	0.082	0.054	5.975	78.824
0	10000	0.217	0.145	0.113	5.495	66.142
0.1	10000	0.264	0.200	0.165	6.984	96.257
1	10000	0.235	0.150	0.114	5.514	66.501

Table 7.7: Summary of test set results for baseline classifier with no adversarial training.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.543 (0.431–0.653)
AUROC > 3	0.482 (0.372–0.589)
AUROC > 6	0.492 (0.351–0.638)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.002 ($p = 0.003$)
Δ AUROC > 3	−0.011 ($p = 0.032$)
Δ AUROC > 6	0.000 ($p = 0.942$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.284$ ($p = 0.002$), Spearman $\rho = -0.297$ ($p = 0.001$)
B2. Intra-patient corr.	Mean Pearson $r = -0.045$, Mean Spearman $\rho = -0.065$
C. Start of surgery	
C1. Mean comparison	Pre: 40.794, Surg: 52.896, Diff: 12.101
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.122$, Mean Spearman $\rho = -0.108$
E. Explainability	
E1. Delta coeff.	−5.662 ($p < 0.001$)
E2. Alpha coeff.	2.123 ($p < 0.001$)
E3. Beta coeff.	−0.316 ($p < 0.001$)

both. There was a clear rise at the start of surgery of 12.1 units, and there was a weak but clear negative relationship with fentanyl effect-site concentration during bolus periods. The surrogate model suggested that the output of this model correlated negatively with delta and beta power, and positively with alpha power. This is the opposite of what would be expected according to the Category E criteria.

Table 7.8: Summary of test set results for discrete classifier with $\lambda_p = 0$, $\lambda_d = 100$.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.553 (0.448–0.661)
AUROC > 3	0.531 (0.419–0.639)
AUROC > 6	0.474 (0.315–0.640)
A2. Pain prediction (nested)	
Δ AUROC > 0	0.020 ($p = 0.010$)
Δ AUROC > 3	0.008 ($p = 0.242$)
Δ AUROC > 6	−0.002 ($p = 0.947$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = -0.071$ ($p = 0.457$), Spearman $\rho = -0.056$ ($p = 0.557$)
B2. Intra-patient corr.	Mean Pearson $r = -0.027$, Mean Spearman $\rho = -0.049$
C. Start of surgery	
C1. Mean comparison	Pre: 42.379, Surg: 49.805, Diff: 7.426
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.042$, Mean Spearman $\rho = -0.022$
E. Explainability	
E1. Delta coeff.	−26.471 ($p < 0.001$)
E2. Alpha coeff.	−7.273 ($p < 0.001$)
E3. Beta coeff.	−8.529 ($p < 0.001$)

Table 7.8 gives the test set results for the best adversarial classifier with $\lambda_p = 0$ and $\lambda_d = 100$. It was similarly poor at predicting postoperative pain to the baseline model. It did appear to have a weaker correlation with OpenIbis, although this correlation was still negative (rather than positive as might be expected from an index of nociception). It demonstrated a smaller rise at the start of surgery (7.4 units), and there was a weaker negative relationship with fentanyl. The explainability analysis associated an increase in this index with reduced power across delta, alpha, and beta bands. Figures 7.7 and 7.8 give examples of how the baseline index and the index with adversarial training behave on examples from the test dataset.

A GradCAM analysis was conducted on the adversarially trained classifier, with the results shown in Figure 7.9. The salient parts of the spectrogram appear to differ substantially between different segments, with no clear bias towards the beginning of the image being more salient. Two of the three examples that were analysed favoured lower frequencies, as would be anticipated, but this was much less consistent than the GradCAM analysis of the EEG-PD-NLME model.

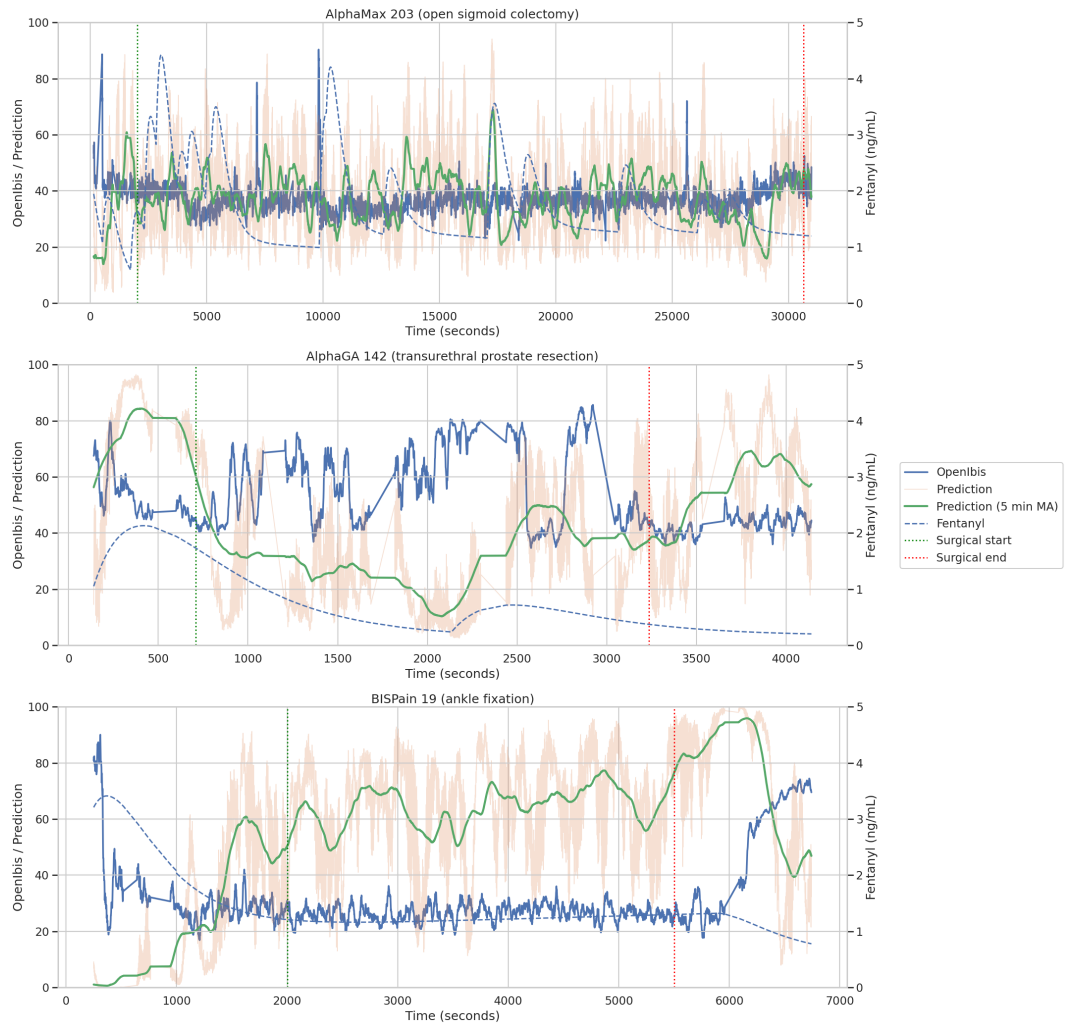


Figure 7.7: Baseline discrete classifier applied to three example EEG recordings from the test dataset.

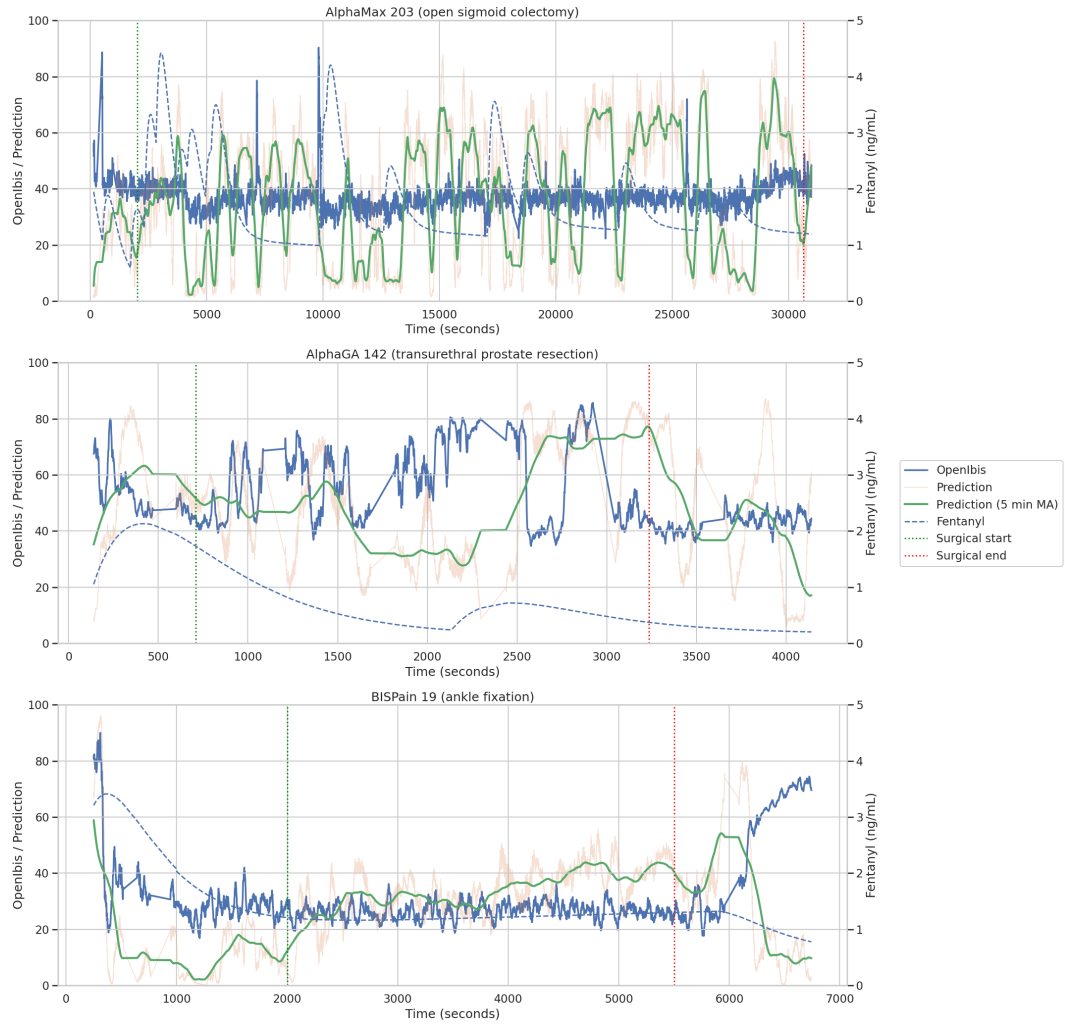


Figure 7.8: Discrete classifier with adversarial training ($\lambda_p = 0$, $\lambda_d = 100$) on the same EEG recordings.

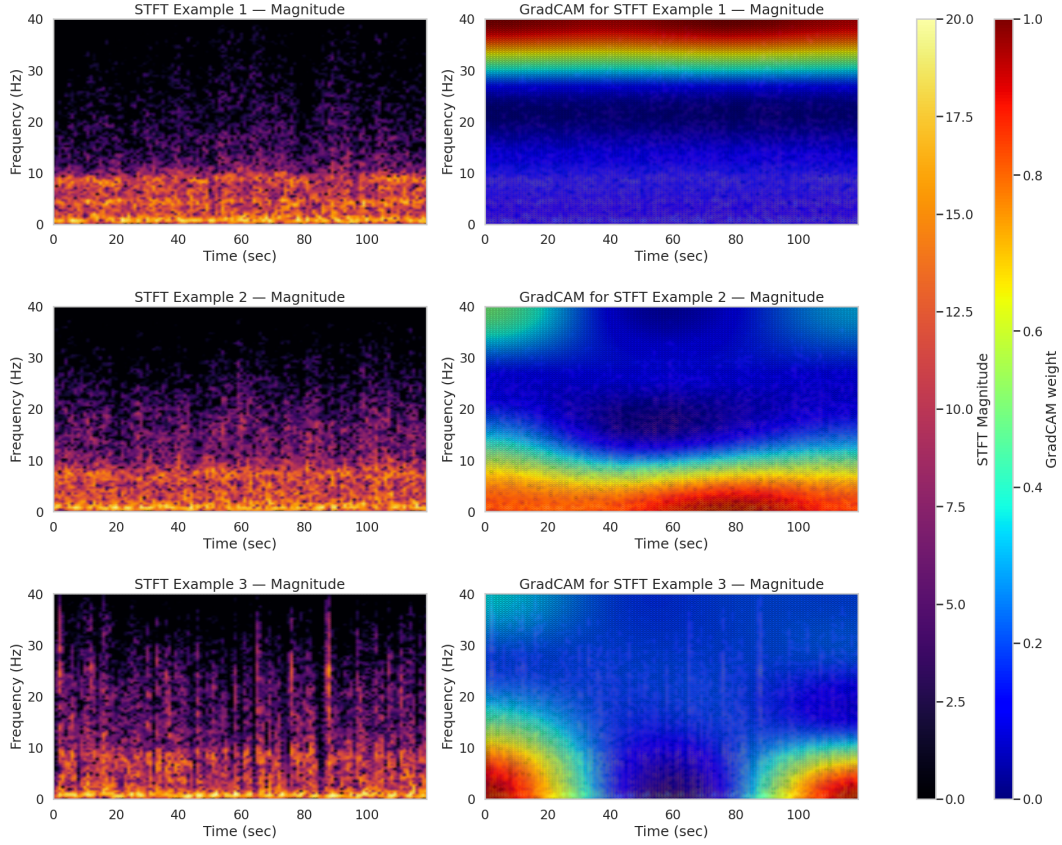


Figure 7.9: GradCAM visualisation of EEG spectrogram predictions with adversarially trained discrete classifier.

7.6 Hypnotic Invariance in the EEG-PD-NLME Model

7.6.1 Methods

Similar invariance techniques were used to create a partially hypnotic-invariant representation for the EEG-PD-NLME model. The basis of this model was the final model used with pre-surgical data (see 6.6). This model used the same pretrained ResNet-18 model with an input length of 120 seconds, and the additional pre-surgical dataset. Hyperparameters were identical to what was used in this previous experiment, with fentanyl MSE being used as the task loss function, and the proportional model being used with a learning rate of 1×10^{-3} , with no weight decay. Early stopping using $\bar{\rho}$ as a stopping metric was used. The balancing constant for the pre-surgical training was $\lambda_b = 0.5$.

The neural network f_ϕ used in the EEG-PD-NLME model was decomposed into an encoder e_{ϕ_e} and a task head t_{ϕ_t} . The encoder, being a ResNet-18, took a spectrogram x as input and output a 512-dimensional representation z . This was used as the input to the task head, which output the raw index of

nociception $\hat{n}$, which then went through the inverse PD model g_η as previously described.

The representation z was also used as an input to a linear adversarial head, h_{ψ_d} , which aimed to predict the 5-minute smoothed OpenIbis index (rescaled from 0–1) at time t (given that the spectrogram input x spans the time period from t to $t + 120$).

The previously employed approach of alternating gradient updates was used here. The first step was training the adversarial head, which used the mean squared error for the depth-of-anaesthetic measure d (in this case smoothed OpenIbis): $\mathcal{L}_d(d, \hat{d})$. This included adversarial training for both pre-surgical periods and bolus periods. This gradient update was as previously:

$$\psi_d \leftarrow \psi_d - \kappa \nabla_{\psi_d} \mathcal{L}_d$$

The gradient update for the encoder and task head was similar to previously, modified to take into account both the fentanyl MSE loss $\mathcal{L}_B$ and the pre-surgical loss $\mathcal{L}_{\tilde{B}}$, as well as the balancing constant λ_b . ϕ , which encompasses the weights of both the encoder and the task head in this context, was updated as follows:

$$\phi \leftarrow \phi - \kappa \nabla_{\phi} (\mathcal{L}_B + \lambda_b \mathcal{L}_{\tilde{B}} - \lambda_d \mathcal{L}_d)$$

The balancing constant λ_b was held at 0.5. The adversarial constant λ_d was varied, with $\lambda_d \in \{0, 0.1, 1, 10, 100, 1000, 10000\}$. For each value of λ_d , the previously mentioned Ridge regression probe was trained on the representation z to predict raw OpenIbis value, and this was evaluated on the validation set using MSE and MAE. Furthermore, each model was evaluated on the combined train-validation set using the unified validation framework. The best performing model was then evaluated on the test set.

7.6.2 Results

Table 7.9: Ridge probe MAE and MSE on OpenIbis predictions as a function of hypnotic depth adversarial weight λ_d .

λ_d	MAE	MSE
0	7.994	121.803
0.1	7.343	109.071
1	7.585	114.044
10	5.583	77.447
100	8.483	139.597
1000	8.762	149.901
10000	5.804	83.690

Table 7.9 gives the Ridge probe results for the NLME trained using an adversarial head with weight λ_d . The effect of the adversarial head was inconsistent, and the effect was small in the best case. MSE was highest with $\lambda_d = 1000$, with $\lambda_d = 100$ also performing well: these were the weights where hypnotic-invariance appeared to be partially achieved. The evaluation of each of these models on the unified validation framework is given in Table D.3. Of the two models where partial invariance was achieved, the best achieving model on the combined train and validation set was the model with $\lambda_d = 100$.

Table 7.10: Summary of test set results for EEG-PD-NLME model with $\lambda_d = 100$ adversarial head.

Metric	Result
A1. Pain prediction (index)	
AUROC > 0	0.388 (0.290–0.496)
AUROC > 3	0.458 (0.355–0.572)
AUROC > 6	0.438 (0.301–0.577)
A2. Pain prediction (nested)	
Δ AUROC > 0	−0.002 ($p = 0.963$)
Δ AUROC > 3	−0.016 ($p = 0.262$)
Δ AUROC > 6	−0.052 ($p = 0.159$)
B. OpenIbis	
B1. Inter-patient corr.	Pearson $r = 0.116$ ($p = 0.220$), Spearman $\rho = 0.102$ ($p = 0.283$)
B2. Intra-patient corr.	Mean Pearson $r = 0.092$, Mean Spearman $\rho = 0.037$
C. Start of surgery	
C1. Mean comparison	Pre: 46.250, Surg: 51.012, Diff: 4.763
D. Fentanyl bolus	
D1. Correlations	Mean Pearson $r = -0.056$, Mean Spearman $\rho = -0.067$
E. Explainability	
E1. Delta coeff.	14.617 ($p < 0.001$)
E2. Alpha coeff.	−13.945 ($p < 0.001$)
E3. Beta coeff.	9.486 ($p < 0.001$)

When this model was evaluated on the test set, with results shown in Table 7.10 and example plots shown in Figure 7.10, it was found that it was not significantly predictive of postoperative pain score. Correlations with OpenIbis were very small on an intra-patient level, but still moderate on an inter-patient level (although lower in magnitude than the EEG-PD-NLME model with no adversarial head). There was a small rise between the pre-surgical period and surgical period, and there were weak negative correlations on average between effect-site fentanyl concentration and the index. The largest association found in the surrogate model was a strong association between delta arousal and the index. There was a weaker association between alpha arousal and the index, and a weak association between beta dropout and the index.

When comparing this against the results from the EEG-PD-NLME with pre-surgical training, there was a reduction in the ability of the model to predict postoperative pain, and correlations with OpenIbis appeared to remain roughly the same as those in the previous model. However, this model now

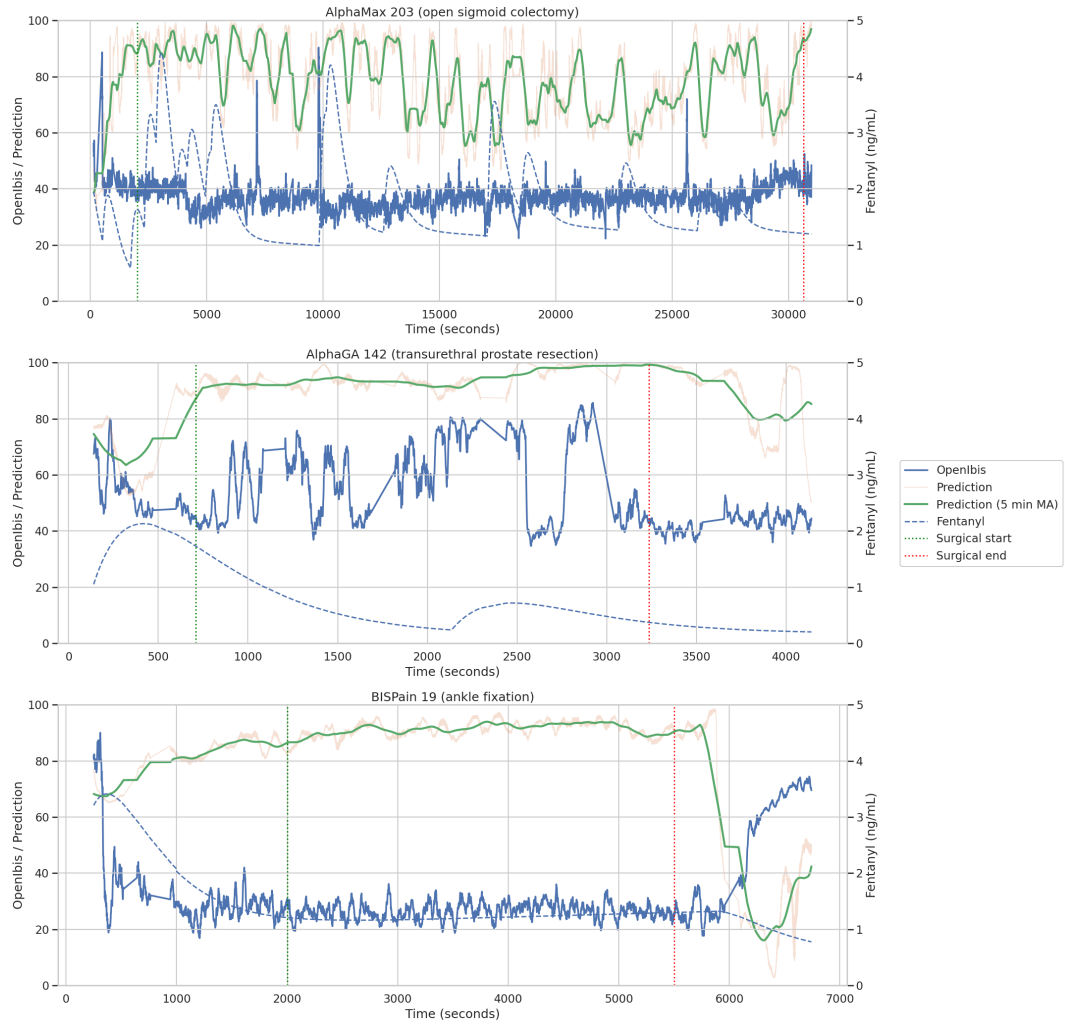


Figure 7.10: EEG-PD-NLME model with adversarial training ($\lambda_d = 100$) example plots.

appeared to perform well on the other metrics. It demonstrated a rise of 4.763 points from the pre-surgical period. It continued to demonstrate a weak negative correlation with fentanyl effect-site concentration during boluses, and the surrogate model coefficients were all in the expected direction.

7.7 Discussion

The main approach introduced in this chapter converted the fentanyl concentration C_e to a binary label y . The strength of this approach is that it may reduce the effect of titration noise on an intra-patient level. When C_e is below threshold, but is rising, this could be a sign of increased nociception relative to baseline rather than decreased nociception, as it would indicate a recent bolus of fentanyl (which may be motivated by increased surgical stimulus). Binarisation of C_e labels aimed to reduce the effect of sub-threshold changes in fentanyl, as $y = 0$ would only occur when fentanyl was above a threshold where nociceptive inputs can be usefully suppressed. There was one experiment conducted with soft y labels rather than hard binary labels, which performed poorly, potentially due to being less able to resolve the intra-patient drug titration paradox.

The results from experiments using binary labels were better, but the results were still inconsistent and poor for low C_f thresholds. At high C_f thresholds, the predictive ability improved, but this meant the model was increasingly resembling a model that classified whether a patient was in the pre-surgical or surgical period, which is likely to be a highly confounded endpoint. The threshold of $C_f = 1.5$ was therefore chosen, not due to predictive ability, but due to better performance on the unified validation framework.

There was a general trend for task metrics to worsen with increasing adversarial strength, while probe metrics also increased (indicating greater representation invariance). As the baseline model with $C_f = 1.5$ only had an AUROC of 0.5833, this limited the extent to which adversarial strength could be increased before completely degrading the predictive ability of the model.

There was a strong cross-invariance phenomenon where adversarial training against patient identity appeared to generate partial hypnotic depth invariance, and adversarial training against hypnotic depth appeared to generate patient identity invariance. This suggests that patient identity and hypnotic depth are strongly linked and likely share features.

No approach in this chapter achieved full adversarial invariance. Even as adversarial strength was increased to very high levels, there was eventually a limiting effect where the probe metrics reached a maximum that was below the expected maximum for a true random or mean prediction probe. Full adversarial invariance attempts to enforce orthogonality between the representation z and the adversarial head outputs ($\hat{d}$ or $\hat{p}$). However, in practice, full adversarial invariance cannot be achieved whilst maintaining reasonable performance on the task, as d and y are correlated (as has been previously discussed), and

y is partially dependent on p .

Part of this dependence is a natural result of autocorrelated signal. There will necessarily be similarities between adjacent portions of the signal, particularly given that in the dataset used in this experiment, there was a 60 second overlap between the adjacent 120-second windows. This means that there will always be more similarity within a patient than between patients, and the binarised outcome y cannot be entirely orthogonal to p .

The goal therefore became to enforce partial invariance in the representation z . From a causal perspective, the aim of partial invariance is to remove the effect of auxiliary backdoor causal pathways whilst maintaining the key causal pathways linking nociception and EEG (i.e. $n \rightarrow x$ and $n \rightarrow d \rightarrow x$). However, partially invariant representations do not necessarily preferentially remove the effect of auxiliary pathways whilst maintaining informative pathways: all pathways appear to be degraded by partial invariance (and degradation of the informative pathways leads to degradation of the quality of the index). For instance, if the test set results from the baseline classifier (Table 7.7) are compared to the results with adversarial training (Table 7.8), the Category C and D criteria are fulfilled more weakly in the model with partial invariance (while the Category B criteria are better fulfilled, reflecting less confounding by hypnotic depth). This trade-off highlights the central tension in utilising adversarial training to generate an index such as this.

