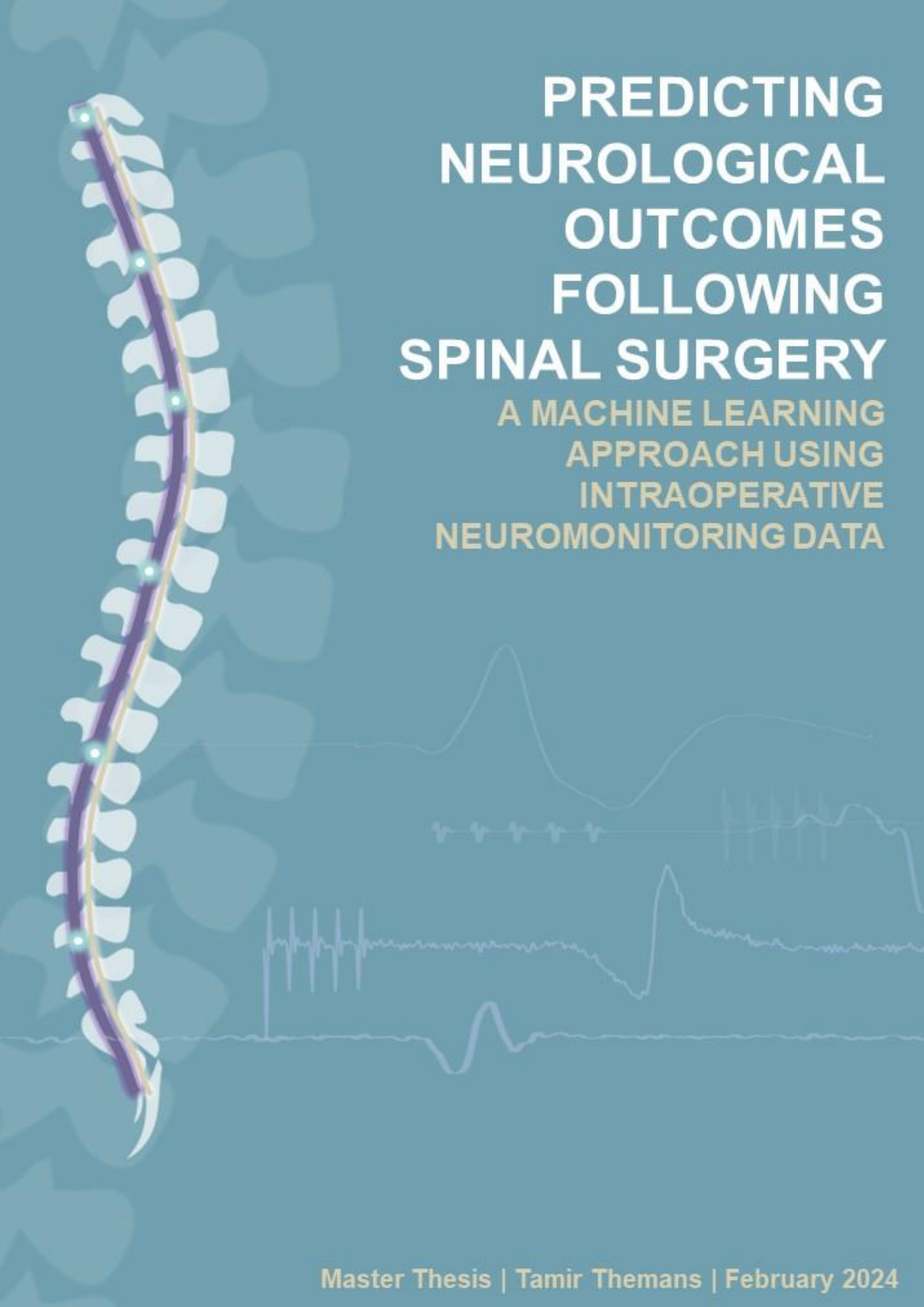

PREDICTING NEUROLOGICAL OUTCOMES FOLLOWING SPINAL SURGERY

A MACHINE LEARNING APPROACH USING INTRAOPERATIVE NEUROMONITORING DATA

Master Thesis | Tamir Themans | February 2024

This page was left blank intentionally

PREDICTING NEUROLOGICAL OUTCOMES FOLLOWING SPINAL SURGERY: A MACHINE LEARING APPROACH USING INTRAOPERATIVE NEUROMONITORING DATA

T.S. (Tamir) Themans

Student number: 4471687

February 16, 2024

Thesis in partial fulfilment of the requirements for the joint degree of Master of Science in

Technical Medicine

Leiden University; Delft University of Technology; Erasmus University Rotterdam

Master thesis project (TM30004; 35 ECTS)

Dept. of Biomechanical Engineering, TUDELFT

July 2023 – February 2024

Supervisor(s):

Dr. Niels van der Gaag

Dr. Ir. Mark van de Ruit

Dr. Valerie Ter Wengel

Thesis committee members:

Dr. Niels van der Gaag, Haga Teaching Hospital (Chair)

Dr. Ir. Mark van de Ruit, TU Delft

Dr. Valerie Ter Wengel, Haaglanden Medical Center

Dr. Pieter Kruizinga, Erasmus Medical Center

Universiteit Leiden

TU Delft Delft University of Technology

Erasmus Universiteit Rotterdam

Table of contents

Preface

This graduation report marks the culmination of my 8.5-year journey of study, which I have thoroughly enjoyed. Since the creation of the study Clinical Technology in 2014, I was certain that it was the study from which I would eventually graduate. Despite not being selected in the draws of 2015 and 2016, I remained hopeful. After attempting to secure a pre-master's from the Bachelor's Mechanical Engineering multiple times, I finally received approval on my fourth attempt. Although warned about the challenges, I can proudly say that I became the first and only Clinical Technologist to graduate from the TU Delft with a Bachelor's degree in Mechanical Engineering.

Throughout my years in college, I had difficulties with programming. However, I couldn't call myself a Clinical Technologist before I began to master this skill. Therefore, I wanted to challenge myself immensely, step out of my comfort zone, and chose to set up a machine learning project as graduation research. As a result, I now feel like I have developed a valuable skill that can be applied in endless possibilities within, as well as outside the hospital.

I would like to express my gratitude to everyone who contributed to the successful completion of this thesis project. Firstly, I want to acknowledge Mark, who, despite not specializing in machine learning, served as an invaluable mentor throughout the entire graduation process. Your ability to approach problems from a different perspective encouraged me to think outside the box, and I am genuinely appreciative of your support. Niels, despite your busy schedule, you gave me certain insights with your medical experience in the brief moments we had together, that highlighted the overarching importance of specific issues, significantly aiding me throughout the graduation process. I am also very grateful for Valerie's substantial help as additional medical support. Whenever I asked her a specific question, I could expect an enthusiastic phone call the same day which often resulted in a detailed conversation that in turn enthused me enormously. I would also like to thank Justin Dauwels for the machine learning expertise I gained from you during this project. Even though you were not officially one of my supervisors, you volunteered to help me, for which I am very grateful. Lastly, Saskia van der Gaag guided me through all the inquiries related to intraoperative neuromonitoring, both in and out of the operation room, and I appreciate that greatly.

Furthermore, I want to thank my mother, sister, fellow students, friends and roommates for your support during graduation. Last but not least, I want to thank myself for the persistence shown over the past 8.5 years. Now the future lies open like a fresh pack of snow in which a path is to be created. Let's see what the future holds for me.

Tamir Themans
Rotterdam, February 2024

List of abbreviations

ADM	Abductor digiti minimi
AH	Abductor hallucis
AP	Average precision
AUC	Area under the curve
CV	Cross validation
C-x	Cervical-x
D-waves	Direct-waves
EHR	Electronic health record
FN	False negative rate
FP	False positive rate
IONM	Intraoperative neuromonitoring
KNN	K-nearest neighbors
L-x	Lumbar-x
MAP	Mean arterial pressure
MEP	Motor evoked potential
MDR	Medical device regulation
ML	Machine learning
NaN	Not a number
RF	Random forest
ROC	Receiver operating characteristic
SSEP	Somatosensory evoked potential
SVM	Support vector machine
S-x	Sacral-x
TA	Tibialis anterior
TN	True negative rate
TP	True positive rate
T-x	Thoracal-x
XGBoost	Extreme gradient boosting