Further development of a nociception index using similar techniques could more thoroughly explore causal inference methods for isolating the informative causal pathways. Other potential confounders could also be investigated. For instance, the use of the start of surgery to define the label y creates confounding by time, as the model could learn other temporal-based features, creating a temporally biased index. Time could therefore also be used as an adversarial target. The adversarial training methods used in this model only used linear adversarial heads and linear probes. It is therefore possible that the adversarial heads may cause nuisance variable information to be encoded in a way that is not linearly recoverable, meaning that linear heads and probes would underestimate the amount of nuisance information retained in z . Non-linear adversarial heads could be used to reduce any residual non-linear information.

Chapter 8

Discussion

8.1 Overview of Findings

This study had the aim of developing a unified validation framework for putative indices of nociception under general anaesthesia, and developing and applying ML approaches to generate an index of nociception.

To summarise the primary findings of each chapter:

- Chapter 2 featured a comprehensive literature review which established that no reliable, open-source index of intraoperative nociception currently exists. Existing indices either have been found to be poor at estimating nociception, such as qNOX [16, 140], or the algorithms are proprietary and are poorly detailed in the literature, such as PTI [141]. Confounding by factors such as hypnotic depth [16] and facial EMG [135, 140] reduce the usefulness and reliability of nociception models.
- Chapter 3 combined data from three studies: BISpain, AlphaGA, and AlphaMax. These data were then cleaned using both standard techniques and an EEG denoising transformer.
- Chapter 4 established the unified validation framework for nociception indices. It was shown that OpenIbis does not fulfil these criteria well. It was also shown that a direct predictor of postoperative pain scores does not fulfil the criteria well due to small sample size and confounding with the hypnotic dimension.
- Chapter 5 applied unsupervised learning techniques, and found that PCA could generate a potential index of nociception orthogonal to the hypnotic dimension. An index developed using HMM features was constructed, but it was found to be highly confounded by the hypnotic dimension.
- Chapter 6 established the novel EEG-PD-NLME model. It was found to recover a weak signal only. A ResNet-18 using an STFT representation of the EEG data was found to perform best.

- Chapter 7 employed adversarial training techniques on a simplified, binarised nociception target, and demonstrated that partial invariance could be achieved at the cost of reduced accuracy on the primary task. Adversarial training techniques were similarly employed on the EEG-PD-NLME model, leading to a better performing model.

8.2 Comparative Analysis

There were nine indices that were evaluated on the test set in this study:

1. OpenIbis (provided for comparison).
2. Direct pain score category prediction, trained using CORAL and a 1D CNN, with an input length of ten seconds in the time-domain (PSCP).
3. Principal component 6 from the PCA decomposition of a 60-second STFT spectrogram (PCA-6).
4. The inverted fractional occupancy of state 5 from the HMM (HMM-5).
5. EEG-PD-NLME using a ResNet-18 feature extractor using a 120-second STFT spectrogram (NLME).
6. EEG-PD-NLME, identical to above but with additional training on the pre-surgical period with $\lambda_b = 0.5$ (NLME+pre).
7. Discrete fentanyl / pre-surgical period classifier using a ResNet-18 feature extractor using a 120-second STFT spectrogram and a fentanyl cut-off of 1.5 ng/mL (Disc.).
8. Discrete fentanyl / pre-surgical period classifier, identical to the above but with an adversarial head for smoothed OpenIbis prediction with strength $\lambda_d = 100$ (Disc.+adv.).
9. EEG-PD-NLME, identical to (6) but with an adversarial head for smoothed OpenIbis prediction with strength $\lambda_d = 100$ (NLME+adv.).

Figure 8.1 gives the pairwise Pearson r values for each of the nine indices. This was calculated by aggregating all predictions across all patients in the test set for each index, and correlating them against each other. There does not broadly appear to be strong concordance between the indices, with most only having very weak positive or negative correlations with each other.

There were three pairwise correlations that were moderately strong and in the positive direction ($r > 0.3$), suggesting some concordance:

- PSCP and HMM-5 ($r = 0.47, p < 0.001$). This correlation is most likely because both indices have strong contamination from the hypnotic dimension. Both have moderately strong negative correlations with OpenIbis.

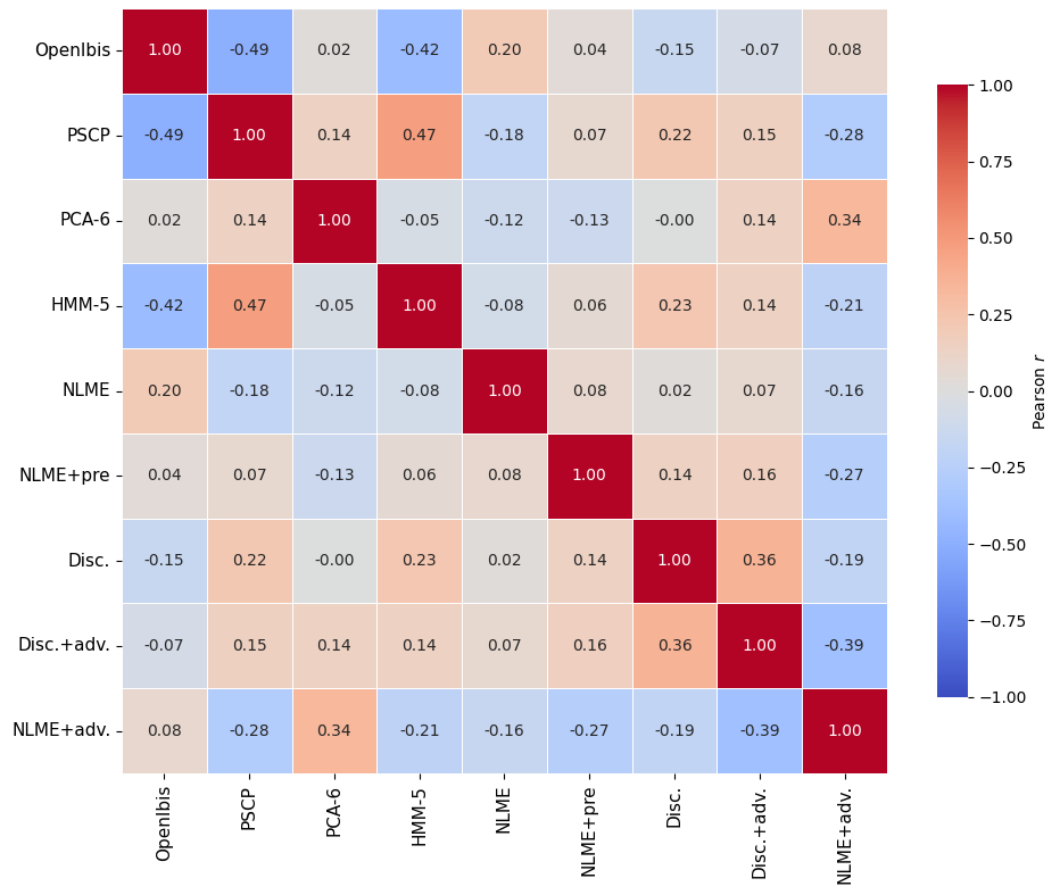


Figure 8.1: Comparison matrix of each of the evaluated indices, with pairwise Pearson r calculations.

- Disc. and Disc.+adv. ($r = 0.36, p < 0.001$). This correlation is reflective of the similar method used to construct these indices.
- PCA-6 and NLME+adv. ($r = 0.34, p < 0.001$). This correlation appears to be separate from the hypnotic dimension, as neither of these indices has a strong correlation with OpenIbis. These indices were not constructed using similar methods. It is possible that this correlation is an indication that these methods both measure similar aspects of nociception. Outside of postoperative pain score prediction, these two indices appear to be the best-performing on the unified validation framework.

There was also a moderate negative correlation between Disc.+adv. and NLME+adv. ($r = -0.39, p < 0.001$). The origin of this correlation is unclear.

8.3 Validation Criteria

There is no gold standard against which to evaluate nociception indices, so the unified validation framework was used. In this section, the performance of each of the models on each of the criteria is evaluated, with each of the evaluation tables restructured into per-criterion tables. This also serves as an evaluation of the criteria themselves, which were developed using clinical intuition and may not be optimal for the development of a nociception index.

8.3.1 Postoperative Pain Prediction (Category A)

Table 8.1: A1. Pain prediction (index): AUROC with 95% CI on the test set. Significant results where the entire CI is above 0.5 are in **bold**.

Model	AUROC > 0	AUROC > 3	AUROC > 6
(1) OpenIbis	0.381 (0.283–0.493)	0.444 (0.337–0.557)	0.526 (0.337–0.700)
(2) PSCP	0.621 (0.516–0.721)	0.571 (0.454–0.678)	0.572 (0.391–0.730)
(3) PCA-6	0.511 (0.405–0.618)	0.590 (0.479–0.700)	0.455 (0.318–0.598)
(4) HMM-5	0.635 (0.528–0.736)	0.523 (0.408–0.634)	0.548 (0.372–0.720)
(5) NLME	0.462 (0.357–0.569)	0.452 (0.345–0.567)	0.525 (0.371–0.686)
(6) NLME+pre	0.606 (0.504–0.706)	0.539 (0.429–0.649)	0.547 (0.382–0.715)
(7) Disc.	0.543 (0.431–0.653)	0.482 (0.372–0.589)	0.492 (0.351–0.638)
(8) Disc.+adv.	0.553 (0.448–0.661)	0.531 (0.419–0.639)	0.474 (0.315–0.640)
(9) NLME+adv.	0.388 (0.290–0.496)	0.458 (0.355–0.572)	0.438 (0.301–0.577)

Table 8.1 and Table 8.2 show the results of each model on the Category A measures (i.e. prediction of postoperative pain). There is no model that appears to be a good predictor of postoperative pain. The model that is specifically trained to predict postoperative pain category from a portion of EEG towards the end of surgery only achieved a weak significant result for prediction of pain score > 0 (AUROC: 0.621; 95% CI: 0.516–0.721), and no significant

Table 8.2: A2. Pain prediction (nested): ΔAUROC relative to a baseline linear model. Significant results at $p < 0.05$ with positive ΔAUROC are in **bold**.

Model	$\Delta\text{AUROC} > 0$	$\Delta\text{AUROC} > 3$	$\Delta\text{AUROC} > 6$
(1) OpenIbis	0.015 ($p = 0.210$)	0.015 ($p = 0.096$)	-0.012 ($p = 0.327$)
(2) PSCP	0.031 ($\mathbf{p} < \mathbf{0.001}$)	0.026 ($\mathbf{p} = \mathbf{0.002}$)	0.048 ($p = 0.075$)
(3) PCA-6	-0.012 ($p = 0.150$)	0.004 ($p = 0.311$)	-0.022 ($p = 0.281$)
(4) HMM-5	0.023 ($\mathbf{p} = \mathbf{0.022}$)	0.002 ($p = 0.198$)	-0.004 ($p = 0.891$)
(5) NLME	-0.001 ($p = 0.728$)	0.005 ($p = 0.505$)	-0.008 ($p = 0.075$)
(6) NLME+pre	0.005 ($p = 0.587$)	0.005 ($p = 0.665$)	-0.020 ($p = 0.313$)
(7) Disc.	0.002 ($\mathbf{p} = \mathbf{0.003}$)	-0.011 ($p = 0.032$)	0.000 ($p = 0.942$)
(8) Disc.+adv.	0.020 ($\mathbf{p} = \mathbf{0.010}$)	0.008 ($p = 0.242$)	-0.002 ($p = 0.947$)
(9) NLME+adv.	-0.002 ($p = 0.963$)	-0.016 ($p = 0.262$)	-0.052 ($p = 0.159$)

result at other thresholds. Most of the other models that had significant results for this criterion (PSCP, HMM-5) were highly correlated with OpenIbis. OpenIbis itself had a weak significant result, but in a negative direction, consistent with the previous observation that deeper anaesthesia appears to be associated with increased postoperative pain.

There were two models that fulfilled every other criterion besides pain score prediction: PCA-6 and NLME+adv. PCA-6 was weakly predictive of pain score when evaluated on the combined train and validation set, but became non-predictive when evaluated on the test set. Every other criterion besides the Category A criteria was consistent between the different splits. This suggests that these models were unable to recover a true signal that is usefully predictive of postoperative pain score. If these models are taken to be valid predictors of intraoperative nociception, as the other criteria seem to suggest, it raises the possibility that intraoperative nociception does not correlate well with postoperative pain score (or even has a slight negative correlation, as was observed with NLME+adv.).

This could firstly be because pain scores are a noisy, indirect measure of nociception. Pain itself is a highly subjective sensory and emotional experience. The quantification of pain onto an 11-point scale adds further subjectivity to the measure. Numerous non-surgical factors have been associated with postoperative pain. In addition to younger age and female sex (which were identified in this study), smoking, depression, anxiety, sleep disturbance, pre-existing pain, preoperative analgesic use, and higher body mass index have all been associated with increased postoperative pain scores [185]. Type of surgery also appears to be predictive of postoperative pain [186]. This may support a correlation between intraoperative nociception and postoperative pain: some surgeries have more intraoperative nociception than others, and therefore more postoperative pain. However, surgical type is also a highly confounded end point, as different types of patients have different surgeries and the characteristics that predict certain types of surgery may also predict postoperative pain perception. Furthermore, changes in intraoperative opioid

administration have been shown to have relatively little effect on postoperative pain [187].

Another potential cause of the lack of association between postoperative pain and intraoperative nociception would be analgesic titration that occurs during the emergence period. Emergence period EEG data were not used for postoperative pain score prediction in this study, as this data is often noisy, and patients are in a shallower hypnotic state than they are for the majority of their surgery. The period that was used for postoperative pain score prediction was the final fifteen minutes of surgery. Between the recording of this end-surgery EEG and the recording of a pain score upon emergence, there is ample opportunity for patients to be given more analgesia, adding more noise to the postoperative pain score.

The effect of drug titration also needs to be considered when analysing the relationship between nociception and postoperative pain scores. This could be illustrated by analogy with the negative association observed between deeper hypnotic status and higher postoperative pain. This is possibly due to a drug titration paradox-like effect: patients undergoing surgeries with higher expected nociceptive stimulus, or who exhibit signs thereof, will receive greater amounts of volatile anaesthetic. Once these patients are awake, they have greater postoperative pain despite having more anaesthetic, as they had greater nociceptive stimulus. A similar effect may occur with opioids: patients undergoing more painful surgeries receive greater levels of opioid, and intraoperatively this would lead to these patients having similar, or even lower nociception levels than patients who had less painful surgeries. However, once surgery is complete, the lingering effects of a more nociceptive surgery last longer than the effects of analgesia, and the disparity in nociception becomes apparent. This could result in a weakened or even reversed association between intraoperative nociception and postoperative pain scores. This is consistent with what was observed with PCA-6 and NLME+adv. Controlling for the type of surgery (based on expected nociceptive stimulus) may be helpful in mitigating this effect.

There is little evidence that nociception monitors, including those that use other non-EEG bio-signals, are able to predict postoperative pain reliably. One study investigated the use of the multimodal Surgical Pleth Index (SPI) and Nociception Level (NOL) index to predict postoperative pain before emergence, and found that they had AUROCs of 0.46 (95% CI: 0.31–0.61) and 0.66 (95% CI: 0.51–0.81) respectively for postoperative pain prediction [188]. A pooled analysis of multiple studies found that the SPI reached an AUROC of 0.72 (95% CI: 0.68–0.76) for prediction of postoperative pain. This is a stronger result (although still only moderate at best), but it is noted that no adjustment was made for confounders such as psychological stress, which may have biased the result [189].

Ultimately, postoperative pain scores appear to be a poor tool for training and validation of intraoperative nociception indices. Due to the heavy

confounding, it is possible that models trained to predict postoperative pain scores will learn patient characteristics and hypnotic depth rather than any useful index of nociception, and any index of nociception learned from postoperative pain scores could be complicated by noise and drug titration. In this study, postoperative pain scores were changed to an ordinal categorical variable to attempt to account for some of their noisiness. However, they were clearly still difficult to predict using any of the putative indices of nociception introduced in this study.

A more objective measure of postoperative pain is the quantity of opioid used postoperatively. It has been found that the use of non-EEG intraoperative nociception monitors can lead to a statistically significant but not clinically significant reduction in the amount of opioid used postoperatively [187]. This may be a less noisy metric to quantify postoperative pain, although it will still be affected by confounding factors such as patient identity and the drug titration paradox.

When evaluating nociception monitors, there may also be value in evaluating the downstream sequelae of uncontrolled postoperative pain. Factors such as the development of chronic pain, immune suppression, and heart strain could also be measured (although these too are likely to be highly confounded) [10].

8.3.2 OpenIbis (Category B)

Table 8.3: B1. OpenIbis inter-patient correlation: Pearson r and Spearman ρ between model output and OpenIbis score, aggregated across patients.

Model	Pearson r (p)	Spearman ρ (p)
(1) OpenIbis	1.000 ($p < 0.001$)	1.000 ($p < 0.001$)
(2) PSCP	-0.703 ($p < 0.001$)	-0.687 ($p < 0.001$)
(3) PCA-6	-0.083 ($p = 0.381$)	-0.072 ($p = 0.450$)
(4) HMM-5	-0.786 ($p < 0.001$)	-0.785 ($p < 0.001$)
(5) NLME	0.519 ($p < 0.001$)	0.526 ($p < 0.001$)
(6) NLME+pre	0.040 ($p = 0.676$)	0.018 ($p = 0.850$)
(7) Disc.	-0.284 ($p = 0.002$)	-0.297 ($p = 0.001$)
(8) Disc.+adv.	-0.071 ($p = 0.457$)	-0.056 ($p = 0.557$)
(9) NLME+adv.	0.116 ($p = 0.220$)	0.102 ($p = 0.283$)

The criteria established in Chapter 4 anticipate a weak positive correlation between an ideal monitor of nociception and OpenIbis. This is not what was observed for most of the indices developed in this study. For the inter-patient OpenIbis correlation, only the EEG-PD-NLME models demonstrated a positive correlation with OpenIbis. On an intra-patient level, only PCA-6, NLME, and NLME+adv. demonstrated an association in the expected direction. All other predictors either demonstrated a weak or strong negative association with OpenIbis.

Table 8.4: B2. OpenIbis intra-patient correlation: mean within-patient Pearson r and Spearman ρ between model output and OpenIbis score.

Model	Mean Pearson r	Mean Spearman ρ
(1) OpenIbis	1.000	1.000
(2) PSCP	-0.364	-0.234
(3) PCA-6	0.120	0.094
(4) HMM-5	-0.182	-0.187
(5) NLME	0.055	0.060
(6) NLME+pre	-0.047	-0.106
(7) Disc.	-0.045	-0.065
(8) Disc.+adv.	-0.027	-0.049
(9) NLME+adv.	0.092	0.037

As was previously discussed, there are multiple possible explanations. Confounding of pain scores and hypnotic depth is a likely contributor: the association between deeper hypnotic state and increased pain scores means that indices that are learned or selected using predictive ability on pain scores are more likely to be contaminated by the hypnotic dimension. The other explanation for this is that the true association between nociception and hypnotic depth is negative. This would be a consequence of the paradoxical increase in delta power that can be observed with nociception. Some of the models that demonstrate a negative association with OpenIbis have positive coefficients for delta power in the surrogate model for Category E: particularly PSCP. This lends credence to the idea that this may at least have a partial role to play in the observed negative associations.

However, particularly strong negative associations with hypnotic depth, such as those observed with PSCP and HMM-5, do not appear to be conducive to a useful index of nociception as evaluated by the other criteria. In terms of the other criteria, both of these indices largely behave as an inverted measure of hypnotic depth, with a positive association with postoperative pain, a rise in the index at the start of surgery, but positive associations with fentanyl effect-site concentration.

The adversarial training does appear to have some effect on reducing the size of the associations that models have with OpenIbis, but it does not fully remove these associations. It is unclear whether this actually improves the index, because it is unclear how much of these associations are driven by confounding. The adversarial training did not seem to improve the EEG-PD-NLME on this particular metric, although it did seem to improve the model across most other metrics.

The limitations of OpenIbis as a measure of hypnotic depth should also be commented upon. As previously mentioned, it is a noisy measure of hypnotic depth and better results for adversarial training were found when averaging it over five minutes. BIS and OpenIbis both utilise gamma band power, which overlaps with the range of facial EMG. This means that there is some contam-

ination of these indices with facial EMG: it has been observed that the BIS falls in fully awake subjects who have been paralysed due to the elimination of facial EMG [190]. OpenIbis is also a simplification of the BIS algorithm, designed for interpretability over faithful reconstruction. It has a high but not perfect correlation with the BIS ($r = 0.9395$), and it has not been validated as a measure of hypnotic depth as extensively as the BIS has.

Correlation with OpenIbis (or any other measure of hypnotic depth) is an essential part of comprehensively evaluating an index of nociception, as it allows one to determine whether the putative index is actually primarily an index of hypnotic depth. Indices with strong correlations with hypnotic depth do not appear to be useful as nociception indices. Some amount of correlation with OpenIbis is acceptable, but it is unclear whether this correlation should necessarily be a positive one, as previously hypothesised.

8.3.3 Start of Surgery (Category C)

Table 8.5: C1. Start of surgery: mean model output in the pre-surgical and surgical periods, and their difference (Surg.–Pre.). Any results where the difference is greater than 5 points are in **bold**.

Model	Pre.	Surg.	Diff.
(1) OpenIbis	40.366	38.825	−1.541
(2) PSCP	52.510	58.044	5.535
(3) PCA-6	41.788	50.928	9.140
(4) HMM-5	54.761	57.720	2.959
(5) NLME	56.929	50.284	−6.645
(6) NLME+pre	52.128	49.509	−2.619
(7) Disc.	40.794	52.896	12.101
(8) Disc.+adv.	42.379	49.805	7.426
(9) NLME+adv.	46.250	51.012	4.763

It is intuitive to expect a sizeable rise in any reasonable index of nociception between the pre-surgical and surgical period. Table 8.5 shows the change in index between these periods for each of the tested indices. A rise was observed in most of the indices, with the largest being observed in Disc., PCA-6, and Disc.+adv. The only indices (besides OpenIbis) that had a fall associated with the start of surgery were NLME and NLME+pre.

A rise in this index does not, however, necessarily indicate that the model is responding to cortical manifestations of nociception. OpenIbis falls slightly between the two periods, potentially as a result of the pre-surgical period being closer to induction. In some of the indices that are highly negatively correlated with OpenIbis (PSCP and HMM-5), a rise would be anticipated here. Other causes of a rise could be confounding by time, facial EMG contamination in response to nociceptive stimuli, or cardiovascular artefacts.

A further difficulty with this criterion is that the pre-surgical period is not necessarily devoid of any nociceptive stimulus, as this evaluation criterion (and multiple models) assume. Although efforts were made to avoid the nociceptive effect of intubation by defining the pre-surgical period as commencing only three minutes after intubation, it is possible that some of the nociceptive effect of intubation will persist. There are also other noxious stimuli that occur prior to incision, such as repositioning of the patient, insertion of catheters, or intravenous cannulation. Patients may be in pain before surgery. These factors may explain why the size of the rise at the start of surgery may not be as large as anticipated [191].

This criterion is ultimately a strong one, as it provides evidence of the index’s ability to respond to nociceptive surgical stimuli. In datasets with more information about intraoperative events, this criterion could be extended to investigate whether prospective indices respond to other intraoperative events.

8.3.4 Fentanyl Bolus Correlations (Category D)

Table 8.6: D1. Fentanyl bolus: mean within-patient Pearson r and Spearman ρ between model output and fentanyl effect-site concentration following bolus events. Any correlations < -0.05 are in **bold**.

Model	Mean Pearson r	Mean Spearman ρ
(1) OpenIbis	-0.258	-0.248
(2) PSCP	0.105	0.090
(3) PCA-6	-0.155	-0.149
(4) HMM-5	0.049	0.055
(5) NLME	-0.032	-0.063
(6) NLME+pre	-0.047	-0.106
(7) Disc.	-0.122	-0.108
(8) Disc.+adv.	-0.042	-0.022
(9) NLME+adv.	-0.056	-0.067

As an analgesic, an increasing effect-site concentration of fentanyl should be associated with a reduction in nociception index. Table 8.6 summarises these results. All the indices follow this correlation, except for PSCP and HMM-5, which appear to have weak positive correlations with fentanyl. This is potentially due to the aforementioned confounding with hypnotic status.

Of the other indices, the strength of the correlation is typically weak at best. The indices with the strongest negative correlations are PCA-6, Disc., and NLME+adv.

The weaknesses of this criterion are that these correlations assume that the background nociceptive stimulus (N_b) is constant throughout the bolus period, and that anaesthetists will always give fentanyl reactively in response to stimuli, rather than proactively. This is not necessarily the case. For instance, if an anaesthetist gives a bolus of fentanyl in anticipation of a particularly

strong nociceptive event, and the event happens while the effect-site fentanyl concentration is rising, this could result in a positive correlation between nociception and fentanyl effect-site concentration for that bolus. It would also be anticipated that the correlation between the index and the fentanyl effect-site concentration would be minimal when the level of nociceptive stimulus is low. The result of these assumptions is that the mean correlation coefficients are likely to indicate weak correlations.

Overall, this criterion is also an important one to have as part of the validation framework. Further development of this part of the validation framework could investigate the way that the bolus-associated periods are defined, or how different intraoperative analgesic agents (including hypnotic agents) interact with each other.

8.3.5 Explainability Analysis (Category E)

Table 8.7: E. Explainability: regression coefficients for delta, alpha, and beta band power in a linear model predicting model output from EEG spectral features. All coefficients significant at $p < 0.001$.