1. Abstract

Background context: Intraoperative neuromonitoring (IONM) has proven effective in reducing postoperative neurological complications. However, current understanding of IONM is limited and its precise meaning in relation to neurological outcomes remains unclear. Machine learning (ML) is a promising solution to analyze the excessive amount of IONM data quickly, objectively and in real-time.
Purpose: The goal is to develop a ML algorithm that can effectively predict neurological outcomes after spinal surgery using IONM data that include both motor evoked potentials (MEPs) and somatosensory evoked potentials (SSEPs), and analyze its key predicting features. To more effectively determine the specific independent contribution of both separate modalities, a separate ML model will be created for both MEP and SSEP in addition to a combined MEP-SSEP model.
Study setting: Retrospective study.
Patient sample: A total of 67 patients were analyzed.
Outcome measures: The neurological status three months postoperatively compared to the preoperative status, categorized into three classes: 'Neurological stable deficits', ‘Neurologically intact’ and 'Neurological improvement'.
Methods: 260 features were obtained from patients who underwent spinal surgery monitored by IONM. During nested cross-validation, the data was split into five folds, for both the inner and the outer loop. The four ML classifiers developed were support vector machine, K-nearest neighbors, random forest and extreme gradient boosting, and tested along the three modalities MEP, SSEP, and MEP-SSEP combination.
Results: Extreme gradient boosting outperformed the other classifiers on all performance metrics. The combined MEP-SSEP model exhibited the highest scores for sensitivity: 70.4%, specificity: 88.3% and accuracy: 87.1%, while the MEP model exhibited the highest performance for precision: 75.6%. Highest predicting scores per individual class were also obtained by this XGBoost classifier on the combined MEP-SSEP model. Key predicting features were the presence or absence of preoperative neurological deficits and last measured signal latency compared to baseline, with a contribution of 29% and 13.5% in the best performing model, respectively.
Conclusion: A reliable prediction of neurological outcomes three months postoperatively can be made combining MEP and SSEP IONM features, provided that the patient's preoperative status is accurately documented and included in the prediction. Though either MEP or SSEP features alone offer predictive value, MEP features show superior predictive values compared to SSEP features when both modalities are accessible, with latency emerging as a prominent predictive IONM feature.

2. Introduction

2.1 Rationale

2.1.1 Intraoperative neuromonitoring

During spinal procedures, surgeons operate in proximity to critical neural structures. This causes a risk of iatrogenic damage, leading to postoperative neurological impairment. [1] To reduce this risk of iatrogenic damage and thereby new neurological damage, intraoperative neuromonitoring (IONM) was introduced in 1975 by Tamaki and Yamane. [2] IONM includes several modalities including somatosensory evoked potentials (SSEPs) and motor evoked potentials (MEPs). When feasible, direct waves (D-waves) are also monitored as an additional modality. These modalities serve to assess neural pathway integrity and detect real-time changes during the spinal procedures, [3-7] allowing for a procedural stop or temporary pause. [1] The use of IONM has been linked to a decreased occurrence of postoperative neurological complications. [4, 8-10]

Nevertheless, despite its benefits, there are certain limitations associated with interpreting IONM. This technique requires specialized equipment, trained personnel, and additional time and resources during surgical procedures, leading to increased costs. Furthermore, the successful application of IONM relies on the expertise of a trained specialist. Different clinicians may interpret signals subjectively, leading to inter-rater variability in predicting neurological outcomes. [7, 11] Because IONM requires rapid and reliable interaction with the surgeon intraoperatively, this subjectivity is considered undesirable. Moreover, in current practice where IONM is used, no clear agreements on warning criteria indicating the risk of reversible or irreversible neurological deficit are described. Current guidelines suggest that for SSEPs, a 50% reduction in amplitude and a latency time increase of over 10% serve as standard warning criteria [6, 9, 12-16], while MEP warning criteria range between 50% and 100% in amplitude reduction as long as D-waves are preserved by over 50%. [1, 6, 9, 17] (App A) Apart from subjectivity and ambiguous alert criteria, considerations such as the general condition of the patient, suboptimal placement of needle electrodes and anaesthesia also contribute to the complex interpretation of IONM signals. [18]

During spinal surgery, the relation between the interpretation of IONM signals and the patient's prognosis becomes a critical focus for surgeons. However, this relation often remains unclear due to the observed variability: While a specific signal collapse may result in postoperative neurological deficits in one patient, others with similar signal collapse may not experience neurological problems. Enhanced and objective criteria in this area could significantly assist surgeons in making intraoperative decisions. Hence, the ability to real-time predict prognosis would be advantageous. Machine learning (ML) techniques might offer valuable contributions towards achieving this goal.