Model	E1. Delta coeff.	E2. Alpha coeff.	E3. Beta coeff.
(1) OpenIbis	-1.786	-3.875	5.962
(2) PSCP	12.639	19.650	0.536
(3) PCA-6	8.888	-4.910	12.928
(4) HMM-5	1.414	6.134	-0.367
(5) NLME	-14.430	-8.401	0.853
(6) NLME+pre	-0.258	6.253	0.580
(7) Disc.	-5.662	2.123	-0.316
(8) Disc.+adv.	-26.471	-7.273	-8.529
(9) NLME+adv.	14.617	-13.945	9.486

The coefficient results from the surrogate models are given in Table 8.7. These are markedly inconsistent, with all coefficients varying in sign and magnitude across indices. Although the typical expected pattern associated with nociception is beta arousal, paradoxical delta arousal, and alpha dropout, these findings are inconsistent and potentially variable at different levels of hypnotic depth. Beta arousal is typically seen at shallower hypnotic depth, whereas delta arousal and alpha dropout are typically seen at deeper hypnotic states [12]. Furthermore, the statistical significance of these coefficients is likely overestimated. They result from aggregating time points across patients, which creates highly autocorrelated and non-independent data points.

The only indices that had the exact expected nociception pattern were PCA-6 and NLME+adv. The basic NLME index had a pattern more typical of a measure of hypnotic depth, with delta dropout rather than delta arousal. The other indices had patterns that were more inconsistent.

Table 8.8: Regression coefficients for delta, alpha, and beta band power in a linear model predicting model output from EEG spectral features, stratified by OpenIbis level. All coefficients significant at $p < 0.001$ unless otherwise noted.

Model	High OpenIbis			Low OpenIbis		
	E1. Delta	E2. Alpha	E3. Beta	E1. Delta	E2. Alpha	E3. Beta
(2) PSCP	14.997	25.212	3.651	13.861	19.964	-12.813
(3) PCA-6	12.988	4.888	10.601	6.292	-12.554	22.315
(4) HMM-5	-7.027	3.101	-0.861	2.390	9.300	-20.039
(5) NLME	-15.745	-6.099	-2.148	-15.747	-11.976	6.924
(6) NLME+pre	-4.892	2.670	1.286	-2.308	7.921	-3.149
(7) Disc.	-16.067	-6.798	-4.280	0.028 [†]	7.098	6.634
(8) Disc.+adv.	-25.951	-5.713	-10.493	-27.700	-9.616	-2.102
(9) NLME+adv.	13.265	-15.162	8.479	17.286	-11.868	13.735

[†] $p = 0.724$ (not significant).

Another regression was performed, stratifying the power bands by hypnotic depth to determine the effect of hypnotic depth on the relationship between band power and each index. For this regression, the normalised band-powers were taken, and split into six interaction predictors: delta power, alpha power, and beta power where OpenIbis value was high (above the mean of 40.7), and the same three measures at low OpenIbis (below the mean). The coefficients of this regression are shown in Table 8.8.

Most band power findings were consistent whether the OpenIbis was high or low. For PCA-6, the alpha depression associated with an increase in the index was only present in a deeper hypnotic state, whereas in a shallower hypnotic state, an increase in the index was associated with an increase in alpha power. It is known that nociception-related alpha dropout is more pronounced at deeper anaesthetic depth [12]. NLME+adv. was consistent across both categories, with the expected pattern of delta arousal, alpha dropout, and beta arousal present both when OpenIbis was high and low.

Other explainable EEG features could be included as part of the surrogate model (or otherwise correlated against model outputs) in order to provide further explainability. Nociception has been linked to changes in cross-frequency modulation, which could be analysed to provide further explainability to these models [112].

Although the surrogate modelling and explainability is useful for improving our confidence in these models, the expected findings are still debatable. This criterion should therefore not be used as the deciding factor for whether a model is useful.

8.3.6 Validation Criteria Summary

To summarise, the most valuable validation criteria appear to be the Category C criterion of a rise at the start of surgery, and the Category D criterion of responsiveness to fentanyl effect-site concentration. These criteria had the most consistency across the attempted approaches. The Category B (OpenIbis) and Category E (explainability) criteria are also valuable, but the expected correlations are more debatable. Finally, good performance on the Category A criterion (postoperative pain prediction) was not achievable with any approach, including the PSCP designed specifically for prediction of postoperative pain. It is a highly confounded and noisy endpoint and the small differences observed between different indices should not be the deciding factor.

There was no clear weighting given to the different factors in the validation framework, making it difficult to objectively rank or compare indices. However, any weighting of the factors would be somewhat arbitrary and based on clinical intuition, as there is no objective ground truth measurement of nociception. The use of multiple criteria allows for the framework to be robust to confounding: no amount of fulfilment of the Category C criterion, for instance, could justify an index that had a positive correlation with fentanyl effect-site concentration.

Depending on data availability, other potential criteria could be used to augment the validation framework. Other intraoperative noxious stimuli such as surgical insufflation could be timed and correlated with indices to determine if the index is responsive. Indices could also be validated by correlating with haemodynamic parameters: one would anticipate blood pressure and heart rate would rise with an increase in nociception.

8.3.7 Model Summary

Of the models, there were two that fulfilled all criteria besides postoperative pain prediction: PCA-6 and NLME+adv. Both had a weak positive correlation with OpenIbis on an intra-patient level. Both had a rise at the start of surgery, with PCA-6 having a larger rise. Both had a negative mean correlation with fentanyl effect-site concentration across all boluses, albeit weak. Both had a surrogate regression model with a positive coefficient for delta and beta power, and a negative coefficient for alpha power.

Either of these models could be investigated further for potential clinical use. PCA-6 is a much more simple model that reflects the underlying structure of EEG data and avoids the complications and confounders involved with supervised training. NLME+adv. is much more complex, but performs similarly, and attempts at adjusting for the confounders are inherent in its structure.

8.4 Requirements for an Index of Nociception

There were several principles learned over the process of constructing the indices in this study. These principles could be applied to further attempts at constructing EEG indices of nociception, or they could be applied to other attempts at generating indices from EEG or other bio-signals.

Firstly, it is important that there is some time-evolving signal to use as a starting point for learning. The PSCP attempted to learn a signal from a static label (the postoperative pain score), which resulted in a highly confounded and poorly performing signal. The unsupervised methods used the signal itself as a basis for index discovery. The EEG-PD-NLME and discrete classifier models used the effect-site concentration of fentanyl as the basis for index development.

The next important consideration is the importance of accounting for confounders. The key confounders in this study were hypnotic depth, patient identity, and clinician behaviour (i.e. the drug titration paradox). The strongest models were the ones that were able to account for all three of these confounders. In the PCA analysis, the first principal component captured much of the hypnotic depth variation, and it is likely that components 2–5 captured many of the patient-specific factors. The drug titration paradox was not of concern as the index was not constructed using drug concentrations. The NLME+adv. model also accounted for these confounders. Patient-specific factors were captured by the random effects in the model. Hypnotic depth was partially adversarially trained against, and the drug titration paradox was addressed by training only on bolus-associated segments. The weakest models, such as HMM-5, had no mechanism for accounting for these confounders.

In building an interpretable index, it is also important to consider calibration of the output. Many of the raw model outputs in this study had arbitrary scaling, so for consistency, a monotonic CDF transform was applied. This led to a uniformly distributed output distribution ranging from 0 to 100, where the output index corresponds to a percentile relative to the distribution established by the combined train and validation set. This calibration technique is closely related to more standard calibration techniques such as isotonic regression [192], although rather than learning conditional label probability, the CDF transform uses only the predicted model output.

The most successful models in this study had input lengths of at least 60 seconds. This could be due to patterns of EEG changes in response to nociception that persist over this time frame, aligning with animal models that suggest nociception responses exist over a time frame of 90 seconds [113]. In constructing a nociception index, there is clear value in having model inputs with extended time scales such as this.

8.5 Limitations

8.5.1 Dataset

The dataset in this study had limited information regarding surgical stimuli levels, besides standardised markers of events such as intubation and surgical start time. Most other indices that use ML to predict nociception under general anaesthesia use standardised stimuli as a ground truth [138, 13]. In this study, there was little information about stimuli (and even if there was, there is no obvious way to weight different types of stimuli and account for the differences between different procedures, patients, and surgeons). As a result, a variety of novel methods had to be employed to develop an index. There are also many variables that were not taken into account that may have an effect on EEG, such as temperature [193], cerebral blood flow [194], and any of the many other drugs that are administered during surgery, including paralytic drugs [190]. Any of these variables have the potential to be confounders.

The dataset also lacked many patient-level factors, such as ethnicity and comorbidities. The type of surgery each patient had was available in the dataset, but was not used in this study. Other surgery-level factors, such as anaesthetist, surgeon, and placement of electrodes would all have had an effect on the EEG.

The EEG in this dataset was single-channel EEG. While this does come with the advantage of easy clinical deployability, it does have several limitations. There is a loss of spatial information, which means that sources cannot be localised [104] and conventional microstates cannot be confidently identified [99]. Conventional blind source separation techniques such as ICA cannot be used and signal-to-noise ratio is lower [78]. Single-channel EEG is less robust to electrode placement, adding another potential confounding variable.

Another confounder, which is likely present in this data but was not fully accounted for, is facial EMG. The data were filtered such that the maximum frequency was 40 Hz, but this does not necessarily remove all the EMG contamination, as EMG does extend into the measured gamma band [110]. The EEGdenoiseNet-cleaned dataset, which is trained to remove EMG contamination, was used in experiments in Chapter 4 (Table 4.5) and Chapter 6 (Table 6.4). In both these cases, the experiments that used EEGdenoiseNet-cleaned data appeared to have inferior performance to models trained on conventionally cleaned EEG data. This is a potential indication that models may have been relying on EMG contamination. The other potential indication as to whether models are using facial EMG is through the explainability techniques. The loadings in Figure 5.5 are primarily concentrated below 15 Hz, indicating that facial EMG is unlikely to be an important feature for PCA-6. The GradCAM map shown in Figure 6.5 seems to suggest that lower frequencies are more salient for the EEG-PD-NLME model, again suggesting that facial EMG contamination is unlikely to be an important feature. The GradCAM map generated for the discrete classifier model in Figure 7.9 is more variable,

and appears to show saliency of high frequencies in Example 1. This could be explained by the model learning facial EMG contamination as a correlate of nociception.

Other artefacts were cleaned from the dataset during the data preparation pipeline, leading to some missingness of the EEG signal. There are also portions of the signal that were missing in the raw EEG. Although this is a relatively small portion of the signal, the missingness is not at random (MNAR), meaning that this could bias the indices generated from this data. The mean missing signal proportion per recording was 0.32%, so it is assumed that the impact on the results would be minimal.

The data were collected primarily from one site (Waikato Hospital), which may limit generalisability of the models to different sites. Other sites may use different hardware to record EEG, and may have different anaesthetic or surgical practices leading to different data distributions.

The observational manifold of this dataset is limited to cases that were selected as appropriate for the three constituent studies that made up this dataset. This is likely to favour routine cases with reasonable analgesic control. This would potentially limit the accuracy and usefulness of these models for edge cases outside the observational manifold of this dataset, such as cases with extremely high nociceptive input and poor nociceptive control.

8.5.2 Modelling

The problem of inferring nociceptive state from EEG is potentially an ill-posed inverse problem. For well-posedness, inverse problems must fulfil the three Hadamard conditions: existence (a solution exists for every possible input), uniqueness (solutions are unique), and stability (small changes in input do not produce large changes in output) [195]. The forward process generating EEG segment x depends not only on nociceptive status n , but also several nuisance variables: hypnotic depth d , patient identity p , and other unobserved phenomena ξ . This gives a mapping $(n, d, p, \xi) \rightarrow x$. It is plausible that, given an observed EEG segment x , multiple values of (n, d, p, ξ) could explain the generated EEG x : changes in alpha power could be due to nociception, they could be due to changes in hypnotic depth, or they could be a result of patient characteristics. This may make n partially non-identifiable.

Removing the influence of patient factors and hypnotic depth by using adversarial techniques can help with confounding, but it will not solve a non-identifiable inverse mapping. Regularisation and the imposition of priors can be useful to identify solutions when problems are fundamentally ill-posed [195]. The EEG-PD-NLME model imposes a prior structure matching the pharmacodynamic Hill model. This model’s inductive bias aims to make n more identifiable.

The other key assumption that this model makes is that the relationship between x and n is stationary. The models that were used are not time-aware

models: for any given EEG signal x , the predicted $\hat{n}$ will always be consistent regardless of the time that the prediction was made, or what the previous predictions were. A more time-aware model may be able to make use of any temporal patterns or autoregression present in $n(t)$ to create better estimates of nociception.

8.5.3 Validation Framework

The unified validation framework introduced in this study was used due to the lack of ground truth nociception data. This was based on a combination of clinical intuition and assumptions made from prior nociception research. The framework was developed alongside the nociception prediction models produced in this study, and the models themselves were used to validate the framework. This does produce a somewhat circular chain of validation between the models and the validation framework. However, due to the lack of ground truth, it is impossible to create a fully reliable and grounded validation framework. Full prospective validation studies would be required to provide well-grounded evidence of the value of an index.

Subgroup analyses were not conducted as part of this study. It is noted that there are differences in EEG by age [101], and there are heritable differences in EEG [100]. This is in addition to the fact that different demographic groups will be subject to different distributions of procedures and differences in anaesthetic care, and therefore different distributions of nociception and analgesic provision. It would therefore be reasonable to evaluate models on different demographic subgroups, including by age, sex, and ethnicity. Due to the small patient-level sample size of the test and validation sets, as well as a lack of demographic information, subgroup analyses could not be completed as part of this study.

8.5.4 Causality

The primary confounders in this study were hypnotic depth, the drug titration paradox, and patient identity. Although efforts were made to mitigate the effects of these confounders, their effect could not be eliminated entirely from most models.

Hypnotic depth specifically is difficult to remove from these models due to its correlation and causal connections to nociception. As discussed in Chapter 7, the relationship between nociception and EEG that this study was attempting to elucidate is composed of two causal pathways: the direct effect of nociception on EEG, and the indirect effect of nociception via effect on the hypnotic dimension. This leads to a natural decomposition of a time-varying nociception index into two parts:

$$n(t) = n_d(t) + n_{\perp}(t)$$

$n_d(t)$ denotes the component of nociception that is correlated with the hypnotic dimension (formally, it is the projection of n onto the dimension spanned by the hypnotic measurements d). $n_{\perp}(t)$ denotes the component of nociception that is orthogonal to the hypnotic dimension.

As the hypnotic dimension is almost fully captured by the first principal component, the signal found in PCA-6 is an estimation of $n_{\perp}(t)$ rather than $n(t)$. NLME+adv. estimates $n(t)$, although the adversarial training partially attenuates the $n_d(t)$ component. NLME+adv. additionally has some confounding from backdoor causal pathways, due to the training only enforcing partial invariance to hypnotic depth.

The other backdoor pathway that was impossible to control for was the direct effect of fentanyl on both EEG and nociception. This is especially difficult given that fentanyl effect-site concentration was used as an indirect marker for nociception levels in most of the models developed in this study. It is unclear how this confounding could be addressed.

As previously mentioned, this model is not time-aware, and no models in this study have any measures to mitigate time bias. Time itself could be a relevant confounding factor in these models, as multiple intraoperative factors such as anaesthetic concentration, surgical stimulation, and levels of noise in the EEG would be expected to change over time.

Another causal factor that may warrant more comprehensive treatment would be the anaesthetist behaviour. It is particularly relevant in this study, as two of the three datasets were from RCTs where a subset of the anaesthetists were instructed to titrate the levels of anaesthesia in response to certain EEG parameters [147, 149]. In BIS Pain, the anaesthetists titrated levels of drug to achieve different target BIS values [147]. In AlphaMax, the patients in the intervention group had anaesthetic titrated to maximise levels of oscillatory alpha power. The relationship between fentanyl and alpha power in this study was complex, varying between a positive and a negative association in the early stages of each surgery, before converging towards a null relationship in the later stages [149]. It is therefore unlikely that this titration behaviour would have had a meaningful impact compared to standard intraoperative titration.

Chapter 9

Conclusion

9.1 Contributions

This study addressed the problem of estimating nociception levels in patients under general anaesthesia, with the eventual aim of developing a clinically usable monitor that could be used for intraoperative titration of analgesia. Previous attempts at estimating nociception levels involved opaque proprietary models, and previous attempts at evaluating putative nociception indices were inconsistent and often based on prediction of postoperative pain scores. There are two key contributions of this study, developed and evaluated in concert to fill these research gaps: a unified validation framework, and novel ML modelling approaches for nociception estimation.

The unified validation framework consisted of five categories: postoperative pain prediction, correlation with hypnotic depth, response to surgical incision, response to fentanyl concentration, and an explainability analysis. Using this framework, it was found that no model could predict postoperative pain well, even when satisfying other criteria. This suggests that nociception at the end of surgery may not correlate well with postoperative pain. This finding challenges the existing assumption that the level of nociception at the end of surgery correlates strongly with postoperative pain scores, and therefore challenges the use of postoperative pain scores for evaluation of intraoperative nociception indices. If further validated, this finding would have strong implications for how future nociception indices are trained and evaluated.

Multiple modelling approaches were used to generate nociception indices. From a modelling perspective, the EEG-PD-NLME model is the most substantial contribution of this study. It combines conventional ML feature extraction and PD modelling, and could be adapted for other inverse problems in the medical domain. Potential applications could include determining the effects of drugs from continuous physiological monitors, or serial radiology or pathology findings. The other major methodological novelty was adversarial training for hypnotic depth invariance, which was shown to have a weak but consistent effect on confounding by hypnotic depth.

The best-performing indices on the test set were principal component 6 from the PCA, and the EEG-PD-NLME with pre-surgical training and adversarial training against hypnotic depth. These two indices were moderately correlated with each other, and they fulfilled all criteria in the framework except for prediction of postoperative pain. The consistency of these indices with each other and all the other criteria provides evidence that intraoperative nociception may not correlate well with postoperative pain.

The framework and the learned models demonstrated methods of learning and evaluating physiological signals in the absence of gold-standard labels. This was achieved by exploiting the structure that exists within EEG signals, and indirect pharmacological correlates.

9.2 Future Work

Of the two models, PCA-6 has the most potential to take forward for further validation. PCA-6 is an inherently interpretable measure and therefore would be preferred over the more complex EEG-PD-NLME model (which incorporates a largely uninterpretable ResNet-18 encoder). Given that the feature loadings are largely invariant over the time dimension of the input, it could be simplified to a pure frequency-domain measure using a power spectrum averaged over 60 seconds.

If intraoperative nociception and postoperative pain are not well-correlated targets, as this study has suggested, this has many implications for the evaluation of potential nociception indices going forward. The framework provided in this study, in particular the criteria from Categories B–E, could serve as a consistent and comprehensive basis for evaluating future indices of nociception, and the performance of the PCA-6 model provides an interpretable baseline for this. Many of the models in this thesis could be further optimised with more comprehensive hyperparameter tuning, and the unified validation framework could be used for this.

Further improvements to the models could be made by considering other data modalities. Multiple-channel EEG provides cleaner data and spatial localisation, which may be able to better identify biomarkers of nociception. This would be particularly useful for identifying EEG microstates. Indices such as SPI and NOL are multimodal and use other physiological signals to quantify nociception. There is therefore potential to combine EEG features with other bio-signals, such as skin conductance, heart rate, blood pressure, and pulse oximetry.

The models in this study used a monotonic CDF transform. This preserves order and ensures the indices are roughly uniformly distributed between 0 and 100. However, these scores are not inherently physiologically interpretable, and it is not clear at what score reasonable nociception control is achieved. Improved calibration techniques could be investigated to address this.

As has been previously mentioned, potential indices such as PCA-6 could

be evaluated further by gathering information on common intraoperative noxious stimuli such as peritoneal insufflation (inflation of the abdominal cavity with gas during laparoscopic surgery). Indices should show a rise in nociception index in response to noxious intraoperative events. Studies could also be conducted looking at the response of these indices to standardised stimuli (such as tetanic stimuli).

Ultimately, the only way that the findings in this study can be validated is through a prospective trial. Such a trial would be an RCT where the study group receives analgesia titrated to achieve better pain control according to one of the potential indices of nociception constructed in this study. If titrating analgesia based on this index shows a decrease in use of opioids (intraoperatively or postoperatively), or a decrease in postoperative pain scores, that would serve as evidence that the index is a genuine marker of nociception. This would serve as a ground truth for validating the other conclusions of this study — most importantly, the notion that postoperative pain does not correlate strongly with intraoperative nociception.

Appendix A

Unsupervised Learning Supplementary Material

This section contains the raw tables for the exploration of PCA and EFA components, as well as the exploration of HMM statistics. The three primary criteria considered were Category A metrics (prediction of postoperative pain, both by AUROC of index alone, and AUROC when combined with demographic data), Category C metrics (change in the index between the pre-surgical period and surgery), and Category D metrics (correlation with fentanyl).

Components with results operating in the expected direction for an ideal index are indicated in **bold**, and components with results operating exactly opposite to the expected direction are indicated in ***bold and italics*** (these indicate the component may still be a valid index with the sign reversed). This includes any significant AUROC or any significant change to AUROC, any absolute surgical change in the index of greater than five, and any r or ρ greater than 0.05 when taking the correlation of the component with fentanyl effect-site concentration.

Table A.1: Complete exploration for PCA components 1–10 for one-second STFT.

Component	A1. AUROC (95% CI)			A2. ΔAUROC (p)			B1. Between-patient OpenBis corr.		B2. Within-patient corr.		C. Surgical change			D. Fent bolus corr.		E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson ρ (p)	Spearman ρ (p)	Mean r	Mean ρ	Presurg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
Comp. 1	0.501 (0.447, 0.552)	0.497 (0.439, 0.551)	0.450 (0.371, 0.525)	0.001 (0.492)	0.000 (0.744)	−0.003 (0.295)	0.438 (<0.001)	0.437 (<0.001)	0.290	0.276	53.607	49.197	−4.410	−0.041	−0.035	−2.756	10.723	15.181
Comp. 2	0.586 (0.536, 0.640)	0.568 (0.513, 0.624)	0.509 (0.427, 0.585)	0.005 (0.121)	0.002 (0.379)	−0.002 (0.407)	−0.127 (0.007)	−0.091 (0.055)	−0.044	0.056	52.497	51.229	−1.258	0.025	0.026	−6.457	18.290	7.318
Comp. 3	0.530 (0.477, 0.585)	0.492 (0.440, 0.547)	0.492 (0.419, 0.566)	0.004 (0.146)	0.000 (0.971)	−0.001 (0.904)	−0.266 (<0.001)	−0.257 (<0.001)	−0.125	−0.109	39.248	50.566	11.708	−0.038	−0.033	3.841	8.956	−3.261
Comp. 4	0.580 (0.527, 0.634)	0.572 (0.517, 0.628)	0.519 (0.435, 0.600)	0.001 (0.377)	0.001 (0.587)	−0.001 (0.471)	−0.138 (<0.001)	−0.185 (<0.001)	−0.012	0.035	51.478	50.660	−0.818	0.014	0.016	−16.627	16.118	−1.782
Comp. 5	0.593 (0.539, 0.648)	0.611 (0.557, 0.664)	0.617 (0.545, 0.692)	0.002 (0.357)	0.004 (0.202)	0.007 (0.216)	−0.117 (0.013)	−0.121 (0.011)	0.039	0.034	52.199	50.326	−1.872	−0.030	−0.033	17.087	−1.611	3.561
Comp. 6	0.547 (0.493, 0.603)	0.569 (0.516, 0.625)	0.550 (0.474, 0.627)	0.001 (0.433)	0.000 (0.957)	−0.001 (0.881)	−0.281 (<0.001)	−0.329 (<0.001)	−0.070	−0.056	52.275	51.096	−0.580	−0.015	−0.018	13.994	16.841	−6.299
Comp. 7	0.447 (0.393, 0.499)	0.463 (0.407, 0.516)	0.474 (0.401, 0.549)	−0.001 (0.495)	0.000 (0.772)	−0.001 (0.941)	−0.040 (0.024)	−0.051 (0.277)	−0.055	−0.068	51.106	50.374	−0.132	0.013	0.011	19.528	−4.025	1.243
Comp. 8	0.468 (0.414, 0.523)	0.463 (0.407, 0.518)	0.483 (0.406, 0.559)	0.000 (0.656)	0.000 (0.933)	0.003 (0.320)	0.253 (<0.001)	0.276 (<0.001)	0.021	0.025	50.836	49.446	−1.390	−0.023	−0.019	−18.375	−5.584	0.656
Comp. 9	0.397 (0.344, 0.447)	0.408 (0.352, 0.462)	0.447 (0.366, 0.528)	0.003 (0.057)	0.001 (0.167)	−0.001 (0.677)	0.212 (<0.001)	0.171 (<0.001)	−0.064	−0.034	55.094	49.945	−5.149	0.057	0.058	13.571	1.513	−0.840
Comp. 10	0.421 (0.371, 0.472)	0.432 (0.379, 0.485)	0.439 (0.363, 0.513)	0.008 (0.006)	0.003 (0.061)	0.005 (0.262)	−0.017 (0.712)	−0.005 (0.916)	−0.067	−0.037	53.366	50.179	−3.187	0.051	0.052	18.723	4.352	−2.225

Table A.2: Complete exploration for PCA components 1–10 for ten-second STFT.

Component	A1. AUROC (95% CI)				A2. ΔAUROC (p)				B1. Between-patient OpenBis corr.		B2. Within-patient corr.		C. Surgical change		D. Fent bolus corr.		E. Spectral regression coeff.		
	>0	>3	>6	>0	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Presurg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
Comp. 1	0.498 (0.445, 0.549)	0.498 (0.440, 0.553)	0.452 (0.373, 0.527)	0.000 (0.608)	0.000 (0.725)	0.000 (0.725)	−0.002 (0.292)	0.451 (<0.001)	0.455 (<0.001)	0.397	0.385	57.001	48.981	−8.020	−0.060	−0.044	−3.466	8.718	16.054
Comp. 2	0.586 (0.534, 0.638)	0.571 (0.516, 0.625)	0.509 (0.427, 0.586)	0.005 (0.161)	0.002 (0.446)	0.002 (0.446)	−0.002 (0.349)	−0.134 (0.004)	−0.096 (0.043)	−0.067	0.086	54.679	51.467	−3.212	0.036	0.045	−2.744	21.236	5.574
Comp. 3	0.574 (0.521, 0.627)	0.574 (0.518, 0.628)	0.516 (0.433, 0.596)	0.001 (0.261)	0.002 (0.286)	0.001 (0.261)	−0.001 (0.603)	−0.073 (0.124)	−0.064 (0.047)	0.013	0.020	50.998	49.881	−1.118	−0.003	−0.000	−7.094	−0.984	−0.468
Comp. 4	0.577 (0.523, 0.629)	0.549 (0.498, 0.604)	0.520 (0.442, 0.595)	0.004 (0.076)	0.001 (0.503)	0.001 (0.503)	−0.000 (0.926)	−0.266 (<0.001)	−0.258 (<0.001)	−0.030	−0.025	46.591	50.270	3.680	−0.011	−0.006	−6.070	2.715	−2.298
Comp. 5	0.552 (0.499, 0.604)	0.518 (0.465, 0.575)	0.494 (0.416, 0.567)	0.004 (0.136)	0.001 (0.778)	0.001 (0.778)	−0.000 (0.712)	−0.245 (<0.001)	−0.223 (<0.001)	−0.164	−0.117	36.548	50.927	14.379	−0.021	−0.003	−13.965	10.134	−9.710
Comp. 6	0.574 (0.522, 0.626)	0.550 (0.498, 0.606)	0.511 (0.429, 0.591)	0.003 (0.098)	0.000 (0.444)	0.000 (0.444)	−0.000 (0.759)	−0.161 (<0.001)	−0.153 (<0.001)	−0.006	0.001	47.078	50.080	3.002	−0.011	−0.003	−12.463	0.806	−3.306
Comp. 7	0.578 (0.525, 0.631)	0.601 (0.546, 0.655)	0.549 (0.467, 0.625)	−0.000 (0.906)	0.002 (0.521)	0.002 (0.521)	−0.002 (0.587)	−0.176 (<0.001)	−0.177 (<0.001)	0.028	0.109	62.046	50.805	−11.241	0.086	0.076	11.053	26.333	2.326
Comp. 8	0.578 (0.525, 0.631)	0.609 (0.553, 0.662)	0.561 (0.484, 0.638)	0.000 (0.726)	0.002 (0.488)	0.002 (0.488)	−0.002 (0.695)	−0.159 (<0.001)	−0.167 (<0.001)	0.030	0.081	61.481	50.687	−10.794	0.054	0.047	12.278	21.688	3.281
Comp. 9	0.435 (0.380, 0.488)	0.404 (0.353, 0.459)	0.405 (0.331, 0.476)	0.000 (0.513)	0.001 (0.611)	0.001 (0.611)	0.002 (0.529)	0.150 (0.001)	0.181 (<0.001)	0.001	−0.004	48.006	49.650	1.645	0.003	0.005	−12.205	−5.948	−2.719
Comp. 10	0.420 (0.368, 0.473)	0.426 (0.370, 0.481)	0.410 (0.335, 0.487)	0.004 (0.163)	−0.000 (0.291)	0.003 (0.228)	0.035 (0.464)	0.020 (0.672)	0.020 (0.672)	−0.110	−0.080	54.584	50.459	−4.125	0.099	0.092	−6.125	5.717	−7.724

Table A.3: Complete exploration for PCA components 1–10 for 30-second STFT.

Component	A1. AUROC (95% CI)			A2. ΔAUROC (p)			B1. Between-patient OpenBis corr.			B2. Within-patient corr.			C. Surgical change			D. Fent bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson ρ (p)	Spearman ρ (p)	Mean r	Mean ρ	Presurg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β			
Comp. 1	0.497 (0.444, 0.549)	0.499 (0.442, 0.554)	0.450 (0.372, 0.526)	−0.000 (0.649)	0.000 (0.724)	−0.002 (0.258)	0.454 (<0.001)	0.458 (<0.001)	0.432	0.427	58.077	48.943	−9.133	−0.086	−0.067	−3.057	8.441	16.443			
Comp. 2	0.586 (0.534, 0.638)	0.572 (0.517, 0.626)	0.510 (0.427, 0.587)	0.005 (0.168)	0.002 (0.458)	−0.001 (0.335)	−0.136 (0.004)	−0.097 (0.039)	−0.064	0.104	55.189	51.470	−3.719	0.024	0.039	−1.807	22.211	5.307			
Comp. 3	0.555 (0.501, 0.606)	0.522 (0.470, 0.579)	0.496 (0.418, 0.569)	0.004 (0.152)	0.000 (0.771)	0.000 (0.771)	−0.239 (<0.001)	−0.216 (<0.001)	−0.186	−0.125	36.307	50.828	14.521	0.010	0.026	−18.581	8.729	−12.063			
Comp. 4	0.578 (0.525, 0.631)	0.607 (0.550, 0.660)	0.558 (0.481, 0.632)	0.000 (0.765)	0.002 (0.483)	−0.002 (0.676)	−0.162 (<0.001)	−0.163 (<0.001)	0.067	0.137	64.445	50.767	−13.678	0.056	0.036	19.964	30.228	6.124			
Comp. 5	0.573 (0.520, 0.626)	0.607 (0.550, 0.660)	0.567 (0.491, 0.636)	0.000 (0.470)	0.001 (0.638)	−0.001 (0.739)	−0.145 (<0.001)	−0.164 (<0.001)	0.043	0.052	56.759	50.606	−6.153	0.004	−0.007	12.717	15.073	2.926			
Comp. 6	0.468 (0.412, 0.524)	0.453 (0.398, 0.510)	0.492 (0.414, 0.574)	0.002 (0.440)	−0.000 (0.806)	−0.001 (0.322)	0.088 (0.063)	0.116 (0.014)	0.001	−0.013	47.312	49.735	2.423	−0.018	−0.017	−2.703	−7.270	0.458			
Comp. 7	0.504 (0.451, 0.559)	0.503 (0.449, 0.560)	0.518 (0.435, 0.596)	0.002 (0.263)	−0.001 (0.367)	0.002 (0.357)	0.168 (<0.001)	0.200 (<0.001)	0.006	0.004	48.723	49.821	0.098	−0.006	−0.004	−2.275	−3.314	0.496			
Comp. 8	0.497 (0.441, 0.552)	0.491 (0.436, 0.550)	0.539 (0.450, 0.611)	0.002 (0.469)	−0.001 (0.600)	0.002 (0.526)	0.078 (0.099)	0.093 (0.050)	0.002	0.002	49.314	49.850	0.536	−0.006	−0.007	−0.359	−3.605	0.928			
Comp. 9	0.585 (0.532, 0.639)	0.621 (0.566, 0.674)	0.596 (0.525, 0.665)	0.000 (0.726)	0.002 (0.466)	−0.002 (0.676)	−0.148 (0.002)	−0.149 (0.002)	0.013	0.018	51.959	50.223	−1.736	−0.003	−0.003	6.778	6.515	1.623			
Comp. 10	0.526 (0.473, 0.579)	0.513 (0.460, 0.574)	0.545 (0.464, 0.623)	0.002 (0.261)	−0.002 (0.530)	0.002 (0.271)	0.073 (0.121)	0.097 (0.041)	0.017	0.007	48.493	49.726	1.233	−0.013	−0.013	−2.485	−6.308	1.627			

Table A.4: Complete exploration for PCA components 1–10 for 60-second STFT.

Component	A1. AUROC (95% CI)				A2. Δ AUROC (p)				B1. Between-patient Openbis corr.		B2. Within-patient corr.		C. Surgical change			D. Pent bolus corr.		E. Spectral regression coeff.		
	>0	>3	>6	>0	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Presurg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β	
Comp. 1	0.498 (0.444, 0.551)	0.500 (0.443, 0.555)	0.450 (0.370, 0.526)	−0.000 (0.666)	0.000 (0.729)	−0.001 (0.238)	−0.001 (0.238)	0.456 (<0.001)	0.456 (<0.001)	0.428	0.435	58.423	49.067	−9.357	−0.113	−0.104	−2.879	8.480	16.584	
Comp. 2	0.587 (0.536, 0.639)	0.573 (0.518, 0.627)	0.510 (0.428, 0.586)	0.005 (0.169)	0.002 (0.468)	−0.001 (0.325)	−0.139 (0.003)	−0.139 (0.003)	−0.212 (0.035)	−0.052	0.113	55.266	51.472	−3.794	0.017	0.028	−1.470	22.498	5.280	
Comp. 3	0.553 (0.500, 0.605)	0.521 (0.469, 0.577)	0.495 (0.416, 0.577)	0.004 (0.161)	0.000 (0.786)	0.000 (0.653)	−0.237 (<0.001)	−0.237 (<0.001)	−0.212 (<0.001)	−0.194	−0.129	36.065	50.630	14.565	0.042	0.050	−10.575	8.274	−12.684	
Comp. 4	0.578 (0.526, 0.632)	0.607 (0.550, 0.660)	0.558 (0.480, 0.634)	0.000 (0.736)	0.002 (0.500)	−0.002 (0.658)	−0.161 (<0.001)	0.002 (0.500)	−0.162 (<0.001)	0.087	0.144	64.552	50.912	−13.640	−0.000	−0.010	20.961	31.085	6.451	
Comp. 5	0.450 (0.397, 0.506)	0.438 (0.384, 0.493)	0.475 (0.398, 0.554)	0.000 (0.500)	−0.000 (0.882)	0.000 (0.411)	0.168 (<0.001)	0.168 (<0.001)	0.174 (<0.001)	−0.046	−0.041	47.085	49.452	2.367	0.037	0.039	−3.963	−12.277	1.247	
Comp. 6	0.423 (0.371, 0.476)	0.432 (0.375, 0.487)	0.419 (0.344, 0.494)	0.004 (0.150)	−0.000 (0.311)	0.003 (0.284)	0.017 (0.715)	0.001 (0.379)	−0.001 (0.979)	−0.127	−0.095	56.035	50.561	−5.473	0.115	0.109	−4.184	7.330	−7.856	
Comp. 7	0.592 (0.539, 0.645)	0.590 (0.536, 0.642)	0.581 (0.510, 0.653)	0.001 (0.464)	0.001 (0.475)	0.001 (0.737)	−0.105 (0.027)	−0.105 (0.027)	−0.106 (0.035)	0.027	0.026	49.917	50.045	0.128	0.005	0.004	2.249	5.064	0.354	
Comp. 8	0.473 (0.418, 0.527)	0.467 (0.412, 0.521)	0.474 (0.395, 0.549)	0.002 (0.322)	−0.001 (0.737)	−0.001 (0.815)	0.178 (<0.001)	0.178 (<0.001)	0.195 (<0.001)	0.008	0.006	49.088	49.463	0.375	0.021	0.017	−4.056	−9.147	2.384	
Comp. 9	0.421 (0.370, 0.474)	0.415 (0.362, 0.469)	0.406 (0.331, 0.483)	0.001 (0.576)	0.004 (0.293)	0.009 (0.232)	0.148 (0.002)	0.148 (0.002)	0.164 (<0.001)	−0.001	−0.001	50.196	49.691	−0.505	0.009	0.010	−3.420	−3.300	−0.341	
Comp. 10	0.404 (0.350, 0.456)	0.400 (0.347, 0.453)	0.402 (0.331, 0.477)	0.001 (0.569)	0.004 (0.292)	0.003 (0.480)	0.288 (<0.001)	0.288 (<0.001)	0.288 (<0.001)	0.015	0.018	50.868	49.526	−1.342	0.012	0.014	−2.118	−3.730	0.678	

Table A.5: Complete exploration for PCA components 1–10 for 120-second STFT.

Component	A1. AUROC (95% CI)				A2. Δ AUROC (p)				B1. Between-patient Openbis corr.		B2. Within-patient corr.		C. Surgical change		D. Pent bolus corr.		E. Spectral regression coeff.		
	>0	>3	>6	>0	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Presurg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
Comp. 1	0.502 (0.448, 0.554)	0.505 (0.448, 0.559)	0.451 (0.372, 0.526)	−0.000 (0.663)	0.001 (0.721)	−0.001 (0.211)	−0.001 (0.211)	0.431 (<0.001)	0.439 (<0.001)	0.396	0.414	58.618	49.392	−9.226	−0.135	−0.132	−2.839	8.793	16.517
Comp. 2	0.589 (0.537, 0.641)	0.574 (0.519, 0.629)	0.509 (0.428, 0.585)	0.005 (0.166)	0.002 (0.478)	−0.001 (0.396)	−0.108 (0.022)	−0.108 (0.022)	−0.108 (0.022)	−0.036	0.112	55.157	51.508	−3.650	0.015	0.015	−1.191	22.769	5.178
Comp. 3	0.550 (0.496, 0.601)	0.516 (0.464, 0.573)	0.490 (0.412, 0.563)	0.003 (0.169)	0.000 (0.824)	0.000 (0.642)	−0.234 (<0.001)	−0.234 (<0.001)	−0.212 (<0.001)	−0.191	−0.130	35.578	50.139	14.561	0.059	0.060	−20.338	7.283	−13.017
Comp. 4	0.578 (0.525, 0.632)	0.605 (0.549, 0.658)	0.554 (0.475, 0.627)	0.000 (0.737)	0.002 (0.551)	−0.002 (0.590)	−0.167 (<0.001)	−0.167 (<0.001)	−0.167 (<0.001)	0.095	0.141	63.936	51.213	−12.723	0.001	−0.004	19.637	31.494	5.541
Comp. 5	0.451 (0.397, 0.508)	0.441 (0.387, 0.496)	0.480 (0.403, 0.562)	0.001 (0.535)	−0.001 (0.798)	0.001 (0.360)	0.181 (<0.001)	0.181 (<0.001)	0.186 (<0.001)	−0.056	−0.081	43.678	48.652	4.974	0.091	0.087	−6.597	−24.392	2.766
Comp. 6	0.423 (0.371, 0.475)	0.432 (0.376, 0.488)	0.423 (0.349, 0.500)	0.003 (0.158)	−0.000 (0.328)	0.002 (0.324)	0.018 (0.703)	0.001 (0.381)	0.001 (0.381)	−0.112	−0.088	55.802	50.622	−5.180	0.096	0.089	−3.846	7.478	−7.999
Comp. 7	0.433 (0.381, 0.487)	0.441 (0.387, 0.494)	0.462 (0.386, 0.539)	0.001 (0.440)	0.000 (0.709)	−0.001 (0.887)	0.036 (0.446)	0.006 (0.896)	0.006 (0.896)	−0.053	−0.042	49.991	50.265	0.274	0.028	0.036	0.885	−4.284	0.301
Comp. 8	0.447 (0.395, 0.503)	0.452 (0.398, 0.504)	0.467 (0.393, 0.543)	0.000 (0.736)	−0.001 (0.880)	−0.001 (0.744)	0.166 (<0.001)	0.146 (0.002)	0.146 (0.002)	−0.014	−0.015	49.890	49.830	−0.060	0.005	0.002	0.245	−8.960	3.325
Comp. 9	0.399 (0.350, 0.452)	0.400 (0.348, 0.453)	0.383 (0.307, 0.462)	0.009 (0.069)	0.008 (0.066)	0.004 (0.122)	0.204 (<0.001)	0.187 (<0.001)	0.187 (<0.001)	−0.003	0.005	50.799	49.576	−1.223	−0.007	−0.009	−2.680	−0.537	−1.081
Comp. 10	0.401 (0.349, 0.455)	0.388 (0.333, 0.444)	0.392 (0.317, 0.470)	0.002 (0.500)	0.005 (0.239)	−0.000 (0.383)	0.383 (<0.001)	0.368 (<0.001)	0.368 (<0.001)	0.005	0.026	53.169	48.937	−4.232	0.034	0.038	−11.869	−9.889	−0.290

Table A.6: Complete exploration for EFA factors 1–10 for one-second STFT.

Factor	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.			B2. Within-patient corr.			C. Surgical change			D. Feat bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β			
Factor 1	0.310 (0.291, 0.391)	0.350 (0.298, 0.403)	0.408 (0.334, 0.483)	0.010 (0.024)	0.003 (0.098)	−0.002 (0.876)	0.863 (<0.001)	0.890 (<0.001)	0.443	0.355	−0.103	−5.596	−0.103	−0.104	−0.104	−3.656	−6.745	9.434			
Factor 2	0.373 (0.322, 0.423)	0.396 (0.341, 0.449)	0.425 (0.353, 0.501)	0.006 (0.143)	0.001 (0.935)	−0.002 (0.832)	0.783 (<0.001)	0.792 (<0.001)	0.398	0.324	−0.075	−7.981	−0.075	−0.075	−0.075	5.183	−3.830	12.893			
Factor 3	0.567 (0.514, 0.619)	0.559 (0.499, 0.614)	0.495 (0.414, 0.570)	0.006 (0.135)	0.003 (0.284)	−0.003 (0.426)	0.077 (0.103)	0.102 (0.031)	0.105	0.168	0.027	50.684	−6.474	0.027	0.031	1.591	11.407	14.708			
Factor 4	0.444 (0.391, 0.496)	0.462 (0.405, 0.517)	0.454 (0.379, 0.528)	0.001 (0.452)	0.001 (0.843)	−0.000 (0.516)	0.496 (<0.001)	0.525 (<0.001)	0.217	0.198	−0.092	57.824	49.097	−8.727	−0.092	1.935	−7.196	14.364			
Factor 5	0.572 (0.518, 0.626)	0.570 (0.515, 0.625)	0.513 (0.426, 0.592)	0.000 (0.678)	0.001 (0.775)	−0.001 (0.331)	−0.192 (<0.001)	−0.174 (<0.001)	−0.096	−0.005	0.046	54.231	51.760	−2.470	0.046	0.049	2.350	24.549	0.058		
Factor 6	0.481 (0.427, 0.534)	0.434 (0.382, 0.489)	0.457 (0.387, 0.526)	0.002 (0.182)	0.000 (0.572)	−0.001 (0.737)	−0.003 (0.944)	0.020 (0.678)	−0.097	−0.075	−0.029	41.977	50.219	8.243	−0.029	−0.018	−1.970	2.338	0.159		
Factor 7	0.502 (0.448, 0.554)	0.469 (0.414, 0.522)	0.441 (0.365, 0.516)	0.003 (0.402)	0.001 (0.754)	0.008 (0.265)	−0.179 (<0.001)	−0.148 (0.002)	−0.118	−0.086	0.014	41.680	50.974	9.293	0.014	0.016	−1.220	11.274	−1.531		
Factor 8	0.578 (0.527, 0.630)	0.562 (0.505, 0.619)	0.578 (0.497, 0.659)	0.003 (0.483)	0.001 (0.634)	−0.001 (0.356)	−0.347 (<0.001)	−0.358 (<0.001)	−0.050	−0.060	−0.033	45.389	51.029	5.640	−0.033	19.955	2.514	−0.755			
Factor 9	0.395 (0.342, 0.446)	0.414 (0.359, 0.468)	0.453 (0.377, 0.533)	0.004 (0.028)	0.000 (0.126)	0.000 (0.769)	0.153 (0.003)	0.138 (0.003)	−0.040	−0.039	0.040	53.269	49.903	−3.367	0.040	0.040	16.020	−3.811	−1.492		
Factor 10	0.441 (0.389, 0.498)	0.437 (0.383, 0.492)	0.446 (0.376, 0.519)	−0.000 (0.818)	0.000 (0.802)	−0.000 (0.926)	0.057 (0.230)	0.001 (0.980)	−0.009	−0.059	−0.041	47.181	50.554	3.373	−0.041	−0.039	−2.652	−4.368	−2.291		

Table A.7: Complete exploration for EFA factors 1–10 for ten-second STFT.

Factor	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.			B2. Within-patient corr.			C. Surgical change			D. Feat bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β			
Factor 1	0.608 (0.556, 0.660)	0.602 (0.547, 0.656)	0.545 (0.466, 0.619)	0.005 (0.112)	0.004 (0.208)	−0.001 (0.637)	−0.117 (0.013)	−0.092 (0.051)	0.082	0.168	−0.010	55.887	50.847	−5.040	−0.010	−0.002	3.796	20.409	9.878		
Factor 2	0.435 (0.384, 0.486)	0.436 (0.379, 0.489)	0.418 (0.341, 0.492)	−0.000 (0.876)	0.001 (0.896)	−0.003 (0.317)	0.697 (<0.001)	0.691 (<0.001)	0.247	0.204	−0.044	51.063	48.312	−2.751	−0.050	−0.044	−12.187	−5.938	6.703		
Factor 3	0.377 (0.327, 0.428)	0.388 (0.333, 0.442)	0.400 (0.331, 0.475)	0.003 (0.271)	0.002 (0.314)	−0.003 (0.282)	0.802 (<0.001)	0.783 (<0.001)	0.241	0.196	−0.039	52.795	48.553	−4.242	−0.039	−0.037	8.647	0.513	11.343		
Factor 4	0.416 (0.365, 0.467)	0.429 (0.373, 0.482)	0.413 (0.338, 0.487)	0.001 (0.717)	0.000 (0.572)	−0.003 (0.258)	0.732 (<0.001)	0.711 (<0.001)	0.207	0.158	−0.046	50.266	48.569	−1.696	−0.046	−0.046	−4.148	−0.994	5.784		
Factor 5	0.385 (0.334, 0.434)	0.391 (0.335, 0.445)	0.401 (0.328, 0.477)	0.001 (0.601)	0.001 (0.417)	−0.003 (0.274)	0.814 (<0.001)	0.806 (<0.001)	0.190	0.143	−0.035	50.702	48.768	−1.934	−0.035	−0.035	−1.911	−2.939	6.999		
Factor 6	0.394 (0.341, 0.443)	0.399 (0.343, 0.452)	0.407 (0.335, 0.482)	0.001 (0.668)	0.001 (0.400)	−0.003 (0.291)	0.812 (<0.001)	0.804 (<0.001)	0.178	0.134	−0.036	50.799	48.888	−1.911	−0.036	−0.035	−1.170	−2.789	7.034		
Factor 7	0.388 (0.339, 0.439)	0.386 (0.331, 0.439)	0.403 (0.334, 0.474)	0.000 (0.945)	0.001 (0.602)	−0.003 (0.488)	0.753 (<0.001)	0.741 (<0.001)	0.186	0.137	−0.034	50.131	48.751	−1.380	−0.034	−0.032	−2.847	−6.384	8.112		
Factor 8	0.440 (0.385, 0.495)	0.416 (0.359, 0.473)	0.450 (0.375, 0.528)	0.003 (0.365)	0.000 (0.993)	−0.001 (0.645)	0.262 (<0.001)	0.289 (<0.001)	0.060	0.003	−0.063	40.682	48.992	8.309	−0.072	−0.054	−12.693	−25.947	6.263		
Factor 9	0.522 (0.468, 0.575)	0.506 (0.451, 0.561)	0.461 (0.381, 0.540)	0.000 (0.649)	−0.000 (0.918)	0.004 (0.233)	−0.146 (0.002)	−0.113 (0.017)	−0.159	−0.061	0.067	47.494	51.520	4.026	0.067	0.067	−9.994	18.406	−5.014		
Factor 10	0.551 (0.498, 0.605)	0.514 (0.463, 0.569)	0.516 (0.441, 0.590)	0.003 (0.090)	0.001 (0.745)	−0.001 (0.791)	−0.252 (<0.001)	−0.236 (<0.001)	−0.136	−0.131	−0.053	32.677	50.416	17.740	−0.053	−0.037	−11.916	0.274	−9.907		

Table A.8: Complete exploration for EFA factors 1–10 for 30-second STFT.

Factor	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.			B2. Within-patient corr.			C. Surgical change			D. Feat bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β			
Factor 1	0.595 (0.545, 0.648)	0.575 (0.520, 0.629)	0.512 (0.433, 0.587)	0.005 (0.077)	0.002 (0.381)	−0.002 (0.383)	−0.219 (<0.001)	−0.166 (<0.001)	−0.201	0.003	0.067	53.891	52.090	−1.800	0.067	0.078	−3.567	20.132	3.395		
Factor 2	0.388 (0.336, 0.437)	0.397 (0.343, 0.449)	0.410 (0.337, 0.485)	0.003 (0.307)	0.001 (0.497)	−0.004 (0.487)	0.834 (<0.001)	0.852 (<0.001)	0.349	0.302	−0.045	52.236	47.065	−5.171	−0.075	−0.068	−3.689	−5.874	12.128		
Factor 3	0.394 (0.343, 0.445)	0.403 (0.348, 0.456)	0.413 (0.340, 0.488)	0.003 (0.363)	0.001 (0.556)	−0.004 (0.462)	0.818 (<0.001)	0.834 (<0.001)	0.338	0.299	−0.041	52.750	47.225	−5.525	−0.061	−0.055	−4.512	−5.806	12.034		
Factor 4	0.393 (0.341, 0.443)	0.402 (0.347, 0.455)	0.413 (0.339, 0.488)	0.003 (0.335)	0.002 (0.521)	−0.004 (0.465)	0.824 (<0.001)	0.839 (<0.001)	0.353	0.304	−0.046	52.715	47.230	−5.485	−0.077	−0.073	−2.892	−4.579	12.466		
Factor 5	0.404 (0.352, 0.454)	0.412 (0.356, 0.465)	0.419 (0.346, 0.493)	0.002 (0.452)	0.001 (0.621)	−0.003 (0.468)	0.793 (<0.001)	0.807 (<0.001)	0.374	0.327	−0.047	53.320	47.614	−5.706	−0.067	−0.063	0.761	−1.394	13.623		
Factor 6	0.393 (0.342, 0.444)	0.402 (0.348, 0.455)	0.414 (0.341, 0.489)	0.003 (0.347)	0.002 (0.533)	−0.003 (0.487)	0.822 (<0.001)	0.838 (<0.001)	0.359	0.307	−0.047	52.727	47.246	−5.480	−0.080	−0.077	−2.908	−4.684	12.572		
Factor 7	0.400 (0.349, 0.450)	0.409 (0.354, 0.461)	0.416 (0.343, 0.490)	0.002 (0.420)	0.001 (0.603)	−0.003 (0.450)	0.807 (<0.001)	0.821 (<0.001)	0.344	0.301	−0.043	53.092	47.373	−5.718	−0.073	−0.068	−3.067	−4.340	12.935		
Factor 8	0.394 (0.342, 0.444)	0.403 (0.348, 0.456)	0.415 (0.341, 0.490)	0.002 (0.341)	0.002 (0.527)	−0.004 (0.478)	0.821 (<0.001)	0.836 (<0.001)	0.365	0.310	−0.055	52.837	47.295	−5.542	−0.076	−0.071	−2.780	−4.447	12.525		
Factor 9	0.395 (0.343, 0.445)	0.403 (0.348, 0.456)	0.415 (0.341, 0.490)	0.003 (0.334)	0.002 (0.518)	−0.004 (0.474)	0.818 (<0.001)	0.833 (<0.001)	0.371	0.315	−0.056	52.941	47.310	−5.631	−0.073	−0.069	−2.368	−3.954	12.545		
Factor 10	0.444 (0.392, 0.498)	0.412 (0.361, 0.465)	0.455 (0.383, 0.531)	0.000 (0.731)	0.004 (0.188)	−0.001 (0.974)	0.032 (0.496)	−0.017 (0.714)	0.124	−0.077	−0.105	29.792	48.099	18.306	−0.105	−0.090	−14.701	−12.488	−14.690		

Table A.9: Complete exploration for EFA factors 1–10 for 60-second STFT.