2.1.2 Machine learning

The increasing utilization of ML in healthcare, as a subset of artificial intelligence (AI), is on the rise due to its capacity to handle extensive datasets and convert analysis into valuable clinical insights, ultimately resulting in improved outcomes, reduced costs, and enhanced patient satisfaction. [19, 20] ML has been extensively researched in the context of neurosurgical conditions, presurgical planning, intraoperative guidance, neurophysiological monitoring, and neurosurgical outcome prediction. [21] Using ML techniques, the excessive amount of IONM data can rapidly be analyzed and interpreted intraoperatively to make predictions about the occurrence of neurological deficits and its key predicting features. ML models are primarily designed to provide precise and consistent predictions for new data by leveraging patterns acquired from existing data. [22] Predicting postoperative outcomes in spinal patients is crucial for developing accurate care plans, minimizing the risk of adverse events, and making informed decisions that support personalized medicine and optimal patient management. [23, 24] A real-time predictive model

could serve as an intraoperative decision support system for surgeons, which could help assess the risk of neurological deficits during surgery and aid in making informed choices.

Only a few articles have used ML in conjunction with IONM to predict specific outcomes. [25-27] However, to our knowledge, only one study has examined the use of IONM data to predict neurological outcomes of patients after spinal surgery with ML. Jamaludin et al. [28] used solely MEP data to categorize patients into those exhibiting positive functional outcomes and those showing no changes from preoperative to postoperative. Only baseline and final IONM measurements were used, along with a limited number of features. Consequently, many potential opportunities remain unexplored within the field of predicting neurological outcomes for patients after spinal surgery using IONM data.

2.2 Objectives

This study investigates the potential of ML techniques in predicting neurological prognosis by using IONM findings combined with additional patient data. Attention will also be directed towards identifying key features that are crucial for predicting the neurological outcome. The goal is therefore to develop a ML algorithm that can effectively predict neurological outcomes after spinal surgery using IONM data, including both motor evoked potentials (MEP) and somatosensory evoked potentials (SSEP). This study is a first step toward a real-time model that can serve as an intraoperative decision support system for surgeons. Research comparing MEP and SSEP values in spinal cord surgery indicates that MEPs outperform SSEPs in predicting postoperative neurological complications. [9, 17, 29] To more effectively determine the specific independent contribution of both separate modalities, a separate ML model is created for both MEP and SSEP in addition to a combined MEP-SSEP model.

3 Design and methods

For the clinical study a retrospective analysis was conducted, encompassing a cohort of patients who underwent spinal surgery at Haga Teaching Hospital in the Netherlands. The surgeries were performed between October 2019 and August 2023, and all procedures were monitored using IONM. Patients without sufficient IONM data or without sufficient pre- and postoperative neurological status were excluded. To ensure confidentiality, all data was anonymized and access to information that could identify individual patients was restricted solely to the research team members. The application of IONM, along with all associated parameters, remained unchanged for this study. All protocols were performed in the standard manner according to the established requirements for the surgical procedure. Due to the retrospective nature of this study, informed consent was not obtained from the patients. For an overview of the steps taken in the methodology of this study, see Figure 1.