Factor	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient		B2. Within-patient corr.		C. Surgical change		D. Feat bolus corr.		E. Spectral regression coeff.			
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
Factor 1	0.390 (0.398, 0.439)	0.400 (0.345, 0.453)	0.413 (0.340, 0.487)	0.003 (0.333)	0.001 (0.550)	−0.004 (0.491)	0.830 (<0.001)	0.848 (<0.001)	0.314	0.284	52.506	47.099	−5.406	−0.027	−0.022	−3.023	−6.001	12.218
Factor 2	0.396 (0.345, 0.446)	0.404 (0.349, 0.456)	0.416 (0.342, 0.491)	0.003 (0.372)	0.001 (0.557)	−0.004 (0.482)	0.815 (<0.001)	0.830 (<0.001)	0.343	0.300	53.102	47.442	−5.661	−0.077	−0.072	−3.375	−5.100	12.577
Factor 3	0.393 (0.341, 0.443)	0.402 (0.348, 0.455)	0.413 (0.339, 0.487)	0.002 (0.342)	0.002 (0.554)	−0.004 (0.477)	0.826 (<0.001)	0.842 (<0.001)	0.312	0.283	52.780	47.230	−5.550	−0.030	−0.028	−3.201	−5.278	12.342
Factor 4	0.393 (0.341, 0.444)	0.402 (0.347, 0.454)	0.415 (0.341, 0.490)	0.003 (0.347)	0.002 (0.543)	−0.003 (0.502)	0.825 (<0.001)	0.841 (<0.001)	0.335	0.295	52.739	47.303	−5.436	−0.070	−0.068	−3.418	−5.022	12.377
Factor 5	0.398 (0.347, 0.448)	0.406 (0.351, 0.459)	0.416 (0.342, 0.489)	0.002 (0.389)	0.001 (0.562)	−0.003 (0.471)	0.804 (<0.001)	0.819 (<0.001)	0.370	0.325	53.264	47.628	−5.636	−0.064	−0.064	−1.152	−3.190	12.899
Factor 6	0.397 (0.346, 0.446)	0.405 (0.350, 0.457)	0.416 (0.342, 0.490)	0.002 (0.385)	0.001 (0.563)	−0.004 (0.481)	0.812 (<0.001)	0.827 (<0.001)	0.354	0.305	53.219	47.511	−5.708	−0.071	−0.069	−3.229	−4.848	12.578
Factor 7	0.395 (0.344, 0.446)	0.404 (0.349, 0.457)	0.415 (0.341, 0.490)	0.003 (0.364)	0.002 (0.560)	−0.004 (0.478)	0.821 (<0.001)	0.836 (<0.001)	0.324	0.292	52.958	47.351	−5.607	−0.064	−0.060	−3.143	−4.937	12.511
Factor 8	0.397 (0.346, 0.448)	0.406 (0.351, 0.459)	0.415 (0.342, 0.489)	0.003 (0.394)	0.001 (0.593)	−0.004 (0.465)	0.813 (<0.001)	0.829 (<0.001)	0.315	0.287	53.151	47.390	−5.761	−0.045	−0.039	−3.132	−4.770	12.725
Factor 9	0.444 (0.391, 0.499)	0.412 (0.362, 0.465)	0.455 (0.385, 0.530)	−0.000 (0.737)	0.004 (0.192)	−0.001 (0.953)	0.028 (0.553)	−0.020 (0.669)	0.069	−0.098	29.665	48.013	18.348	−0.084	−0.081	−15.507	−13.222	−15.368

Table A.10: Complete exploration for EFA factors 1–10 for 120-second STFT.

Factor	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.		B2. Within-patient corr.		C. Surgical change		D. Feat bolus corr.		E. Spectral regression coeff.			
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
Factor 1	0.594 (0.545, 0.646)	0.575 (0.520, 0.628)	0.510 (0.430, 0.584)	0.006 (0.078)	0.002 (0.398)	−0.002 (0.549)	−0.210 (<0.001)	−0.157 (<0.001)	−0.136	0.029	54.193	51.980	−2.214	0.063	0.064	−3.324	19.564	4.207
Factor 2	0.393 (0.342, 0.443)	0.400 (0.344, 0.452)	0.412 (0.339, 0.487)	0.002 (0.322)	0.002 (0.496)	−0.003 (0.460)	0.817 (<0.001)	0.833 (<0.001)	0.325	0.291	53.310	47.650	−5.659	−0.085	−0.084	−3.787	−5.832	12.348
Factor 3	0.395 (0.344, 0.445)	0.402 (0.347, 0.455)	0.413 (0.340, 0.488)	0.003 (0.318)	0.002 (0.483)	−0.004 (0.441)	0.808 (<0.001)	0.823 (<0.001)	0.355	0.317	53.427	47.830	−5.598	−0.065	−0.067	−2.486	−4.596	12.501
Factor 4	0.396 (0.345, 0.446)	0.407 (0.352, 0.459)	0.413 (0.341, 0.489)	0.002 (0.442)	0.001 (0.706)	−0.003 (0.485)	0.814 (<0.001)	0.833 (<0.001)	0.271	0.252	53.079	47.276	−5.803	0.007	0.008	−3.637	−5.797	12.763
Factor 5	0.394 (0.343, 0.445)	0.402 (0.347, 0.454)	0.414 (0.341, 0.490)	0.002 (0.346)	0.002 (0.530)	−0.004 (0.479)	0.817 (<0.001)	0.833 (<0.001)	0.306	0.284	53.307	47.585	−5.722	−0.076	−0.076	−3.729	−5.729	12.552
Factor 6	0.390 (0.339, 0.440)	0.399 (0.345, 0.452)	0.414 (0.340, 0.488)	0.003 (0.299)	0.001 (0.504)	−0.004 (0.480)	0.826 (<0.001)	0.842 (<0.001)	0.292	0.275	53.113	47.431	−5.682	−0.035	−0.032	−3.552	−5.843	12.355
Factor 7	0.391 (0.340, 0.440)	0.401 (0.346, 0.453)	0.413 (0.340, 0.488)	0.002 (0.346)	0.001 (0.580)	−0.004 (0.483)	0.829 (<0.001)	0.847 (<0.001)	0.280	0.261	52.825	47.261	−5.564	0.004	0.006	−3.626	−6.013	12.413
Factor 8	0.394 (0.342, 0.444)	0.405 (0.349, 0.458)	0.413 (0.339, 0.487)	0.003 (0.389)	0.001 (0.635)	−0.004 (0.471)	0.823 (<0.001)	0.840 (<0.001)	0.275	0.256	52.923	47.273	−5.650	−0.004	−0.003	−3.412	−5.556	12.538
Factor 9	0.393 (0.341, 0.443)	0.402 (0.347, 0.455)	0.415 (0.340, 0.490)	0.002 (0.349)	0.002 (0.565)	−0.004 (0.480)	0.823 (<0.001)	0.840 (<0.001)	0.281	0.265	53.043	47.408	−5.635	−0.012	−0.007	−3.548	−5.704	12.450
Factor 10	0.444 (0.392, 0.499)	0.413 (0.363, 0.466)	0.458 (0.386, 0.533)	−0.000 (0.752)	0.004 (0.196)	−0.001 (0.927)	0.016 (0.739)	−0.034 (0.471)	−0.010	−0.122	29.616	47.737	18.121	−0.040	−0.038	−15.806	−12.719	−16.296

Table A.11: Complete exploration for Fractional Occupancies and HMM statistics.

Component	A1. AUROC (95% CI)			A2. ΔAUROC (p)			B1. Between-patient Openbis corr.			B2. Within-patient corr.			C. Surgical change			D. Fent bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β			
FO State 1	0.541 (0.489, 0.593)	0.540 (0.486, 0.596)	0.573 (0.501, 0.650)	-0.000 (0.907)	0.000 (0.769)	-0.001 (0.164)	-0.586 (<0.001)	-0.576 (<0.001)	-0.298	-0.252	48.017	51.546	3.529	0.058	0.053	6.350	-6.555	-11.141			
FO State 2	0.606 (0.554, 0.658)	0.580 (0.523, 0.634)	0.555 (0.482, 0.626)	0.005 (0.280)	0.001 (0.781)	-0.001 (0.810)	-0.815 (<0.001)	-0.816 (<0.001)	-0.372	-0.289	47.443	53.821	6.378	0.087	0.091	2.562	6.031	-10.798			
FO State 3	0.398 (0.348, 0.447)	0.401 (0.346, 0.453)	0.414 (0.340, 0.489)	0.002 (0.363)	0.002 (0.429)	-0.004 (0.474)	0.825 (<0.001)	0.840 (<0.001)	0.418	0.361	51.523	46.402	-5.121	-0.071	-0.070	-3.869	-1.270	13.264			
FO State 4	0.508 (0.456, 0.560)	0.504 (0.447, 0.561)	0.462 (0.379, 0.537)	0.000 (0.656)	0.000 (0.732)	-0.004 (0.368)	0.218 (<0.001)	0.324 (<0.001)	-0.066	0.071	54.840	50.198	-4.641	0.044	0.038	-6.510	11.390	6.476			
FO State 5	0.344 (0.297, 0.393)	0.357 (0.308, 0.410)	0.412 (0.337, 0.493)	0.008 (0.022)	0.003 (0.198)	0.001 (0.891)	0.698 (<0.001)	0.781 (<0.001)	0.132	0.161	49.825	46.406	-3.419	-0.029	-0.036	-1.921	-6.994	0.043			
FO State 6	0.583 (0.531, 0.636)	0.548 (0.492, 0.603)	0.510 (0.438, 0.582)	0.004 (0.329)	0.001 (0.942)	-0.000 (0.698)	-0.751 (<0.001)	-0.709 (<0.001)	-0.354	-0.268	46.940	53.159	6.219	0.089	0.092	-0.150	7.071	-10.010			
Switching Rate	0.476 (0.425, 0.531)	0.452 (0.399, 0.506)	0.489 (0.415, 0.564)	0.009 (0.928)	0.001 (0.906)	0.007 (0.396)	-0.129 (0.006)	0.122 (0.010)	-0.159	-0.082	48.516	50.830	2.413	0.049	0.049	-1.518	0.966	-3.515			
State Entropy	0.453 (0.401, 0.506)	0.437 (0.381, 0.490)	0.462 (0.385, 0.537)	0.001 (0.858)	0.000 (0.845)	0.003 (0.570)	-0.123 (0.009)	0.169 (<0.001)	-0.196	-0.110	49.140	50.512	1.373	0.067	0.065	-2.569	0.570	-4.940			

Table A.12: Complete exploration for transition matrix statistics.

Component	A1. AUROC (95% CI)				A2. ΔAUROC (p)				B1. Between-patient Openbis corr.			B2. Within-patient corr.			C. Surgical change			D. Fent bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β				
S1→S1	0.541 (0.490, 0.592)	0.540 (0.486, 0.596)	0.572 (0.500, 0.649)	0.000 (0.945)	0.000 (0.907)	-0.000 (0.175)	0.000 (0.907)	0.000 (0.919)	0.000 (0.994)	-0.589 (<0.001)	-0.577 (<0.001)	48.348	51.535	3.186	0.059	0.056	6.167	-5.610	-10.637				
S1→S2	0.600 (0.500, 0.653)	0.580 (0.527, 0.635)	0.546 (0.463, 0.624)	0.004 (0.373)	0.001 (0.419)	-0.000 (0.994)	0.000 (0.994)	0.001 (0.919)	0.000 (0.994)	-0.589 (<0.001)	-0.583 (<0.001)	49.965	51.554	1.589	0.049	0.052	-1.984	7.022	-2.838				
S1→S3	0.413 (0.361, 0.463)	0.418 (0.361, 0.472)	0.415 (0.341, 0.488)	0.002 (0.555)	0.001 (0.618)	0.000 (0.994)	0.000 (0.897)	0.001 (0.618)	0.002 (0.361)	0.778 (<0.001)	0.805 (<0.001)	51.047	47.130	-3.917	-0.055	-0.052	-5.144	1.942	9.178				
S1→S4	0.464 (0.414, 0.516)	0.461 (0.404, 0.515)	0.439 (0.352, 0.502)	-0.000 (0.971)	0.000 (0.897)	-0.001 (0.268)	0.000 (0.971)	0.000 (0.897)	0.000 (0.971)	0.540 (<0.001)	0.580 (<0.001)	51.711	48.597	-3.114	-0.028	-0.021	-6.980	6.198	7.498				
S1→S5	0.383 (0.334, 0.432)	0.393 (0.339, 0.446)	0.409 (0.333, 0.484)	0.003 (0.228)	0.002 (0.429)	-0.003 (0.427)	0.003 (0.228)	0.002 (0.429)	0.003 (0.427)	0.820 (<0.001)	0.857 (<0.001)	50.945	46.674	-4.271	-0.053	-0.049	-4.659	0.920	6.765				
S1→S6	0.552 (0.499, 0.605)	0.538 (0.483, 0.593)	0.481 (0.403, 0.556)	0.002 (0.479)	0.001 (0.546)	-0.005 (0.343)	0.002 (0.479)	0.001 (0.546)	0.002 (0.343)	0.007 (0.156)	0.067 (0.156)	50.800	51.213	0.413	0.049	0.052	-2.967	7.452	-1.001				
S2→S1	0.502 (0.450, 0.555)	0.511 (0.455, 0.568)	0.563 (0.488, 0.642)	0.000 (0.667)	0.000 (0.558)	0.002 (0.107)	0.000 (0.667)	0.000 (0.558)	0.002 (0.107)	0.412 (<0.001)	0.373 (<0.001)	47.789	50.298	2.509	0.033	0.032	4.353	-7.378	-9.227				
S2→S2	0.597 (0.543, 0.647)	0.573 (0.514, 0.629)	0.543 (0.471, 0.613)	0.004 (0.396)	0.001 (0.833)	-0.000 (0.901)	0.000 (0.901)	0.001 (0.833)	0.000 (0.901)	0.808 (<0.001)	0.799 (<0.001)	48.667	53.241	4.574	0.082	0.087	2.934	5.722	-9.565				
S2→S3	0.407 (0.355, 0.456)	0.409 (0.355, 0.462)	0.415 (0.339, 0.490)	0.002 (0.434)	0.002 (0.434)	-0.003 (0.336)	0.002 (0.434)	0.002 (0.434)	0.003 (0.336)	0.812 (<0.001)	0.823 (<0.001)	51.245	46.819	-4.427	-0.065	-0.068	-4.443	0.013	10.869				
S2→S4	0.477 (0.426, 0.520)	0.471 (0.414, 0.526)	0.434 (0.355, 0.508)	0.000 (0.942)	0.000 (0.775)	-0.001 (0.216)	0.000 (0.942)	0.000 (0.775)	0.001 (0.216)	0.515 (<0.001)	0.539 (<0.001)	52.854	48.881	-3.973	-0.018	-0.019	-6.097	6.583	8.293				
S2→S5	0.355 (0.307, 0.406)	0.367 (0.315, 0.421)	0.410 (0.338, 0.488)	0.007 (0.068)	0.003 (0.244)	-0.002 (0.732)	0.003 (0.244)	0.003 (0.244)	0.002 (0.732)	0.818 (<0.001)	0.846 (<0.001)	50.494	45.333	-5.161	-0.060	-0.068	-2.301	-7.152	4.390				
S2→S6	0.590 (0.532, 0.639)	0.565 (0.510, 0.620)	0.545 (0.471, 0.621)	0.004 (0.341)	0.001 (0.477)	0.001 (0.514)	0.000 (0.901)	0.001 (0.477)	0.001 (0.514)	0.769 (<0.001)	0.699 (<0.001)	47.546	51.599	4.053	0.032	0.040	-1.151	2.069	-5.559				
S3→S1	0.560 (0.510, 0.614)	0.537 (0.503, 0.612)	0.577 (0.504, 0.654)	0.001 (0.650)	0.000 (0.591)	-0.002 (0.235)	0.000 (0.901)	0.000 (0.591)	0.002 (0.235)	0.698 (<0.001)	0.688 (<0.001)	46.961	51.691	4.730	0.058	0.052	5.240	-3.705	-12.227				
S3→S2	0.612 (0.562, 0.663)	0.602 (0.547, 0.656)	0.580 (0.507, 0.652)	0.005 (0.186)	0.002 (0.346)	-0.003 (0.469)	0.005 (0.186)	0.002 (0.346)	0.003 (0.469)	0.850 (<0.001)	0.859 (<0.001)	47.358	53.419	6.061	0.072	0.074	3.259	3.422	-11.677				
S3→S3	0.401 (0.350, 0.449)	0.401 (0.346, 0.452)	0.413 (0.338, 0.489)	0.002 (0.370)	0.002 (0.386)	0.003 (0.430)	0.002 (0.370)	0.002 (0.386)	0.003 (0.430)	0.819 (<0.001)	0.828 (<0.001)	51.366	46.637	-4.729	-0.070	-0.067	-3.571	0.967	12.941				
S3→S4	0.606 (0.555, 0.657)	0.577 (0.523, 0.632)	0.513 (0.433, 0.590)	0.007 (0.142)	0.001 (0.651)	-0.000 (0.327)	0.000 (0.901)	0.001 (0.651)	0.000 (0.327)	0.562 (<0.001)	0.509 (<0.001)	50.798	53.207	2.409	0.086	0.094	-3.538	12.171	-4.821				
S3→S5	0.431 (0.379, 0.484)	0.429 (0.376, 0.484)	0.478 (0.402, 0.556)	0.000 (0.522)	0.000 (0.824)	0.002 (0.283)	0.000 (0.824)	0.000 (0.824)	0.002 (0.283)	0.008 (0.861)	0.122 (0.010)	48.045	48.913	0.868	0.037	0.031	2.861	-6.084	-8.066				
S3→S6	0.611 (0.561, 0.662)	0.602 (0.547, 0.657)	0.575 (0.497, 0.650)	0.004 (0.285)	0.002 (0.472)	0.002 (0.737)	0.000 (0.901)	0.002 (0.472)	0.002 (0.737)	0.856 (<0.001)	0.856 (<0.001)	47.158	52.976	5.819	0.072	0.075	3.169	4.109	-11.571				
S4→S1	0.511 (0.459, 0.564)	0.516 (0.460, 0.572)	0.561 (0.487, 0.640)	0.000 (0.786)	0.000 (0.839)	0.001 (0.154)	0.000 (0.839)	0.001 (0.154)	0.000 (0.839)	0.477 (<0.001)	0.455 (<0.001)	46.723	50.555	3.832	0.032	0.029	5.344	-7.763	-10.541				
S4→S2	0.578 (0.526, 0.630)	0.573 (0.518, 0.628)	0.580 (0.505, 0.655)	0.001 (0.475)	0.001 (0.412)	0.002 (0.197)	0.000 (0.901)	0.001 (0.412)	0.002 (0.197)	0.765 (<0.001)	0.749 (<0.001)	47.469	52.108	4.638	0.050	0.055	3.797	-0.700	-10.219				
S4→S3	0.376 (0.325, 0.427)	0.381 (0.328, 0.435)	0.412 (0.339, 0.489)	0.004 (0.200)	0.002 (0.360)	0.004 (0.678)	0.004 (0.200)	0.002 (0.360)	0.004 (0.678)	0.851 (<0.001)	0.872 (<0.001)	50.721	45.771	-4.950	-0.081	-0.080	-2.670	-4.081	12.075				
S4→S4	0.511 (0.459, 0.565)	0.505 (0.449, 0.561)	0.458 (0.377, 0.535)	0.001 (0.653)	0.000 (0.800)	0.000 (0.316)	0.000 (0.800)	0.001 (0.653)	0.000 (0.316)	0.196 (<0.001)	0.292 (<0.001)	54.714	50.360	-4.354	0.043	0.040	-6.197	10.400	5.471				
S4→S5	0.359 (0.304, 0.402)	0.369 (0.320, 0.425)	0.439 (0.354, 0.510)	0.008 (0.038)	0.002 (0.277)	0.002 (0.472)	0.000 (0.901)	0.002 (0.277)	0.002 (0.472)	0.611 (<0.001)	0.654 (<0.001)	48.149	46.392	-2.757	-0.031	-0.042	0.819	-10.844	-1.458				
S4→S6	0.574 (0.523, 0.626)	0.569 (0.516, 0.625)	0.580 (0.505, 0.654)	0.000 (0.841)	0.000 (0.737)	-0.003 (0.387)	0.000 (0.841)	0.000 (0.737)	0.003 (0.387)	0.756 (<0.001)	0.734 (<0.001)	47.239	51.616	4.377	0.045	0.049	3.818	-0.524	-10.360				
S5→S1	0.566 (0.514, 0.618)	0.565 (0.512, 0.621)	0.590 (0.518, 0.666)	0.000 (0.804)	0.000 (0.692)	0.000 (0.755)	0.000 (0.804)	0.000 (0.692)	0.000 (0.755)	0.699 (<0.001)	0.707 (<0.001)	47.255	51.962	4.706	0.054	0.054	5.723	-2.805	-10.852				
S5→S2	0.628 (0.576, 0.678)	0.618 (0.562, 0.672)	0.598 (0.522, 0.670)	0.005 (0.205)	0.002 (0.331)	0.003 (0.366)	0.005 (0.205)	0.002 (0.331)	0.003 (0.366)	0.865 (<0.001)	0.875 (<0.001)	47.937	53.905	5.968	0.082	0.087	3.350	5.253	-9.651				
S5→S3	0.424 (0.372, 0.474)	0.424 (0.370, 0.476)	0.418 (0.341, 0.491)	0.001 (0.481)	0.002 (0.397)	-0.003 (0.277)	0.001 (0.481)	0.002 (0.397)	0.003 (0.277)	0.766 (<0.001)	0.750 (<0.001)	51.005	47.586	-3.419	-0.065	-0.060	-4.185	1.214	11.632				
S5→S4	0.562 (0.510, 0.613)	0.549 (0.493, 0.604)	0.492 (0.411, 0.567)	0.003 (0.339)	0.001 (0.857)	0.002 (0.279)	0.003 (0.339)	0.001 (0.857)	0.002 (0.279)	0.112 (0.018)	0.047 (0.317)	53.016	51.742	-1.275	0.046	0.051	-5.367	10.476	3.325				
S5→S5	0.344 (0.297, 0.395)	0.357 (0.307, 0.410)	0.409 (0.333, 0.489)	0.008 (0.025)	0.002 (0.185)	0.000 (0.994)	0.008 (0.025)	0.002 (0.185)	0.000 (0.994)	0.694 (<0.001)	0.772 (<0.001)	50.160	46.468	-3.692	-0.030	-0.035	-1.788	-6.193	-0.266				
S5→S6	0.615 (0.563, 0.665)	0.605 (0.550, 0.659)	0.569 (0.495, 0.642)	0.003 (0.361)	0.001 (0.427)	0.000 (0.782)	0.003 (0.361)	0.001 (0.427)	0.000 (0.782)	0.863 (<0.001)	0.869 (<0.001)	48.101	52.905	4.804	0.075	0.079	2.351	4.377	-8.614				
S6→S1	0.538 (0.486, 0.589)	0.537 (0.483, 0.595)	0.577 (0.504, 0.654)	0.000 (0.927)	0.000 (0.871)	0.000 (0.141)	0.000 (0.927)	0.000 (0.871)	0.000 (0.141)	0.551 (<0.001)	0.543 (<0.001)	48.217	51.272	3.055	0.034	0.034	5.497	-5.726	-9.734				
S6→S2	0.622 (0.570, 0.672)	0.594 (0.541, 0.646)	0.567 (0.494, 0.636)	0.006 (0.216)	0.001 (0.477)	0.000 (0.505)	0.006 (0.216)	0.001 (0.477)	0.000 (0.505)	0.746 (<0.001)	0.718 (<0.001)	48.199	52.034	3.835	0.042	0.047	-2.460	5.707	-5.830				
S6→S3	0.408 (0.357, 0.458)	0.411 (0.356, 0.463)	0.417 (0.342, 0.491)	0.002 (0.515)	0.001 (0.511)	0.004 (0.448)	0.002 (0.515)	0.001 (0.511)	0.004 (0.448)	0.807 (<0.001)	0.821 (<0.001)	50.929	46.894	-4.035	-0.064	-0.064	-4.496	0.474	10.988				
S6→S4	0.491 (0.439, 0.543)	0.484 (0.429, 0.540)	0.444 (0.365, 0.519)	0.000 (0.738)	0.000 (0.963)	0.002 (0.280)	0.000 (0.738)	0.000 (0.963)	0.002 (0.280)	0.459 (<0.001)	0.472 (<0.001)	52.519	49.232	-3.287	-0.007	-0.008	-6.572	7.247	7.603				
S6→S5	0.359 (0.309, 0.410)	0.369 (0.317, 0.422)	0.411 (0.338, 0.489)	0.005 (0.088)	0.003 (0.241)	0.002 (0.761)	0.005 (0.088)	0.003 (0.241)	0.002 (0.761)	0.829 (<0.001)	0.850 (<0.001)	50.593	46.457	-4.135	-0.060	-0.064	-2.699	-5.700	4.250				
S6→S6	0.567 (0.514, 0.622)	0.537 (0.484, 0.591)	0.485 (0.426, 0.566)	0.003 (0.473)	0.001 (0.910)	0.000 (0.603)	0.003 (0.473)	0.001 (0.910)	0.000 (0.603)	0.735 (<0.001)	0.664 (<0.001)	48.152	52.571	4.418	0.087	0.092	0.512	6.227	-8.692				