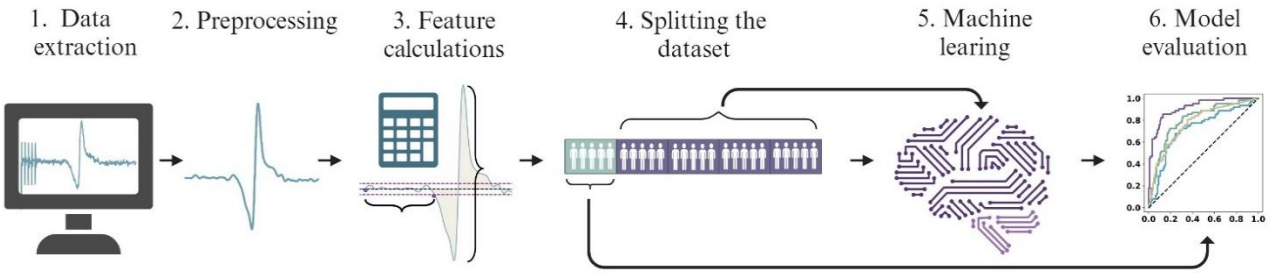

Figure 1: Schematic representation of the used methodology to conduct this study. 1: Data extraction of intraoperative neuromonitoring and additional features, 2: personal signal preprocessing to enable feature calculations, 3: Feature calculations, 4: Splitting the dataset in 80% training set and 20% testing set, 5: Machine learning algorithm trains on the training set, 6: Model evaluation based on predictions made by the machine learning algorithm and actual outcomes from the test-set. Created with BioRender.com

3.1 Data extraction

In order to develop a ML algorithm, various types of data needed to be collected, including IONM features, additional patient-related features and the neurological status three months postoperatively compared to the preoperative status. The latter is described as the outcome value and was categorized into three classes: 'Neurological stable deficits', ‘Neurologically intact’ and 'Neurological improvement'. For the purpose of this study, patients experiencing neurological deterioration were categorized as 'Neurological stable deficits'. Additional patient features and the outcome value were retrospectively extracted from the electronic health record (EHR) (see section 3.3 Feature selection). Lastly, IONM signals were retrieved from Medtronic's NIM-Eclipse®, a system used for generating, recording and storing IONM data during the surgical procedures. For a detailed user manual of the NIM-Eclipse®, refer to this link.

3.1.1 Specific data extraction IONM

All surgical procedures targeted the spine at or below the cervical-3 (C3) level, and all baseline measurements were determined by the physician assistant. Regarding the MEP signals, the overlapping monitored muscles in all procedures were the Tibialis Anterior (TA), which is innervated by the nerve roots L4, L5 and S1, the Abductor Hallucis (AH), which is innervated mostly by the S2 nerve root, and the Gastrocnemius, which is mostly innervated by the S1 and S2 nerve roots. [30] The placement of the reference electrode was standardized for surgeries concerning the thoracic and lumbar regions, positioned on the hand muscle Abductor Digiti Minimi (ADM), which is innervated by the C8-T1 nerve roots, for both the left and right sides. During surgeries involving regions above C8, the ADM was not considered a

reference muscle; instead, it was treated as a regularly measured muscle. These four muscle groups, measured bilaterally, correspond to eight measured muscle groups per patient included in this study. To activate these muscle groups, a corkscrew electrode is bilaterally positioned on the patient's motor cortex. A pulse train stimulus of four to five pulses within the range of 150 to 650 volts was applied to generate a MEP signal, while needle electrodes were placed in the measured muscles to monitor the measured muscle responses with a sampling frequency of 10.000 Hz. The train rate was set to 333 pulses per second and the pulse duration was set to 75 µs. The MEP signals are typically measured during the surgery, with an average interval of every 10 minutes. However, when operating near critical structures, the frequency of this action may increase, depending on the preference of either the physician assistant or the surgeon.

Regarding the SSEP signals, different tracts labeled under distinct names were documented for each patient. However, each patient presented one or more signals representing the left leg (L: Cz'-Fpz, L: C3'-C4', L: C3'-Fpz, L: Fpz-Cz', L: C4'-C3', L: C3'-Cz, L: C4'-Cz), as well as the right leg (R: Cz'-Fpz, R: C3'-C4', R: Fpz-Cz', R: C4'-C3', R: C4'-Cz, R: C3'-Cz). The decision was made to incorporate, from each patient, one corresponding tract for both the left and right leg, resulting in two SSEP pathways per patient. In patients where multiple signals per lower extremity were measured, the signal with the largest feature value was included (see section 3.3 Feature selection). To activate the sensory pathways, needle electrodes are inserted into the corresponding nerves, while corkscrew electrodes on the head located above the somatosensory cortex are employed to capture the signals with a sampling frequency of 10.000 Hz. Throughout the surgical procedure, the sensory pathways received continuous stimulation at a specific frequency of one Hz, with a pulse duration of 200 µs. Every 300 measurements were averaged to create one SSEP signal, which was subsequently stored using the NIM-Eclipse® system. The intensity was set within the range of 7-20 mA, and the pulse duration was fixed at 200 µs.