Appendix B

EEG-PD-NLME Loss Derivations

B.1 Proportional Model MLE Loss Derivation

The derivation of the loss function begins with the normally distributed residuals of $\hat{n}$:

$$\hat{n}_i + \epsilon_i = n_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

Consider, firstly, the proportional model. $\hat{n}_i$ can be substituted with the output of f_ϕ given a specific EEG segment x_i , and n_i can be substituted with the output of the PD model from Equation 6.1, assuming that segment i belongs to bolus b and patient p . This gives:

$$f_\phi(x_i) + \epsilon_i = \frac{N_b C_{50,p}^{\gamma_p}}{C_{50,p}^{\gamma_p} + C_{e,i}^{\gamma_p}}$$

This can then be rearranged as in Equation 6.2. This gives:

$$C_{e,i} = g_\eta(n_i) = g_\eta(f_\phi(x_i) + \epsilon_i) = C_{50,p} \left(\frac{N_b - f_\phi(x_i) - \epsilon_i}{f_\phi(x_i) + \epsilon_i} \right)^{\frac{1}{\gamma_p}}$$

Using the fact that ϵ_i is assumed to be distributed as a Gaussian, a probability density function (PDF) for n_i can be calculated, given an observed EEG segment x_i :

$$P(n_i | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(n_i - f_\phi(x_i))^2}{2\sigma^2}\right)$$

However, n_i is not directly observed. A change of variables needs to take place to calculate $P(C_{e,i} | x_i)$. This can be calculated using g_η , which takes n_i as an input and outputs $C_{e,i}$. The general formula for a transform of variables in this context is:

$$P(y | x) = p(g^{-1}(y) | x) \left| \frac{d}{dy} g^{-1}(y) \right|$$

g_η is differentiable, invertible, and bijective, hence it is valid to use this change of variables formula. Substituting in the relevant functions and variables for this context gives:

$$P(C_{e,i} | x_i) = p(g_\eta^{-1}(C_{e,i}) | x_i) \left| \frac{d}{dC_{e,i}} g_\eta^{-1}(C_{e,i}) \right|$$

The first part, $p(g_\eta^{-1}(C_{e,i}) | x_i) = p(n_i | x_i)$ is the PDF that was previously derived. The second part requires the derivative of $g_\eta^{-1}(C_{e,i})$. $g_\eta^{-1}(C_{e,i})$ is the forward PD model in Equation 6.1, and it can be differentiated with respect to $C_{e,i}$:

$$\frac{d}{dC_{e,i}} g_\eta^{-1}(C_{e,i}) = -\frac{N_b C_{50,p}^{\gamma_p} \gamma_p C_{e,i}^{\gamma_p-1}}{(C_{50,p}^{\gamma_p} + C_{e,i}^{\gamma_p})^2}.$$

Combining them back into the PDF gives:

$$P(C_{e,i} | x_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(g_\eta^{-1}(C_{e,i}) - f_\phi(x_i))^2}{2\sigma^2}\right) \cdot \frac{N_b C_{50,p}^{\gamma_p} \gamma_p C_{e,i}^{\gamma_p-1}}{(C_{50,p}^{\gamma_p} + C_{e,i}^{\gamma_p})^2}$$

Negative log-likelihood for a single instance is therefore:

$$\begin{aligned} \ell_i(C_{e,i} | x_i) &= \frac{1}{2} \log(2\pi\sigma^2) + \frac{(g_\eta^{-1}(C_{e,i}) - f_\phi(x_i))^2}{2\sigma^2} - \log(N_b) - \gamma_p \log(C_{50,p}) \\ &\quad - \log(\gamma_p) - (\gamma_p - 1) \log(C_{e,i}) + 2 \log(C_{50,p}^{\gamma_p} + C_{e,i}^{\gamma_p}) \end{aligned}$$

$b(i)$ denotes the bolus associated with input x_i , and $p(i)$ denotes the patient associated with input x_i . The loss function for the whole dataset, calculated by summing the negative log-likelihood across all data points is:

$$\begin{aligned} \mathcal{L}(\eta, \phi) &= \sum_{i=1}^N \ell_i(C_{e,i} | x_i) \\ &= \frac{N}{2} \log(2\pi\sigma^2) + \sum_{i=1}^N \left[\frac{(g_\eta^{-1}(C_{e,i}) - f_\phi(x_i))^2}{2\sigma^2} - \log(N_{b(i)}) - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ &\quad \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right] \end{aligned}$$

The one remaining decision point for this loss function is what σ should be set to. In this study, three options were explored:

- Assume the error in $\hat{n}_i$ will be homoscedastic, and set a specific value of σ for this.
- Assume the error in $\hat{n}_i$ will be homoscedastic, but calculate an optimal value for it using the summed negative log-likelihood.
- Assume the error in $\hat{n}_i$ will be heteroscedastic but proportional to the value of $\hat{n}_i$.

In the first case, the first term in the loss function becomes a constant and therefore can be removed. The full loss function in this case thus becomes:

$$\mathcal{L}(\eta, \phi) = \sum_{i=1}^N \left[\frac{(g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2}{2\sigma^2} - \log(N_{b(i)}) - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

In this case, σ becomes a tunable parameter that determines the balance between fitting the fixed effects ϕ and the random effects η . When σ is high, the first term inside the summation shrinks, allowing the Jacobian random effect terms to have relatively greater influence. When σ is small, the first term increases, hence the ability of f_{ϕ} to fit n is more highly weighted. This matches the intuition for how σ operates in a PDF sense: a higher σ allows for larger errors in the estimation of n .

In the second case, the partial derivative of the likelihood function $\mathcal{L}(\eta, \phi)$ is taken with respect to σ , and this is set to 0 in order to determine the value of σ that maximises likelihood:

$$\frac{\partial \mathcal{L}(\eta, \phi)}{\partial \sigma} = \frac{N}{\sigma} - \frac{\sum_{i=1}^N (g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2}{\sigma^3} = 0$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^N (g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2}{N}$$

Substituting this back into the original formula and removing constants gives a final loss:

$$\mathcal{L}(\eta, \phi) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{i=1}^N (g_{\eta}^{-1}(C_{e,i}) - f_{\phi}(x_i))^2 \right) + \sum_{i=1}^N \left[-\log(N_{b(i)}) - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

In the third case, the variance σ^2 is taken to be proportional to the square of n (i.e. $\sigma^2 = kn^2$). This gives:

$$\mathcal{L}(\eta, \phi) = \sum_{i=1}^N \left[\frac{1}{2} \log(2\pi k n_i^2) + \frac{(n_i - f_\phi(x_i))^2}{2k n_i^2} - \log(N_{b(i)}) - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

The function can be similarly differentiated with respect to k , the result set to zero, and solved as previously to determine an optimal k value:

$$\frac{\partial \mathcal{L}(\eta, \phi)}{\partial k} = \frac{N}{2k} - \sum_{i=1}^N \frac{(n_i - f_\phi(x_i))^2}{2k^2 n_i^2} = 0$$

$$k = \frac{1}{N} \sum_{i=1}^N \frac{(n_i - f_\phi(x_i))^2}{n_i^2}$$

Substituting this back in gives another potential loss function:

$$\mathcal{L}(\eta, \phi) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{i=1}^N \frac{(n_i - f_\phi(x_i))^2}{n_i^2} \right) + \sum_{i=1}^N \left[\log(n_i) - \log N_{b(i)} - \gamma_{p(i)} \log C_{50,p(i)} \right. \\ \left. - \log \gamma_{p(i)} - (\gamma_{p(i)} - 1) \log C_{e,i} + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

B.2 Absolute Model MLE Loss Derivation

The derivations of the loss functions for the absolute model are similar, using Equation 6.3 and Equation 6.4 as a basis for the calculation.

The change of variables is carried out in a similar manner, although the different equation for g_η means that the derivative of g_η^{-1} is different:

$$\frac{d}{dC_{e,i}} g_\eta^{-1}(C_{e,i}) = - \frac{C_{50,p}^{\gamma_p} \gamma_p C_{e,i}^{\gamma_p - 1}}{(C_{50,p}^{\gamma_p} + C_{e,i}^{\gamma_p})^2}$$

Note that this is almost the same as the derivative used in the proportional model, differing only by a factor of N_b . The final loss function is the same as in the proportional case, just with this factor removed.

$$\mathcal{L}(\eta, \phi) = \sum_{i=1}^N \ell_i(C_{e,i} \mid x_i)$$

$$= \frac{N}{2} \log(2\pi\sigma^2) + \sum_{i=1}^N \left[\frac{(g_\eta^{-1}(C_{e,i}) - f_\phi(x_i))^2}{2\sigma^2} - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

The homoscedastic choices of σ are the same as before, although the fact that n can become negative causes the heteroscedastic case to fail, as this loss function takes a logarithm of n .

The remaining derivation is similar to the proportional case, differing only by a factor of N_b . For an arbitrarily chosen σ , this loss function is hence:

$$\mathcal{L}(\eta, \phi) = \sum_{i=1}^N \left[\frac{(g_\eta^{-1}(C_{e,i}) - f_\phi(x_i))^2}{2\sigma^2} - \gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

For an optimally chosen σ , the loss function is:

$$\mathcal{L}(\eta, \phi) = \frac{N}{2} \log \left(\frac{1}{N} \sum_{i=1}^N (g_\eta^{-1}(C_{e,i}) - f_\phi(x_i))^2 \right) + \sum_{i=1}^N \left[-\gamma_{p(i)} \log(C_{50,p(i)}) \right. \\ \left. - \log(\gamma_{p(i)}) - (\gamma_{p(i)} - 1) \log(C_{e,i}) + 2 \log(C_{50,p(i)}^{\gamma_{p(i)}} + C_{e,i}^{\gamma_{p(i)}}) \right]$$

Appendix C

EEG-PD-NLME Supplementary Material

This appendix contains plots of the random effect distributions for some of the experiments in Chapter 6.

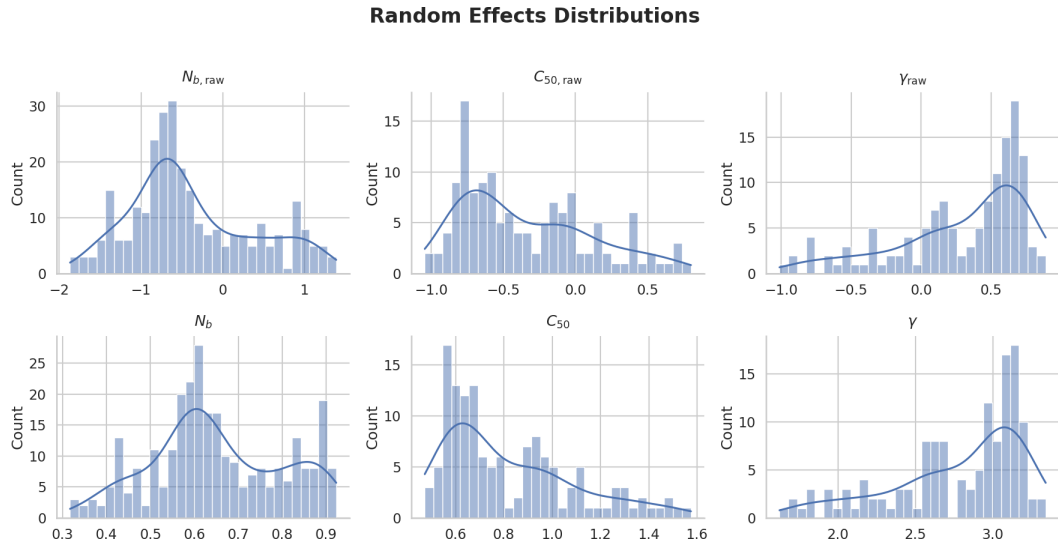


Figure C.1: Experiment 1 random effect distribution.

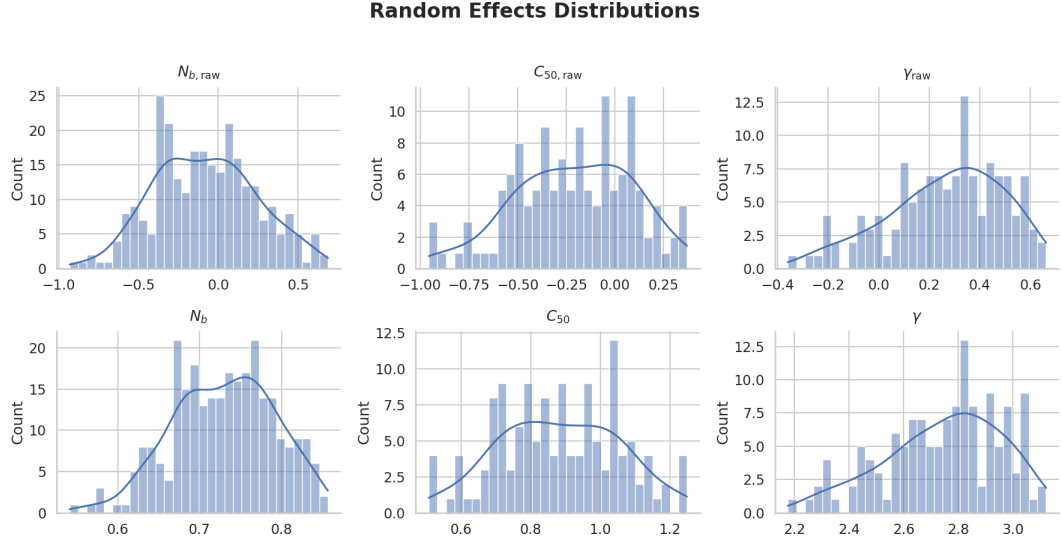


Figure C.2: Random effect distribution for 120-second ResNet-18 model.

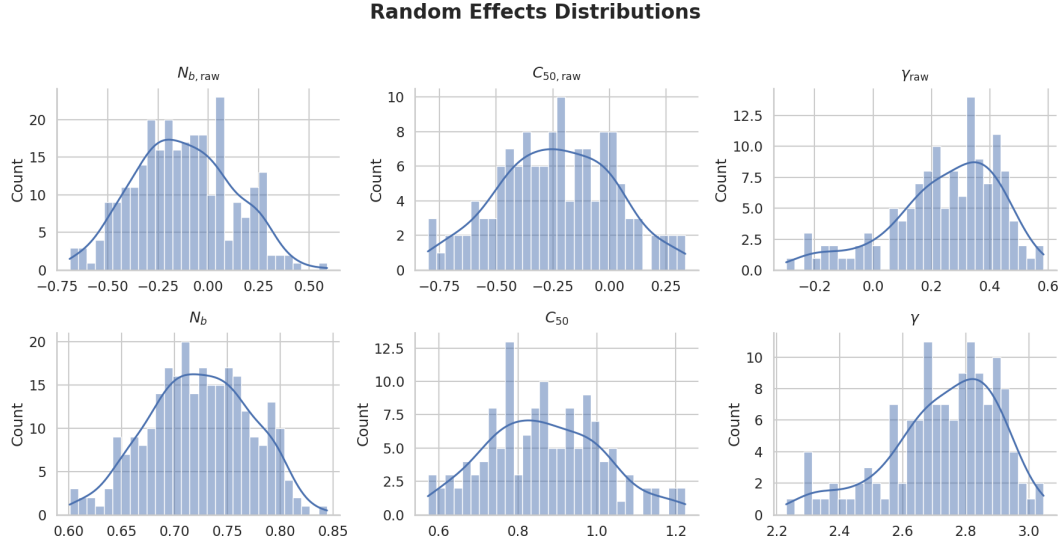


Figure C.3: Random effect distribution for 120-second ResNet-18 model with pre-surgical regularisation ($\lambda_b = 0.5$).

Appendix D

Representation Learning Supplementary Material

This appendix contains raw tables featuring the results of model performance on the unified validation framework, using the combined train and validation set. The models in this appendix are those in Chapter 7 (representation learning). The criteria for bold and italic text are identical to those of Appendix A.

Table D.1: Baseline binary classification model.

Model	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.		B2. Within-patient corr.		C. Surgical change			D. Feat bolus corr.		E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
$C_I = 1.0$	0.427 (0.372, 0.480)	0.406 (0.352, 0.460)	0.399 (0.326, 0.474)	0.001 (0.427)	0.000 (0.985)	-0.004 (0.407)	0.406 (<0.001)	0.388 (<0.001)	0.062	0.040	37.630	48.879	11.249	-0.055	-0.046	-23.339	-16.977	-0.918
$C_I = 1.5$	0.608 (0.557, 0.662)	0.605 (0.552, 0.660)	0.570 (0.492, 0.647)	0.003 (0.022)	0.003 (0.040)	0.004 (0.387)	-0.143 (0.002)	-0.188 (<0.001)	-0.008	-0.021	30.783	51.716	20.963	-0.102	-0.097	1.363	13.515	-1.379
$C_I = 2.0$	0.577 (0.524, 0.629)	0.541 (0.483, 0.596)	0.521 (0.443, 0.597)	0.002 (0.025)	-0.001 (0.395)	-0.001 (0.893)	-0.101 (0.033)	-0.138 (<0.001)	0.055	-0.014	21.282	51.309	30.027	-0.122	-0.116	-7.449	-2.902	-3.452
$C_I = 2.5$	0.566 (0.512, 0.620)	0.520 (0.466, 0.574)	0.508 (0.432, 0.584)	0.003 (0.016)	-0.001 (0.483)	-0.001 (0.877)	-0.059 (0.288)	-0.087 (0.065)	0.006	-0.049	22.803	50.420	27.617	-0.091	-0.085	-15.121	-8.075	-4.456
$C_I = 3.0$	0.507 (0.513, 0.621)	0.538 (0.484, 0.596)	0.551 (0.481, 0.624)	0.000 (0.039)	-0.001 (0.336)	0.003 (0.297)	-0.099 (0.035)	-0.105 (0.026)	0.009	-0.035	20.523	49.393	28.869	-0.063	-0.057	-15.988	-5.660	-6.042
$C_I = 3.5$	0.577 (0.524, 0.629)	0.544 (0.491, 0.600)	0.530 (0.457, 0.607)	0.005 (0.009)	-0.002 (0.195)	-0.001 (0.644)	-0.128 (0.007)	-0.155 (0.001)	0.021	-0.060	21.600	48.711	27.111	-0.116	-0.110	-16.426	-11.332	-6.955
$C_I = 4.0$	0.581 (0.529, 0.631)	0.552 (0.499, 0.606)	0.549 (0.479, 0.619)	0.004 (0.006)	-0.002 (0.100)	-0.003 (0.294)	-0.153 (0.001)	-0.159 (<0.001)	-0.008	-0.062	22.360	48.120	25.760	-0.123	-0.111	-19.379	-7.715	-9.006
Smooth y	0.405 (0.411, 0.519)	0.414 (0.362, 0.470)	0.390 (0.322, 0.463)	0.002 (0.287)	0.001 (0.382)	0.013 (0.136)	0.093 (0.049)	0.042 (0.380)	-0.088	-0.104	37.110	52.998	15.888	-0.055	-0.049	-20.468	-13.081	-4.065

Table D.2: Binary classification model with varying adversarial head strength (λ_p and λ_d).

λ_p	λ_d	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.		B2. Within-patient corr.		C. Surgical change		D. Font bolus corr.		E. Spectral regression coeff.			
		>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Pre-surg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β
0	0	0.622 (0.571, 0.674)	0.604 (0.549, 0.657)	0.535 (0.456, 0.614)	0.012 (0.003)	0.005 (0.032)	-0.001 (0.942)	-0.204 (<0.001)	-0.257 (<0.001)	-0.015	-0.043	30.989	53.988	23.000	-0.143	-0.134	-3.364	6.741	1.128
0.1	0	0.594 (0.543, 0.645)	0.569 (0.516, 0.624)	0.500 (0.423, 0.582)	0.007 (0.067)	0.003 (0.531)	0.003 (0.220)	-0.182 (<0.001)	-0.280 (<0.001)	0.028	0.025	38.666	52.047	13.381	-0.054	-0.036	-7.181	7.199	-3.579
1	0	0.568 (0.514, 0.622)	0.526 (0.470, 0.587)	0.459 (0.376, 0.539)	0.001 (0.319)	0.002 (0.580)	0.006 (0.066)	-0.037 (0.439)	-0.037 (0.440)	0.028	0.068	53.028	50.056	-2.973	-0.000	0.005	-5.902	12.229	2.778
0	10	0.560 (0.505, 0.613)	0.533 (0.477, 0.589)	0.499 (0.417, 0.579)	0.001 (0.082)	0.000 (0.507)	-0.001 (0.711)	-0.109 (0.021)	-0.073 (0.122)	-0.061	0.018	40.540	50.253	9.713	-0.055	-0.045	-21.801	5.814	-6.183
0.1	10	0.509 (0.457, 0.563)	0.481 (0.429, 0.534)	0.473 (0.402, 0.544)	0.004 (0.046)	0.000 (0.319)	0.001 (0.894)	0.288 (<0.001)	0.319 (<0.001)	0.118	0.130	40.028	49.483	9.455	-0.005	-0.076	-22.636	-5.452	2.247
1	10	0.566 (0.514, 0.617)	0.540 (0.485, 0.593)	0.551 (0.476, 0.631)	0.004 (0.147)	0.001 (0.520)	0.001 (0.401)	-0.288 (<0.001)	-0.342 (<0.001)	-0.147	-0.103	37.786	51.781	13.995	-0.023	-0.021	-7.394	5.728	-10.323
0	100	0.555 (0.503, 0.608)	0.521 (0.471, 0.573)	0.497 (0.425, 0.572)	0.008 (0.010)	0.000 (0.242)	-0.006 (0.947)	-0.059 (0.216)	-0.046 (0.331)	0.003	-0.018	38.002	49.288	11.287	-0.060	-0.055	-23.670	-4.128	-7.290
0.1	100	0.563 (0.508, 0.617)	0.549 (0.493, 0.606)	0.509 (0.426, 0.591)	0.002 (0.280)	0.001 (0.626)	-0.002 (0.480)	-0.039 (0.405)	-0.019 (0.684)	0.095	0.126	49.750	50.210	0.460	-0.043	-0.036	-0.170	13.994	7.795
1	100	0.552 (0.499, 0.606)	0.526 (0.470, 0.587)	0.504 (0.423, 0.582)	0.001 (0.304)	0.000 (0.968)	-0.002 (0.699)	-0.021 (0.662)	-0.006 (0.892)	-0.005	0.072	47.485	50.323	2.838	-0.078	-0.065	-12.338	14.070	-0.865
0	1000	0.570 (0.517, 0.621)	0.549 (0.491, 0.605)	0.508 (0.427, 0.585)	0.005 (0.148)	0.001 (0.504)	-0.006 (0.543)	-0.030 (0.532)	-0.037 (0.430)	0.046	0.080	56.169	51.990	-4.179	-0.108	-0.111	-4.599	10.942	6.719
0.1	1000	0.515 (0.462, 0.568)	0.487 (0.434, 0.541)	0.482 (0.414, 0.557)	-0.000 (0.458)	-0.000 (0.846)	-0.001 (0.802)	0.034 (0.176)	0.004 (0.938)	0.112	0.061	40.703	48.155	7.452	-0.039	-0.038	-14.763	4.429	-6.115
1	1000	0.484 (0.431, 0.536)	0.475 (0.415, 0.531)	0.490 (0.412, 0.569)	0.001 (0.663)	0.000 (0.949)	-0.001 (0.729)	0.194 (<0.001)	0.131 (0.005)	0.124	0.056	47.807	50.257	2.450	-0.074	-0.076	-1.539	-7.337	2.491
0	10000	0.570 (0.518, 0.623)	0.540 (0.485, 0.598)	0.489 (0.405, 0.568)	0.003 (0.300)	0.000 (0.815)	0.001 (0.185)	-0.211 (<0.001)	-0.233 (<0.001)	-0.136	-0.081	43.185	52.666	9.481	0.026	0.027	-12.873	12.072	-5.418
0.1	10000	0.476 (0.421, 0.529)	0.491 (0.437, 0.547)	0.494 (0.418, 0.570)	-0.000 (0.922)	0.002 (0.372)	-0.006 (0.526)	-0.050 (0.292)	-0.050 (0.294)	-0.136	-0.100	38.550	50.366	12.386	-0.079	-0.072	-13.197	2.651	-12.541
1	10000	0.526 (0.470, 0.578)	0.524 (0.468, 0.581)	0.536 (0.456, 0.616)	-0.000 (0.989)	-0.000 (0.888)	-0.000 (0.808)	0.097 (0.040)	0.066 (0.100)	0.155	0.150	50.140	49.767	-0.672	-0.102	-0.107	2.438	18.256	3.327

Table D.3: NLMF with varying adversarial head strength.

λ_d	A1. AUROC (95% CI)			A2. Δ AUROC (p)			B1. Between-patient Openbis corr.			B2. Within-patient corr.			C. Surgical change			D. Feat bolus corr.			E. Spectral regression coeff.		
	>0	>3	>6	>0	>3	>6	Pearson r (p)	Spearman ρ (p)	Mean r	Mean ρ	Presurg.	Surg.	Diff.	Mean r	Mean ρ	δ	α	β			
0	0.407 (0.355, 0.461)	0.410 (0.358, 0.466)	0.416 (0.349, 0.484)	0.004 (0.270)	0.004 (0.265)	0.001 (0.507)	0.368 (<0.001)	0.336 (<0.001)	0.089	0.080	49.316	48.853	-0.464	-0.493	-0.526	-4.710	-3.185	1.614			
0.1	0.518 (0.464, 0.569)	0.542 (0.484, 0.598)	0.494 (0.414, 0.578)	-0.000 (0.884)	0.000 (0.528)	0.003 (0.596)	0.038 (0.426)	-0.034 (0.471)	0.053	0.037	48.410	49.483	1.073	-0.479	-0.520	-1.287	-0.696	3.024			
1	0.391 (0.338, 0.441)	0.394 (0.341, 0.446)	0.429 (0.353, 0.506)	0.003 (0.264)	0.003 (0.323)	-0.002 (0.784)	0.077 (<0.001)	0.704 (<0.001)	0.226	0.200	49.668	47.949	-1.719	-0.271	-0.278	-2.091	-10.493	11.410			
10	0.459 (0.403, 0.513)	0.468 (0.411, 0.522)	0.436 (0.365, 0.503)	0.001 (0.520)	-0.000 (0.605)	0.004 (0.178)	0.510 (<0.001)	0.471 (<0.001)	0.375	0.319	55.590	50.064	-0.826	-0.104	-0.094	-3.681	1.301	13.959			
100	0.454 (0.400, 0.506)	0.481 (0.425, 0.539)	0.511 (0.429, 0.597)	0.000 (1.000)	0.000 (1.000)	0.208 (<0.001)	0.208 (<0.001)	0.192 (<0.001)	0.113	0.044	51.885	49.121	-2.764	-0.081	-0.087	10.660	-16.438	6.124			
1000	0.464 (0.411, 0.518)	0.497 (0.442, 0.551)	0.514 (0.434, 0.593)	0.001 (0.487)	0.000 (0.620)	-0.002 (0.489)	0.199 (<0.001)	0.178 (<0.001)	0.191	0.120	56.199	48.925	-7.273	-0.070	-0.075	16.536	-13.870	12.262			
10000	0.585 (0.537, 0.635)	0.560 (0.512, 0.610)	0.548 (0.481, 0.608)	0.004 (0.199)	0.001 (0.514)	-0.001 (0.687)	-0.690 (<0.001)	-0.643 (<0.001)	-0.479	-0.368	78.714	87.175	8.461	0.299	0.306	1.521	12.388	-18.191			

References

- [1] S. Chen, Y. Zolo, L. Ngulube, M. Isiagi, and S. Maswime, “Global surgery and climate change: how global surgery can prioritise both the health of the planet and its people,” *BMC surgery*, vol. 25, no. 1, p. 21, 2025.
- [2] L. Uhrig, S. Dehaene, and B. Jarraya, “Cerebral mechanisms of general anesthesia,” in *Annales francaises d’anesthesie et de reanimation*, vol. 33, no. 2. Elsevier, 2014, pp. 72–82.
- [3] E. N. Brown, K. J. Pavone, and M. Naranjo, “Multimodal general anesthesia: theory and practice,” *Anesthesia & analgesia*, vol. 127, no. 5, pp. 1246–1258, 2018.
- [4] A. R. Caffrey and E. P. Borrelli, “The art and science of drug titration,” *Therapeutic advances in drug safety*, vol. 11, p. 2042098620958910, 2020.
- [5] M. X. Cohen, “Where does eeg come from and what does it mean?” *Trends in neurosciences*, vol. 40, no. 4, pp. 208–218, 2017.
- [6] A. W. Gelb, W. W. Morriss, W. Johnson, and A. F. Merry, “World health organization-world federation of societies of anaesthesiologists (who-wfsa) international standards for a safe practice of anesthesia,” *Canadian Journal of Anesthesia/Journal canadien d’anesthésie*, vol. 65, no. 6, pp. 698–708, 2018.
- [7] Y. Sun, C. Wei, V. Cui, M. Xiu, and A. Wu, “Electroencephalography: clinical applications during the perioperative period,” *Frontiers in medicine*, vol. 7, p. 251, 2020.
- [8] B. Musizza and S. Ribaric, “Monitoring the depth of anaesthesia,” *Sensors*, vol. 10, no. 12, pp. 10 896–10 935, 2010.
- [9] L. A. Santa Cruz Mercado, R. Liu, K. M. Bharadwaj, J. J. Johnson, R. Gutierrez, P. Das, G. Balanza, H. Deng, A. Pandit, T. A. Stone *et al.*, “Association of intraoperative opioid administration with postoperative pain and opioid use,” *JAMA surgery*, vol. 158, no. 8, pp. 854–864, 2023.
- [10] T. J. Gan, “Poorly controlled postoperative pain: prevalence, consequences, and prevention,” *Journal of pain research*, pp. 2287–2298, 2017.

- [11] N. Ferry, L. E. Hancock, J. M. Hendrix, and S. T. Dhanjal, “Opioid anesthesia,” in *StatPearls [Internet]*. StatPearls Publishing, 2025.
- [12] P. S. García, M. Kreuzer, D. Hight, and J. W. Sleight, “Effects of noxious stimulation on the electroencephalogram during general anaesthesia: a narrative review and approach to analgesic titration,” *British journal of anaesthesia*, vol. 126, no. 2, pp. 445–457, 2021.
- [13] J. Altés, J. F. Valencia, S. Jaramillo, and P. L. Gambùs, “Prediction model of response to standardized painful stimulation during general anesthesia: Applying machine learning to the eeg signal,” in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021, pp. 1–6.
- [14] O. M. Abdel Deen, S.-Z. Fan, and J.-S. Shieh, “A multimodal deep learning approach to intraoperative nociception monitoring: Integrating electroencephalogram, photoplethysmography, and electrocardiogram,” *Sensors*, vol. 25, no. 4, p. 1150, 2025.
- [15] O. M. A. Deen, S.-Z. Fan, F. Doctor, and J.-S. Shieh, “A quantum multilayer perceptron for intraoperative nociception prediction,” *IEEE Transactions on Artificial Intelligence*, 2026.
- [16] T. Ledowski and I. Schmitz-Rode, “Predicting acute postoperative pain by the qnox score at the end of surgery: a prospective observational study,” *British Journal of Anaesthesia*, vol. 124, no. 2, pp. 222–226, 2020.
- [17] Y. Wu, “Extraction of objective and quantitative indexes of pain, anxiety, depression and other brain function states from electroencephalogram,” *China Med Eng*, vol. 25, no. 4, pp. 1–7, 2017.
- [18] M. Saeidi, W. Karwowski, F. V. Farahani, K. Fiok, R. Taiar, P. A. Hancock, and A. Al-Juaid, “Neural decoding of eeg signals with machine learning: a systematic review,” *Brain sciences*, vol. 11, no. 11, p. 1525, 2021.
- [19] R. Dhiman *et al.*, “Machine learning techniques for electroencephalogram based brain-computer interface: A systematic literature review,” *Measurement: Sensors*, vol. 28, p. 100823, 2023.
- [20] Y. Si, “Machine learning applications for electroencephalograph signals in epilepsy: a quick review,” *Acta Epileptologica*, vol. 2, no. 1, p. 5, 2020.
- [21] G. R. L. Guevara, I. Vargas-Huicochea, and A. C. Rodriguez-Machain, “Systematic review of machine learning and deep learning models for eeg-based detection of depression,” *Journal of Psychiatric Research*, 2026.

- [22] K. A. I. Aboalayon, M. Faezipour, W. S. Almuhammadi, and S. Moslehpour, "Sleep stage classification using eeg signal analysis: a comprehensive survey and new investigation," *Entropy*, vol. 18, no. 9, p. 272, 2016.
- [23] Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert, "Deep learning-based electroencephalography analysis: a systematic review," *Journal of neural engineering*, vol. 16, no. 5, p. 051001, 2019.
- [24] X. Xu, X.-F. Zhang, Z.-H. Yu, J. Liu, L. Nie, and J.-L. Song, "Comparison of surgical pleth index-guided analgesia versus conventional analgesia technique in general anesthesia surgeries: A systematic review and meta-analysis," *Journal of Clinical Anesthesia*, vol. 103, p. 111800, 2025.
- [25] J. Lv, J. Zhang, K. Zhang, and J. Zheng, "Predictive value of eeg-derived pain threshold index for acute postoperative pain in children," *Frontiers in pediatrics*, vol. 10, p. 1052532, 2022.
- [26] Y. Son, "Molecular mechanisms of general anesthesia," *Korean journal of anesthesiology*, vol. 59, no. 1, pp. 3–8, 2010.
- [27] E. N. Brown, P. L. Purdon, and C. J. Van Dort, "General anesthesia and altered states of arousal: a systems neuroscience analysis," *Annual review of neuroscience*, vol. 34, no. 1, pp. 601–628, 2011.
- [28] T. B. Folino, E. Mucio, A. O. Safadi, and L. J. Parks, "Propofol," in *StatPearls [Internet]*. StatPearls Publishing, 2023.
- [29] L. Voss and J. Sleight, "Monitoring consciousness: the current status of eeg-based depth of anaesthesia monitors," *Best practice & research Clinical anaesthesiology*, vol. 21, no. 3, pp. 313–325, 2007.
- [30] H. Liu, Z. Yang, Y. Chen, F. Yang, X. Cao, G. Zhou, and Y. Zhang, "Neural oscillations and memory: unraveling the mechanisms of anesthesia-induced amnesia," *Frontiers in Neuroscience*, vol. 18, p. 1492103, 2024.
- [31] B. Riedel, J. Dubowitz, J. Yeung, S. Jhanji, S. Kheterpal, and M. S. Avidan, "On the horns of a dilemma: choosing total intravenous anaesthesia or volatile anaesthesia," pp. 284–289, 2022.
- [32] M. Ramne, M. Elam, L. S. Löken, and M. Ortiz-Catalan, "Neurophysiology of pain for non-neurophysiologists: a systematic review," 2023.
- [33] S. A. Armstrong and M. J. Herr, "Physiology, nociception," 2019.
- [34] I. Ghanty and S. Schraag, "The quantification and monitoring of intra-operative nociception levels in thoracic surgery: a review," *Journal of Thoracic Disease*, vol. 11, no. 9, p. 4059, 2019.

- [35] G. Lichtner, R. Auksztulewicz, E. Kirilina, H. Velten, D. Mavrodis, M. Scheel, F. Blankenburg, and F. von Dincklage, “Effects of propofol anesthesia on the processing of noxious stimuli in the spinal cord and the brain,” *Neuroimage*, vol. 172, pp. 642–653, 2018.
- [36] R. Ivascu, L. I. Torsin, L. Hostiuc, C. Nitipir, D. Corneci, and M. Dutu, “The surgical stress response and anesthesia: a narrative review,” *Journal of Clinical Medicine*, vol. 13, no. 10, p. 3017, 2024.
- [37] M. Hirose, H. Okutani, K. Hashimoto, R. Ueki, N. Shimode, N. Kariya, Y. Takao, and T. Tatara, “Intraoperative assessment of surgical stress response using nociception monitor under general anesthesia and postoperative complications: a narrative review,” *Journal of clinical medicine*, vol. 11, no. 20, p. 6080, 2022.
- [38] C. L. Wu, M. Naqibuddin, A. J. Rowlingson, S. A. Lietman, R. M. Jermyn, and L. A. Fleisher, “The effect of pain on health-related quality of life in the immediate postoperative period,” *Anesthesia & Analgesia*, vol. 97, no. 4, pp. 1078–1085, 2003.
- [39] Y. Kivrak, H. Kose-Ozlece, M. F. Ustundag, and M. Asoglu, “Pain perception: predictive value of sex, depression, anxiety, somatosensory amplification, obesity, and age,” *Neuropsychiatric disease and treatment*, pp. 1913–1918, 2016.
- [40] L. Crawford, E. Mills, N. Meylakh, P. Macey, V. Macefield, and L. Henderson, “Brain activity changes associated with pain perception variability,” *Cerebral Cortex*, vol. 33, no. 7, pp. 4145–4155, 2023.
- [41] S. J. Gibson and M. Farrell, “A review of age differences in the neurophysiology of nociception and the perceptual experience of pain,” *The Clinical journal of pain*, vol. 20, no. 4, pp. 227–239, 2004.
- [42] C. J. Paller, C. M. Campbell, R. R. Edwards, and A. S. Dobs, “Sex-based differences in pain perception and treatment,” *Pain medicine*, vol. 10, no. 2, pp. 289–299, 2009.
- [43] R. L. van Boekel, K. C. Vissers, R. van der Sande, E. Bronkhorst, J. G. Lerou, and M. A. Steegers, “Moving beyond pain scores: Multidimensional pain assessment is essential for adequate pain management after surgery,” *PLoS One*, vol. 12, no. 5, p. e0177345, 2017.
- [44] E. M. Rosier, M. J. Iadarola, and R. C. Coghill, “Reproducibility of pain measurement and pain perception,” *Pain*, vol. 98, no. 1-2, pp. 205–216, 2002.

- [45] T. L. Sandritter, M. McLaughlin, M. Artman, and J. Lowry, “The interplay between pharmacokinetics and pharmacodynamics,” *Pediatrics in Review*, vol. 38, no. 5, pp. 195–206, 2017.
- [46] L. Z. Benet and P. Zia-Amirhosseini, “Basic principles of pharmacokinetics,” *Toxicologic pathology*, vol. 23, no. 2, pp. 115–123, 1995.
- [47] J. Bovill and P. Sebel, “Pharmacokinetics of high-dose fentanyl: a study in patients undergoing cardiac surgery,” *British journal of anaesthesia*, vol. 52, no. 8, pp. 795–801, 1980.
- [48] M. Holz and A. Fahr, “Compartment modeling,” *Advanced Drug Delivery Reviews*, vol. 48, no. 2-3, pp. 249–264, 2001.
- [49] M. Bardol, S. Pan, S. M. Walker, J. F. Standing, and J. M. Dawes, “Pharmacokinetic pharmacodynamic modeling of analgesics and sedatives in children,” *Pediatric Anesthesia*, vol. 33, no. 10, pp. 781–792, 2023.
- [50] J. Lötsch, “Pharmacokinetic–pharmacodynamic modeling of opioids,” *Journal of pain and symptom management*, vol. 29, no. 5, pp. 90–103, 2005.
- [51] F. Hof and L. Bridge, “Exact solutions and equi-dosing regimen regions for multi-dose pharmacokinetics models with transit compartments,” *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 48, no. 1, pp. 99–131, 2021.
- [52] A. Aranake, G. Mashour, and M. Avidan, “Minimum alveolar concentration: ongoing relevance and clinical utility,” *Anaesthesia*, vol. 68, no. 5, pp. 512–522, 2013.
- [53] S. A. Lobo, J. Ojeda, A. Dua, K. Singh, and J. Lopez, “Minimum alveolar concentration,” 2018.
- [54] T. Rana, “Minimum alveolar concentration,” in *Introduction to Diseases, Diagnosis, and Management of Dogs and Cats*. Elsevier, 2023.
- [55] R. Kennedy, M. McKellow, R. French, and J. Sleight, “Sevoflurane end-tidal to effect-site equilibration in women determined by response to laryngeal mask airway insertion,” *Anesthesia & Analgesia*, vol. 117, no. 4, pp. 786–791, 2013.
- [56] J. A. Rogers, H. Maas, and A. P. Pitarch, “An introduction to causal inference for pharmacometricians,” *CPT: Pharmacometrics & Systems Pharmacology*, vol. 12, no. 1, pp. 27–40, 2023.

- [57] T. W. Schnider, C. F. Minto, and M. Filipovic, “The drug titration paradox: correlation of more drug with less effect in clinical data,” *Clinical Pharmacology & Therapeutics*, vol. 110, no. 2, pp. 401–408, 2021.
- [58] S. C. Goulooze, E. H. Krekels, C. A. Knibbe, and M. van Noort, “The drug titration paradox in the presence of intra-individual variation: can we estimate the true concentration-effect relationship?” *The AAPS Journal*, vol. 27, no. 3, p. 70, 2025.
- [59] N. R. Kristensen and H. Agersø, “Dose/exposure–response modeling in dose titration trials: overcoming the titration paradox,” *CPT: Pharmacometrics & Systems Pharmacology*, vol. 11, no. 12, pp. 1592–1603, 2022.
- [60] T. W. Schnider and C. F. Minto, “Beware the drug titration paradox. comment on br j anaesth 2021; 127: 704–12,” *British Journal of Anaesthesia*, vol. 128, no. 6, pp. e335–e337, 2022.
- [61] C. Aldecoa, G. Bettelli, F. Bilotta, R. D. Sanders, R. Audisio, A. Borozdina, A. Cherubini, C. Jones, H. Kehlet, A. MacLulich *et al.*, “European society of anaesthesiology evidence-based and consensus-based guideline on postoperative delirium,” *European Journal of Anaesthesiology—EJA*, vol. 34, no. 4, pp. 192–214, 2017.
- [62] M. Checketts, R. Alladi, K. Ferguson, L. Gemmell, J. Handy, A. Klein, N. Love, U. Misra, C. Morris, M. Nathanson *et al.*, “Recommendations for standards of monitoring during anaesthesia and recovery 2015: Association of anaesthetists of great britain and ireland,” *Anaesthesia*, vol. 71, no. 1, pp. 85–93, 2016.
- [63] F. Radtke, M. Franck, J. Lendner, S. Krüger, K. Wernecke, and C. Spies, “Monitoring depth of anaesthesia in a randomized trial decreases the rate of postoperative delirium but not postoperative cognitive dysfunction,” *British journal of anaesthesia*, vol. 110, no. suppl_1, pp. i98–i105, 2013.
- [64] Y. Punjasawadwong, W. Chau-in, M. Laopaiboon, S. Punjasawadwong, and P. Pin-on, “Processed electroencephalogram and evoked potential techniques for amelioration of postoperative delirium and cognitive dysfunction following non-cardiac and non-neurosurgical procedures in adults,” *Cochrane Database of Systematic Reviews*, no. 5, 2018.
- [65] T. J. Gan, P. S. Glass, A. Windsor, F. Payne, C. Rosow, P. Sebel, and P. Manberg, “Bispectral index monitoring allows faster emergence and improved recovery from propofol, alfentanil, and nitrous oxide anesthesia,” *Anesthesiology*, vol. 87, no. 4, pp. 808–815, 1997.

- [66] C. Bennett, L. J. Voss, J. P. Barnard, and J. W. Sleight, "Practical use of the raw electroencephalogram waveform during general anesthesia: the art and science," *Anesthesia & Analgesia*, vol. 109, no. 2, pp. 539–550, 2009.
- [67] S. Mathur, J. Patel, S. Goldstein, J. M. Hendrix, and A. Jain, "Bispectral index," in *StatPearls [Internet]*. StatPearls Publishing, 2023.
- [68] A. Aho, A. Yli-Hankala, L.-P. Lyytikäinen, and V. Jäntti, "Facial muscle activity, response entropy, and state entropy indices during noxious stimuli in propofol–nitrous oxide or propofol–nitrous oxide–remifentanyl anaesthesia without neuromuscular block," *British journal of anaesthesia*, vol. 102, no. 2, pp. 227–233, 2009.
- [69] T. Ledowski, "Objective monitoring of nociception: a review of current commercial solutions," *British journal of anaesthesia*, vol. 123, no. 2, pp. e312–e321, 2019.
- [70] S. Kreuer and W. Wilhelm, "The narcotrend monitor," *Best Practice & Research Clinical Anaesthesiology*, vol. 20, no. 1, pp. 111–119, 2006.
- [71] R. Sharma and H. K. Meena, "Emerging trends in eeg signal processing: A systematic review," *SN Computer Science*, vol. 5, no. 4, p. 415, 2024.
- [72] H. Zhang, Q.-Q. Zhou, H. Chen, X.-Q. Hu, W.-G. Li, Y. Bai, J.-X. Han, Y. Wang, Z.-H. Liang, D. Chen *et al.*, "The applied principles of eeg analysis methods in neuroscience and clinical neurology," *Military Medical Research*, vol. 10, no. 1, p. 67, 2023.
- [73] X. Jiang, G.-B. Bian, and Z. Tian, "Removal of artifacts from eeg signals: a review," *Sensors*, vol. 19, no. 5, p. 987, 2019.
- [74] M. I. Al-Kadi, M. B. I. Reaz, and M. A. M. Ali, "Evolution of electroencephalogram signal analysis techniques during anesthesia," *Sensors*, vol. 13, no. 5, pp. 6605–6635, 2013.
- [75] D. Safieddine, A. Kachenoura, L. Albera, G. Birot, A. Karfoul, A. Pasnicu, A. Biraben, F. Wendling, L. Senhadji, and I. Merlet, "Removal of muscle artifact from eeg data: comparison between stochastic (ica and cca) and deterministic (emd and wavelet-based) approaches," *EURASIP Journal on Advances in Signal Processing*, vol. 2012, no. 1, p. 127, 2012.
- [76] V. A. Raj, T. Parupudi, A. Thalengala, and S. G. Nayak, "A comprehensive review of deep learning models for denoising eeg signals: challenges, advances, and future directions," *Discover Applied Sciences*, vol. 7, no. 11, p. 1268, 2025.

- [77] H. Zhang, M. Zhao, C. Wei, D. Mantini, Z. Li, and Q. Liu, “Eegdenoisnet: a benchmark dataset for deep learning solutions of eeg denoising,” *Journal of Neural Engineering*, vol. 18, no. 5, p. 056057, 2021.
- [78] Y. Li, W. Zeng, W. Dong, D. Han, L. Chen, H. Chen, Z. Kang, S. Gong, H. Yan, W. T. Siok *et al.*, “A tale of single-channel electroencephalogram: Devices, datasets, signal processing, applications, and future directions,” *IEEE Transactions on Instrumentation and Measurement*, 2025.
- [79] A. Chaddad, Y. Wu, R. Kateb, and A. Bouridane, “Electroencephalography signal processing: A comprehensive review and analysis of methods and techniques,” *Sensors*, vol. 23, no. 14, p. 6434, 2023.
- [80] D. F. Hight, H. A. Kaiser, J. W. Sleight, and M. S. Avidan, “An updated introduction to electroencephalogram-based brain monitoring during intended general anesthesia,” *Canadian Journal of Anesthesia/Journal canadien d’anesthésie*, vol. 67, no. 12, pp. 1858–1878, 2020.
- [81] M. C. Kim, G. Fricchione, E. Brown, and O. Akeju, “Role of electroencephalogram oscillations and the spectrogram in monitoring anaesthesia,” *BJA education*, vol. 20, no. 5, pp. 166–172, 2020.
- [82] S. Hagihira, “Changes in the electroencephalogram during anaesthesia and their physiological basis,” *British journal of anaesthesia*, vol. 115, no. suppl.1, pp. i27–i31, 2015.
- [83] P. L. Purdon, E. T. Pierce, E. A. Mukamel, M. J. Prerau, J. L. Walsh, K. F. K. Wong, A. F. Salazar-Gomez, P. G. Harrell, A. L. Sampson, A. Cimenser *et al.*, “Electroencephalogram signatures of loss and recovery of consciousness from propofol,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 12, pp. E1142–E1151, 2013.
- [84] A. Sandre and S. Troller-Renfree, “Power struggles: Absolute vs. relative eeg power in developmental neuroscience,” *Relative EEG Power in Developmental Neuroscience*, 2025.
- [85] T. Donoghue, J. Dominguez, and B. Voytek, “Electrophysiological frequency band ratio measures conflate periodic and aperiodic neural activity,” *eneuro*, vol. 7, no. 6, 2020.
- [86] D. Hight, L. J. Voss, P. S. Garcia, and J. Sleight, “Changes in alpha frequency and power of the electroencephalogram during volatile-based general anesthesia,” *Frontiers in systems neuroscience*, vol. 11, p. 36, 2017.
- [87] Z. Ricci, L. Gobbi, E. L. Rosa, E. Filippini, M. Lui, M. Fischer, D. Colosimo, and S. Romagnoli, “Analysis of sef95 derived from two processed eeg devices during pediatric anesthesia for non-cardiac surgery,” *Journal of Clinical Monitoring and Computing*, pp. 1–6, 2025.

- [88] A. Shanker, J. H. Abel, G. Schamberg, and E. N. Brown, “Etiology of burst suppression eeg patterns,” *Frontiers in Psychology*, vol. 12, p. 673529, 2021.
- [89] F. Amzica, “What does burst suppression really mean?” *Epilepsy & Behavior*, vol. 49, pp. 234–237, 2015.
- [90] N. Pawar and O. L. Barreto Chang, “Burst suppression during general anesthesia and postoperative outcomes: mini review,” *Frontiers in systems neuroscience*, vol. 15, p. 767489, 2022.
- [91] J. Bruhn, “Burst suppression ratio is the only determinant for bis values below 30,” *Canadian Journal of Anesthesia*, vol. 49, no. 7, pp. 755–756, 2002.
- [92] R.-u. Alam, H. Zhao, A. Goodwin, O. Kavehei, and A. McEwan, “Differences in power spectral densities and phase quantities due to processing of eeg signals,” *Sensors*, vol. 20, no. 21, p. 6285, 2020.
- [93] L. Zakaria, A. Desowska, C. B. Berde, and L. Cornelissen, “Electroencephalographic delta and alpha oscillations reveal phase-amplitude coupling in paediatric patients undergoing sevoflurane-based general anaesthesia,” *British Journal of Anaesthesia*, vol. 130, no. 5, pp. 595–602, 2023.
- [94] C. Kalogeropoulos, K. Theofilatos, and S. Mavroudi, “From neurons to networks: A holistic review of electroencephalography (eeg) from neurophysiological foundations to ai techniques,” *Signals*, vol. 7, no. 1, p. 17, 2026.
- [95] T. Kleinert, T. Koenig, K. Nash, and E. Wascher, “On the reliability of the eeg microstate approach,” *Brain topography*, vol. 37, no. 2, pp. 271–286, 2024.
- [96] C. M. Michel and T. Koenig, “Eeg microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: a review,” *Neuroimage*, vol. 180, pp. 577–593, 2018.
- [97] L. Br chet and C. M. Michel, “Eeg microstates in altered states of consciousness,” *Frontiers in psychology*, vol. 13, p. 856697, 2022.
- [98] Y. Zhang, H. Wang, F. Yan, D. Song, Q. Wang, Y. Wang, and L. Huang, “Frequency-and state-dependent dynamics of eeg microstates during propofol anesthesia,” *NeuroImage*, vol. 310, p. 121159, 2025.
- [99] A. P. Lapointe, D. Li, A. G. Hudetz, and P. E. Vlisides, “Microstate analyses as an indicator of anesthesia-induced unconsciousness,” *Clinical Neurophysiology*, vol. 147, pp. 81–87, 2023.