3.2 Raw data preprocessing

Prior to computing the ML features, the raw IONM data underwent preprocessing, which consisted of two parts. The first part involved direct raw data preprocessing within the NIM-Eclipse® system during signal measurements. Here, a band-pass filter was applied including a low-pass filter set at 800 Hz and a high-pass filter set at one Hz. The second part consisted of personal raw data preprocessing in Python 3.8 [31], starting with organizing all acquired signals per patient for specific muscles and sensory pathways. Any signals measured prior to the last indicated baseline measure were excluded from the dataset. Each signal obtained consisted of 1000 samples collected over a duration of 100 ms. To eliminate potential interference from the pulse train that could affect the feature calculations, the first 140/1000 samples of each signal were set to zero. Following this step, a fourth order zero-phase Butterworth low-pass filter was employed with a cutoff frequency of 300 Hz to filter out noise from the signal. For a visualization of the personal preprocessing of a single signal, see Figure 2.

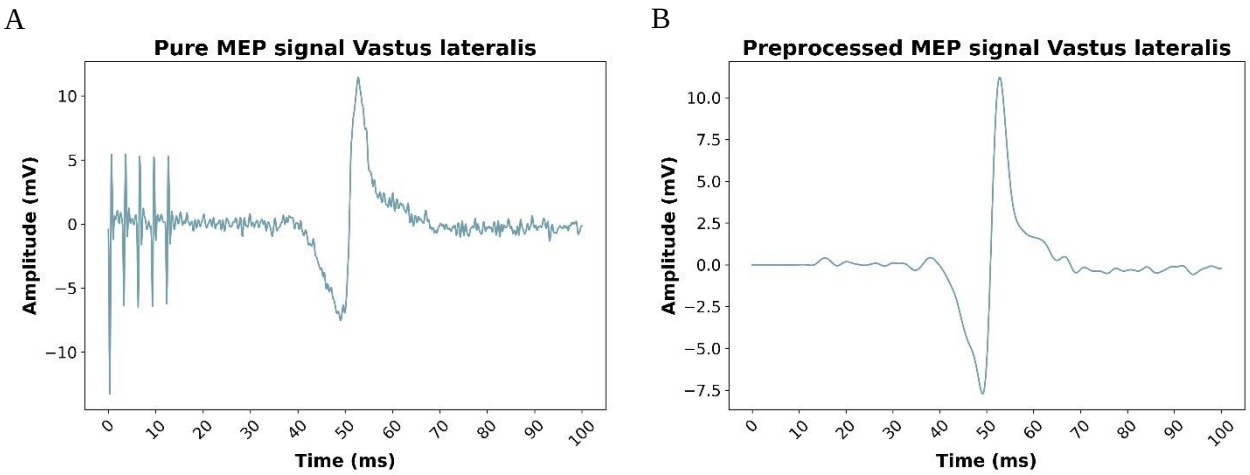

Figure 2: MEP signal obtained from a random patients' Vastus lateralis, A: Before the personal raw data preprocessing, B: After the personal raw data preprocessing. Note: In the left image, the NIM-eclipse already pre-processed the signal with a band-pass filter (1-800 Hz).

3.3 Feature selection

For every patient, 260 features were identified: 27 features for eight muscles for the MEP signals (216 features), 20 features for two tracts for the SSEP signals (40 features), and four additional features to be discussed below. This resulted in 220 features for the MEP ML model, 44 features for the SSEP model and 260 for the combined MEP-SSEP model. IONM features were calculated based on different signal characteristics, specifically peak-to-peak, area under the curve (AUC), and latency parameters. For a detailed overview of the calculated features, see Table 1. The four additional features considered in this study were primarily derived from a prior literature review aimed at identifying features significantly contributing to the prediction of neurological outcomes after spinal surgery. Among these features, the most crucial included the mean intraoperative mean arterial pressure (MAP), [32, 33] the presence or absence of preoperative neurological deficits, [8, 17, 34-36] and patient age. [19, 34] In collaboration with neurosurgeons at the Haga Teaching Hospital, it was decided to also consider the surgical area, described as 'Above thoracic 12', or 'Below thoracic 12'. When the surgical area spans vertebrae both above and below thoracic 12, the area is considered as 'Above thoracic 12'.