- [100] X.-T. Tran, T.-N. Vo, S.-T. Vu, T.-T. Tran, M.-D. Nguyen, T. Do, and C.-T. Lin, “Inter-and intra-subject variability in eeg: A systematic survey,” *arXiv preprint arXiv:2602.01019*, 2026.
- [101] H. Hoshi and Y. Shigihara, “Age-and gender-specific characteristics of the resting-state brain activity: a magnetoencephalography study,” *Aging (Albany NY)*, vol. 12, no. 21, p. 21613, 2020.
- [102] A. Chiang, C. Rennie, P. Robinson, S. Van Albada, and C. Kerr, “Age trends and sex differences of alpha rhythms including split alpha peaks,” *Clinical Neurophysiology*, vol. 122, no. 8, pp. 1505–1517, 2011.
- [103] C. Han, V. C. Cheung, and R. H. Chan, “Aging amplifies sex differences in low alpha and low beta eeg oscillations,” *Neuroimage*, vol. 312, p. 121231, 2025.
- [104] C. Wusthoff, R. Shellhaas, and R. Clancy, “Limitations of single-channel eeg on the forehead for neonatal seizure detection,” *Journal of Perinatology*, vol. 29, no. 3, pp. 237–242, 2009.
- [105] A. Craik, Y. He, and J. L. Contreras-Vidal, “Deep learning for electroencephalogram (eeg) classification tasks: a review,” *Journal of neural engineering*, vol. 16, no. 3, p. 031001, 2019.
- [106] E. Vafaei and M. Hosseini, “Transformers in eeg analysis: a review of architectures and applications in motor imagery, seizure, and emotion classification,” *Sensors*, vol. 25, no. 5, p. 1293, 2025.
- [107] J. Lai, J. Wei, L. Yao, and Y. Wang, “A simple review of eeg foundation models: Datasets, advancements and future perspectives,” *arXiv preprint arXiv:2504.20069*, 2025.
- [108] J. Sukhbaatar, S. Imamura, I. Inoue, S. Murakami, K. M. Hassan, S. Han, I. Chanpornpakdi, and T. Tanaka, “Singlem: Single-channel large eeg model,” *arXiv preprint arXiv:2509.17920*, 2025.
- [109] P. Zis, A. Liampas, A. Artemiadis, G. Tsalamandris, P. Neophytou, Z. Unwin, V. K. Kimiskidis, G. M. Hadjigeorgiou, G. Varrassi, Y. Zhao *et al.*, “Eeg recordings as biomarkers of pain perception: where do we stand and where to go?” *Pain and therapy*, vol. 11, no. 2, pp. 369–380, 2022.
- [110] F. Chouchou, C. Perchet, and L. Garcia-Larrea, “Eeg changes reflecting pain: is alpha suppression better than gamma enhancement?” *Neurophysiologie Clinique*, vol. 51, no. 3, pp. 209–218, 2021.
- [111] E. MacKay, J. Sleight, L. Voss, and J. Barnard, “Episodic waveforms in the electroencephalogram during general anaesthesia: a study of patterns

- of response to noxious stimuli,” *Anaesthesia and Intensive Care*, vol. 38, no. 1, pp. 102–112, 2010.
- [112] C.-N. Ho, N. E. Huang, J.-Y. Chen, and A. C. Yang, “Exploring nociceptive–analgesic balance and eeg modulation patterns during general anesthesia using holo-hilbert spectral analysis,” *Pain Research and Management*, vol. 2026, no. 1, p. 5504074, 2026.
 - [113] J. Reiser, M. Kreuzer, J. Werner, A. M. Saller, J. Fischer, S. Senf, P. Deffner, N. Abendschön, T. Groll, A. Grott *et al.*, “Nociception-induced changes in electroencephalographic activity and fos protein expression in piglets undergoing castration under isoflurane anaesthesia,” *Animals*, vol. 12, no. 18, p. 2309, 2022.
 - [114] S. Qiu, X. Lyu, Q. Zheng, H. He, R. Jin, and W. Peng, “Temporal dynamics of electroencephalographic microstates during sustained pain,” *Cerebral Cortex*, vol. 33, no. 13, pp. 8594–8604, 2023.
 - [115] O. Akeju, M. B. Westover, K. J. Pavone, A. L. Sampson, K. E. Hartnack, E. N. Brown, and P. L. Purdon, “Effects of sevoflurane and propofol on frontal electroencephalogram power and coherence,” *Anesthesiology*, vol. 121, no. 5, p. 990, 2014.
 - [116] P. S. Sebel, J. G. Bovill, A. Wauquier, and P. Rog, “Effects of high-dose fentanyl anesthesia on the electroencephalogram,” *Anesthesiology*, vol. 55, no. 3, pp. 203–211, 1981.
 - [117] G. A. Balanza, K. M. Bharadwaj, A. C. Mullen, A. M. Beck, E. C. Work, F. J. McGovern, T. T. Houle, T. P. Eric, and P. L. Purdon, “An electroencephalogram biomarker of fentanyl drug effects,” *PNAS nexus*, vol. 1, no. 4, p. pgac158, 2022.
 - [118] J. K. Danel, J. Rosada-Kurasinska, M. M. Copik, S. Zdanowski, W. Gola, H. Misiolek, A. Bartkowska-Śniatkowska, and S. Białka, “Objective monitoring of acute pain and nociception in anaesthesia and intensive care: evidence and applications,” *Anaesthesiology Intensive Therapy*, vol. 57, p. e365, 2025.
 - [119] M. K. Kim, G. J. Choi, K. S. Oh, S. P. Lee, and H. Kang, “Pain assessment using the analgesia nociception index (ani) in patients undergoing general anesthesia: a systematic review and meta-analysis,” *Journal of personalized medicine*, vol. 13, no. 10, p. 1461, 2023.
 - [120] D. Wildemeersch, N. Peeters, V. Saldien, M. Vercauteren, and G. Hans, “Pain assessment by pupil dilation reflex in response to noxious stimulation in anaesthetized adults,” *Acta Anaesthesiologica Scandinavica*, vol. 62, no. 8, pp. 1050–1056, 2018.

- [121] A. Jakuscheit, M. J. Posch, S. Gkaitatzis, L. Neumark, M. Hackbarth, M. Schneider, G. Lichtner, J. H. Baars, and F. von Dincklage, "Utility of nociceptive flexion reflex threshold and bispectral index to predict movement responses under propofol anaesthesia," *Somatosensory & Motor Research*, vol. 34, no. 2, pp. 139–144, 2017.
- [122] F. Meijer, M. Honing, T. Roor, S. Toet, P. Calis, E. Olofsen, C. Martini, M. van Velzen, L. Aarts, M. Niesters *et al.*, "Reduced postoperative pain using nociception level-guided fentanyl dosing during sevoflurane anaesthesia: a randomised controlled trial," *British Journal of Anaesthesia*, vol. 125, no. 6, pp. 1070–1078, 2020.
- [123] A. Cividjian, J. Martinez, E. Combourieu, P. Precloux, A. Beraud, Y. Rochette, M. Cler, L. Bourdon, J. Escarment, and L. Quintin, "Beat-by-beat cardiovascular index to predict unexpected intraoperative movement in anesthetized unparalyzed patients: a retrospective analysis," *Journal of clinical monitoring and computing*, vol. 21, no. 2, pp. 91–101, 2007.
- [124] F. Von Dincklage, K. Send, M. Hackbarth, B. Rehberg, and J. Baars, "Comparison of the nociceptive flexion reflex threshold and the bispectral index as monitors of movement responses to noxious stimuli under propofol mono-anaesthesia," *British journal of anaesthesia*, vol. 102, no. 2, pp. 244–250, 2009.
- [125] N. Yang, Y. Yue, J. Z Pan, M.-Z. Zuo, Y. Shi, S.-Z. Zhou, W.-P. Peng, and J.-D. Gao, "Changes in the bispectral index in response to loss of consciousness and no somatic movement to nociceptive stimuli in elderly patients," *Chinese Medical Journal*, vol. 129, no. 04, pp. 410–416, 2016.
- [126] M. Shoushtarian, M. M. Sahinovic, A. R. Absalom, A. F. Kalmar, H. E. Vereecke, D. T. Liley, and M. M. Struys, "Comparisons of electroencephalographically derived measures of hypnosis and antinociception in response to standardized stimuli during target-controlled propofol-remifentanyl anesthesia," *Anesthesia & Analgesia*, vol. 122, no. 2, pp. 382–392, 2016.
- [127] D. M. Mathews, L. Clark, J. Johansen, E. Matute, and C. V. Seshagiri, "Increases in electroencephalogram and electromyogram variability are associated with an increased incidence of intraoperative somatic response," *Anesthesia & Analgesia*, vol. 114, no. 4, pp. 759–770, 2012.
- [128] M. M. Sahinovic, D. J. Eleveld, A. F. Kalmar, E. H. Heeremans, T. De Smet, C. V. Seshagiri, A. R. Absalom, H. E. Vereecke, and M. M. Struys, "Accuracy of the composite variability index as a measure of the balance between nociception and antinociception during anesthesia," *Anesthesia & Analgesia*, vol. 119, no. 2, pp. 288–301, 2014.

- [129] I. Takamatsu, M. Ozaki, and T. Kazama, “Entropy indices vs the bispectral indexTM for estimating nociception during sevoflurane anaesthesia,” *BJA: British Journal of Anaesthesia*, vol. 96, no. 5, pp. 620–626, 2006.
- [130] M. Struys, C. Vanpeteghem, M. Huiku, K. Uutela, N. Blyaert, and E. Mortier, “Changes in a surgical stress index in response to standardized pain stimuli during propofol–remifentanyl infusion,” *British journal of anaesthesia*, vol. 99, no. 3, pp. 359–367, 2007.
- [131] E. Dierckens, M. Fleyfel, E. Robin, A. Legrand, M. Borel, L. Gambier, B. Vallet, and G. Lebuffe, “Is entropy a monitor for the guidance of intraoperative analgesia?” in *Annales Francaises D’anesthesie et de Re-animation*, vol. 26, no. 2, 2006, pp. 113–118.
- [132] M. Rantanen, A. Yli-Hankala, M. van Gils, H. Yppärilä-Wolters, P. Takala, M. Huiku, M. Kymäläinen, E. Seitsonen, and I. Korhonen, “Novel multiparameter approach for measurement of nociception at skin incision during general anaesthesia,” *British journal of anaesthesia*, vol. 96, no. 3, pp. 367–376, 2006.
- [133] T. J. Sarén-Koivuniemi, A. M. Yli-Hankala, and M. J. van Gils, “Increased variation of the response index of nociception during noxious stimulation in patients during general anaesthesia,” *Computer methods and programs in biomedicine*, vol. 104, no. 2, pp. 154–160, 2011.
- [134] U. Melia, M. Vallverdú, M. Jospin, E. W. Jensen, J. F. Valencia, F. Clariá, P. L. Gambús, and P. Caminal, “Prediction of nociceptive responses during sedation by time-frequency representation,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2013, pp. 2547–2550.
- [135] U. Melia, M. Vallverdu, X. Borrat, J. F. Valencia, M. Jospin, E. W. Jensen, P. Gambús, and P. Caminal, “Prediction of nociceptive responses during sedation by linear and non-linear measures of eeg signals in high frequencies,” *PLoS One*, vol. 10, no. 4, p. e0123464, 2015.
- [136] J. F. Valencia, U. S. Melia, M. Vallverdú, X. Borrat, M. Jospin, E. W. Jensen, A. Porta, P. L. Gambús, and P. Caminal, “Assessment of nociceptive responsiveness levels during sedation-analgesia by entropy analysis of eeg,” *Entropy*, vol. 18, no. 3, p. 103, 2016.
- [137] M. Shoushtarian, D. P. McGlade, L. J. Delacretaz, and D. T. Liley, “Evaluation of the brain anaesthesia response monitor during anaesthesia for cardiac surgery: a double-blind, randomised controlled trial using two doses of fentanyl,” *Journal of clinical monitoring and computing*, vol. 30, no. 6, pp. 833–844, 2016.

- [138] E. W. Jensen, J. Valencia, A. López, T. Anglada, M. Agustí, Y. Ramos, R. Serra, M. Jospin, P. Pineda, and P. Gambus, “Monitoring hypnotic effect and nociception with two eeg-derived indices, qcon and qnox, during general anaesthesia,” *Acta Anaesthesiologica Scandinavica*, vol. 58, no. 8, pp. 933–941, 2014.
- [139] U. Melia, E. Gabarron, M. Agustí, N. Souto, P. Pineda, J. Fontanet, M. Vallverdu, E. W. Jensen, and P. Gambus, “Comparison of the qcon and qnox indices for the assessment of unconsciousness level and noxious stimulation response during surgery,” *Journal of clinical monitoring and computing*, vol. 31, no. 6, pp. 1273–1281, 2017.
- [140] H. Kong, D.-D. Ma, J.-H. Ma, Y.-X. Zhang, H. Zhang, and D.-X. Wang, “Qnox index for quantification of intraoperative nociception and analgesia: a prospective single-centre validation study,” *British journal of anaesthesia*, vol. 134, no. 3, pp. 736–745, 2025.
- [141] L. Wu, S. Wang, Y. Wang, K. Zhang, J. Bai, and J. Zheng, “Prediction of hemodynamic reactivity by electroencephalographically derived pain threshold index in children undergoing general anesthesia: a prospective observational study,” *Journal of pain research*, pp. 3245–3255, 2019.
- [142] R. Wang, Y. Deng, S. Zhou, and J. Zhang, “Eeg-derived pain threshold index for prediction of postoperative pain in patients undergoing laparoscopic urological surgery: a comparison with surgical pleth index,” *Journal of Clinical Monitoring and Computing*, vol. 35, no. 6, pp. 1395–1402, 2021.
- [143] Y. Jiang, J.-m. Ding, X.-x. Hao, P.-p. Fang, and X.-S. Liu, “Eeg-derived pain threshold index-guided versus standard care during propofol-remifentanil anesthesia: A randomized controlled trial,” *Heliyon*, vol. 9, no. 8, 2023.
- [144] C. Lysakowski, L. Dumont, M. Pellegrini, F. Clergue, and E. Tassonyi, “Effects of fentanyl, alfentanil, remifentanil and sufentanil on loss of consciousness and bispectral index during propofol induction of anaesthesia,” *British journal of anaesthesia*, vol. 86, no. 4, pp. 523–527, 2001.
- [145] I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw, “Emg contamination of eeg: spectral and topographical characteristics,” *Clinical neurophysiology*, vol. 114, no. 9, pp. 1580–1593, 2003.
- [146] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and P. Consortium, “Explainability for artificial intelligence in healthcare: a multidisciplinary perspective,” *BMC medical informatics and decision making*, vol. 20, no. 1, p. 310, 2020.

- [147] C. Law, G. Jacobson, M. Kluger, M. Chaddock, M. Scott, and J. Sleigh, “Randomized controlled trial of the effect of depth of anaesthesia on postoperative pain,” *British journal of anaesthesia*, vol. 112, no. 4, pp. 675–680, 2014.
- [148] D. F. Hight, A. L. Gaskell, M. Kreuzer, L. J. Voss, P. S. García, and J. W. Sleigh, “Transient electroencephalographic alpha power loss during maintenance of general anaesthesia,” *British journal of anaesthesia*, vol. 122, no. 5, pp. 635–642, 2019.
- [149] A. L. Gaskell, A. Banerji, R. M. Pullon, J. Termaat, P. S. García, and J. W. Sleigh, “Eeg alpha power and delirium in the postanesthesia care unit in older adults: the alphamax trial part 1-effect of desflurane and fentanyl titration during maintenance,” *British Journal of Anaesthesia*, 2025.
- [150] S. L. Shafer, J. R. Varvel, N. Aziz, and J. C. Scott, “Pharmacokinetics of fentanyl administered by computer-controlled infusion pump,” *Anesthesiology*, vol. 73, no. 6, pp. 1091–1102, 1990.
- [151] S. L. Shafer and J. R. Varvel, “Pharmacokinetics, pharmacodynamics, and rational opioid selection,” *Anesthesiology*, vol. 74, no. 1, pp. 53–63, 1991.
- [152] P. Wiczling, A. Bienert, P. Sobczyński, R. Hartmann-Sobczyńska, K. Bieda, A. Marcinkowska, M. Malatyńska, R. Kaliszan, and E. Grześkowiak, “Pharmacokinetics and pharmacodynamics of propofol in patients undergoing abdominal aortic surgery,” *Pharmacological Reports*, vol. 64, no. 1, pp. 113–122, 2012.
- [153] J. X. Mazoit, K. Butscher, and K. Samii, “Morphine in postoperative patients: pharmacokinetics and pharmacodynamics of metabolites,” *Anesthesia & Analgesia*, vol. 105, no. 1, pp. 70–78, 2007.
- [154] I. D. McKay, L. J. Voss, J. W. Sleigh, J. P. Barnard, and E. K. Johannsen, “Pharmacokinetic-pharmacodynamic modeling the hypnotic effect of sevoflurane using the spectral entropy of the electroencephalogram,” *Anesthesia & Analgesia*, vol. 102, no. 1, pp. 91–97, 2006.
- [155] X. Pu, P. Yi, K. Chen, Z. Ma, D. Zhao, and Y. Ren, “Eegdnet: Fusing non-local and local self-similarity for eeg signal denoising with transformer,” *Computers in Biology and Medicine*, vol. 151, p. 106248, 2022.
- [156] C. W. Connor, “Open reimplementations of the bis algorithms for depth of anesthesia,” *Anesthesia & Analgesia*, vol. 135, no. 4, pp. 855–864, 2022.

- [157] G. R. Palos, T. R. Mendoza, G. M. Mobley, S. B. Cantor, and C. S. Cleeland, “Asking the community about cutpoints used to describe mild, moderate, and severe pain,” *The Journal of Pain*, vol. 7, no. 1, pp. 49–56, 2006.
- [158] C. Y. Lee, “Nested logistic regression models and δ auc applications: Change-point analysis,” *Statistical methods in medical research*, vol. 30, no. 7, pp. 1654–1666, 2021.
- [159] S.-H. Moon, “Sedation regimens for gastrointestinal endoscopy,” *Clinical endoscopy*, vol. 47, no. 2, p. 135, 2014.
- [160] K. Sriganesh, K. A. Theerth, M. Reddy, D. Chakrabarti, and G. S. U. Rao, “Analgesia nociception index and systemic haemodynamics during anaesthetic induction and tracheal intubation: A secondary analysis of a randomised controlled trial,” *Indian Journal of Anaesthesia*, vol. 63, no. 2, pp. 100–105, 2019.
- [161] K. Karako and W. Tang, “Applications of and issues with machine learning in medicine: Bridging the gap with explainable ai,” *Bioscience trends*, vol. 18, no. 6, pp. 497–504, 2024.
- [162] N. Freyer, D. Groß, and M. Lipprandt, “The ethical requirement of explainability for ai-dss in healthcare: a systematic review of reasons,” *BMC Medical Ethics*, vol. 25, no. 1, p. 104, 2024.
- [163] N. Burkart and M. F. Huber, “A survey on the explainability of supervised machine learning,” *Journal of Artificial Intelligence Research*, vol. 70, pp. 245–317, 2021.
- [164] W. Cao, V. Mirjalili, and S. Raschka, “Rank consistent ordinal regression for neural networks with application to age estimation,” *Pattern Recognition Letters*, vol. 140, pp. 325–331, 2020.
- [165] A. F. Rogobete, O. H. Bedreag, M. Papurica, S. E. Popovici, L. M. Bratu, A. Rata, C. R. Barsac, A. Maghiar, D. N. Garofil, M. Negrea *et al.*, “Multiparametric monitoring of hypnosis and nociception-antinociception balance during general anesthesia—a new era in patient safety standards and healthcare management,” *Medicina*, vol. 57, no. 2, p. 132, 2021.
- [166] M. S. Fabuš, *Spatiotemporal brain dynamics induced by propofol and ketamine in humans*. University of Oxford (United Kingdom), 2023.
- [167] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

- [168] M. W. Watkins, “Exploratory factor analysis: A guide to best practice,” *Journal of black psychology*, vol. 44, no. 3, pp. 219–246, 2018.
- [169] S. Aviyente, “Time-frequency decomposition methods for event-related potential analysis,” *The Oxford handbook of EEG frequency*, pp. 65–87, 2022.
- [170] J. W. Osborne and A. B. Costello, “Sample size and subject to item ratio in principal components analysis,” *Practical Assessment, Research, and Evaluation*, vol. 9, no. 1, 2004.
- [171] L. R. Rabiner, “A tutorial on hidden markov models and selected applications in speech recognition,” *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.
- [172] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, “A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains,” *The annals of mathematical statistics*, vol. 41, no. 1, pp. 164–171, 1970.
- [173] D. Vidaurre, L. Masaracchia, N. Y. Larsen, L. R. Ruijters, S. Alonso, C. Ahrends, and M. W. Woolrich, “The gaussian-linear hidden markov model: A python package,” *Imaging Neuroscience*, vol. 3, p. imag_a.00460, 2025.
- [174] E. J. Kuip, M. L. Zandvliet, S. L. Koolen, R. H. Mathijssen, and C. C. van der Rijt, “A review of factors explaining variability in fentanyl pharmacokinetics; focus on implications for cancer patients,” *British journal of clinical pharmacology*, vol. 83, no. 2, pp. 294–313, 2017.
- [175] S. B. Duffull, D. F. Wright, and H. R. Winter, “Interpreting population pharmacokinetic-pharmacodynamic analyses—a clinical viewpoint,” *British journal of clinical pharmacology*, vol. 71, no. 6, pp. 807–814, 2011.
- [176] L. Keutzer, H. You, A. Farnoud, J. Nyberg, S. G. Wicha, G. Maher-Edwards, G. Vlasakakis, G. K. Moghaddam, E. M. Svensson, M. P. Menden *et al.*, “Machine learning and pharmacometrics for prediction of pharmacokinetic data: differences, similarities and challenges illustrated with rifampicin,” *Pharmaceutics*, vol. 14, no. 8, p. 1530, 2022.
- [177] R. J. Bauer, “Nonmem tutorial part ii: estimation methods and advanced examples,” *CPT: pharmacometrics & systems pharmacology*, vol. 8, no. 8, pp. 538–556, 2019.
- [178] J. Bae, M. Kwon, Y.-H. Lee, E.-K. Lee, B.-M. Choi, and G.-J. Noh, “An allometric pharmacokinetic model and minimum effective analgesic concentration of fentanyl in patients undergoing major abdominal surgery,” *British Journal of Anaesthesia*, vol. 125, no. 6, pp. 976–985, 2020.

- [179] F. Bahrami, R. M. Rossi, and T. Defraeye, “Predicting transdermal fentanyl delivery using physics-based simulations for tailored therapy based on the age,” *Drug Delivery*, vol. 29, no. 1, pp. 950–969, 2022.
- [180] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-cam: visual explanations from deep networks via gradient-based localization,” *International journal of computer vision*, vol. 128, no. 2, pp. 336–359, 2020.
- [181] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” *Advances in neural information processing systems*, vol. 27, 2014.
- [182] Y. Ganin and V. Lempitsky, “Unsupervised domain adaptation by back-propagation,” in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.
- [183] O. Özdenizci, Y. Wang, T. Koike-Akino, and D. Erdoğmuş, “Learning invariant representations from eeg via adversarial inference,” *IEEE access*, vol. 8, pp. 27 074–27 085, 2020.
- [184] P. M. Schumacher, J. Dossche, E. P. Mortier, M. Luginbuehl, T. W. Bouillon, and M. M. Struys, “Response surface modeling of the interaction between propofol and sevoflurane,” *Anesthesiology*, vol. 111, no. 4, pp. 790–804, 2009.
- [185] M. M. Yang, R. L. Hartley, A. A. Leung, P. E. Ronksley, N. Jetté, S. Casha, and J. Riva-Cambrin, “Preoperative predictors of poor acute postoperative pain control: a systematic review and meta-analysis,” *BMJ open*, vol. 9, no. 4, p. e025091, 2019.
- [186] H. Y. V. Ip, A. Abrishami, P. W. Peng, J. Wong, and F. Chung, “Predictors of postoperative pain and analgesic consumption: a qualitative systematic review,” *Anesthesiology*, vol. 111, no. 3, pp. 657–677, 2009.
- [187] M. A. Pietrancosta and V. Martinez, “A systematic review and meta-analysis of postoperative opioid consumption, pain and complications following intraoperative autonomic nervous system-based nociception monitoring,” *European Journal of Pain*, vol. 30, no. 2, p. e70227, 2026.
- [188] K. C. Koschmieder, S. Funcke, M. Shadloo, H. O. Pinnschmidt, G. Greiwe, M. Fischer, and R. Nitzschke, “Validation of three nociception indices to predict immediate postoperative pain before emergence from general anaesthesia: a prospective double-blind, observational study,” *British Journal of Anaesthesia*, vol. 130, no. 4, pp. 477–484, 2023.
- [189] K.-C. Hung, Y.-T. Huang, J.-R. Kuo, C.-W. Hsu, M. Yew, J.-Y. Chen, M.-C. Lin, I.-W. Chen, and C.-K. Sun, “Elevated surgical pleth index at

- the end of surgery is associated with postoperative moderate-to-severe pain: a systematic review and meta-analysis,” *Diagnostics*, vol. 12, no. 9, p. 2167, 2022.
- [190] M. Messner, U. Beese, J. Romstöck, M. Dinkel, and K. Tschaikowsky, “The bispectral index declines during neuromuscular block in fully awake persons,” *Anesthesia & Analgesia*, vol. 97, no. 2, pp. 488–491, 2003.
 - [191] F. H. van Loon, L. A. Puijn, W. H. van Aarle, A. T. Dierick-van Daele, and A. R. Bouwman, “Pain upon inserting a peripheral intravenous catheter: size does not matter,” *The journal of vascular access*, vol. 19, no. 3, pp. 258–265, 2018.
 - [192] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, “Classifier calibration: a survey on how to assess and improve predicted class probabilities,” *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, 2023.
 - [193] M. B. Westover, S. Ching, V. M. Kumaraswamy, O. Akeju, E. Pierce, S. S. Cash, R. Kilbride, E. N. Brown, and P. L. Purdon, “The human burst suppression electroencephalogram of deep hypothermia,” *Clinical Neurophysiology*, vol. 126, no. 10, pp. 1901–1914, 2015.
 - [194] E. Manquat, H. Ravaux, M. Kindermans, J. Joachim, J. Serrano, C. Touchard, J. Mateo, A. Mebazaa, E. Gayat, F. Vallée *et al.*, “Impact of impaired cerebral blood flow autoregulation on electroencephalogram signals in adults undergoing propofol anaesthesia: a pilot study,” *BJA open*, vol. 1, p. 100004, 2022.
 - [195] A. Kirsch *et al.*, *An introduction to the mathematical theory of inverse problems*. Springer, 2011, vol. 120.