3.3.1 Feature calculations

Peak-to-peak values were obtained by calculating the difference between the highest and lowest point of a signal after preprocessing, and AUCs describe the absolute area under the curve up to the line where y=0, as shown in Figure 3. Latency indicates the delay between the last pulse of the pulse train and the first moment when the signal crosses a predetermined threshold. When the signal first exceeds this threshold, it is assumed that the MEP has been initiated. To determine this threshold, a segment of each signal representing the baseline of that signal was selected. The baseline ranges of the four muscles are:

- Tibialis anterior: 14 – 33 ms
- Abductor hallucis: 14 – 42 ms
- Gastrocnemius: 14 – 32 ms
- Abductor digiti minimi: 14 – 22 ms

The intervals were chosen to cover the time between the last pulse in the pulse train and the first moment an evoked potential could naturally occur, thereby indicating a baseline value. The threshold value was determined as the mean ± 3.5 times the standard deviation of this baseline segment. The corresponding time at which this point occurred was recorded. Consequently, the latency of the signal represents the time interval between the last pulse in the sequence of the pulse train and the noted time point.

Table 1: Calculation of the 256 IONM features. Feature origin: This specifies the signal property from which the features were calculated. Feature calculation description: This specifies which signal changes relative to each other were calculated and included as individual features. Modality: This column describes which modalities the feature calculations were calculated for. AUC: area under the curve, MEP: motor evoked potential, SSEP: somatosensory evoked potential.

Feature origin	**Feature calculation description**	**Modality**	
		8 MEP muscles	**2 SSEP tracts**
Peak-to-peak:	Maximum amplitude drop of a signal compared to the preceding signal	✓	✓
Peak-to-peak:	Maximum amplitude increment of a signal compared to the preceding signal	✓	✓
Peak-to-peak:	Maximum % amplitude drop of a signal compared to the preceding signal	✓	✓
Peak-to-peak:	Maximum % amplitude increment of a signal compared to the preceding signal	✓	✓
Peak-to-peak:	Maximum amplitude drop of a signal compared to the initial signal	✓	✓
Peak-to-peak:	Maximum amplitude increment of a signal compared to the initial signal	✓	✓
Peak-to-peak:	Maximum % amplitude drop of a signal compared to the initial signal	✓	✓
Peak-to-peak:	Maximum % amplitude increment of a signal compared to the initial signal	✓	✓
Peak-to-peak:	Amplitude difference of the last measured signal compared to the baseline signal	✓	✓
Peak-to-peak:	% Amplitude difference of the last measured signal compared to the baseline signal	✓	✓
AUC	Maximum AUC decrease of a signal compared to the preceding signal	✓	✓
AUC	Maximum AUC increment of a signal compared to the preceding signal	✓	✓
AUC	Maximum % AUC decrease of a signal compared to the preceding signal	✓	✓
AUC	Maximum % AUC increment of a signal compared to the preceding signal	✓	✓
AUC	Maximum AUC decrease of a signal compared to the baseline signal	✓	✓
AUC	Maximum AUC increment of a signal compared to the baseline signal	✓	✓
AUC	Maximum % AUC decrease of a signal compared to the baseline signal	✓	✓
AUC	Maximum % AUC increment of a signal compared to the baseline signal	✓	✓
AUC	AUC decrease of the last measured signal compared to the baseline signal	✓	✓
AUC	% Collapse of the last measured signal compared to the baseline signal	✓	✓
Latency	Maximum delay of a signal compared to a preceding signal	✓	
Latency	Maximum % delay of a signal compared to a preceding signal	✓	
Latency	Maximum delay of a signal compared to the baseline signal	✓	
Latency	Maximum % delay of a signal compared to the baseline signal	✓	
Latency	Delay of the last measured signal compared to the baseline signal	✓	
Latency	% Delay of the last measured signal compared to the baseline signal	✓	
Latency	Longest delay of all signals	✓	

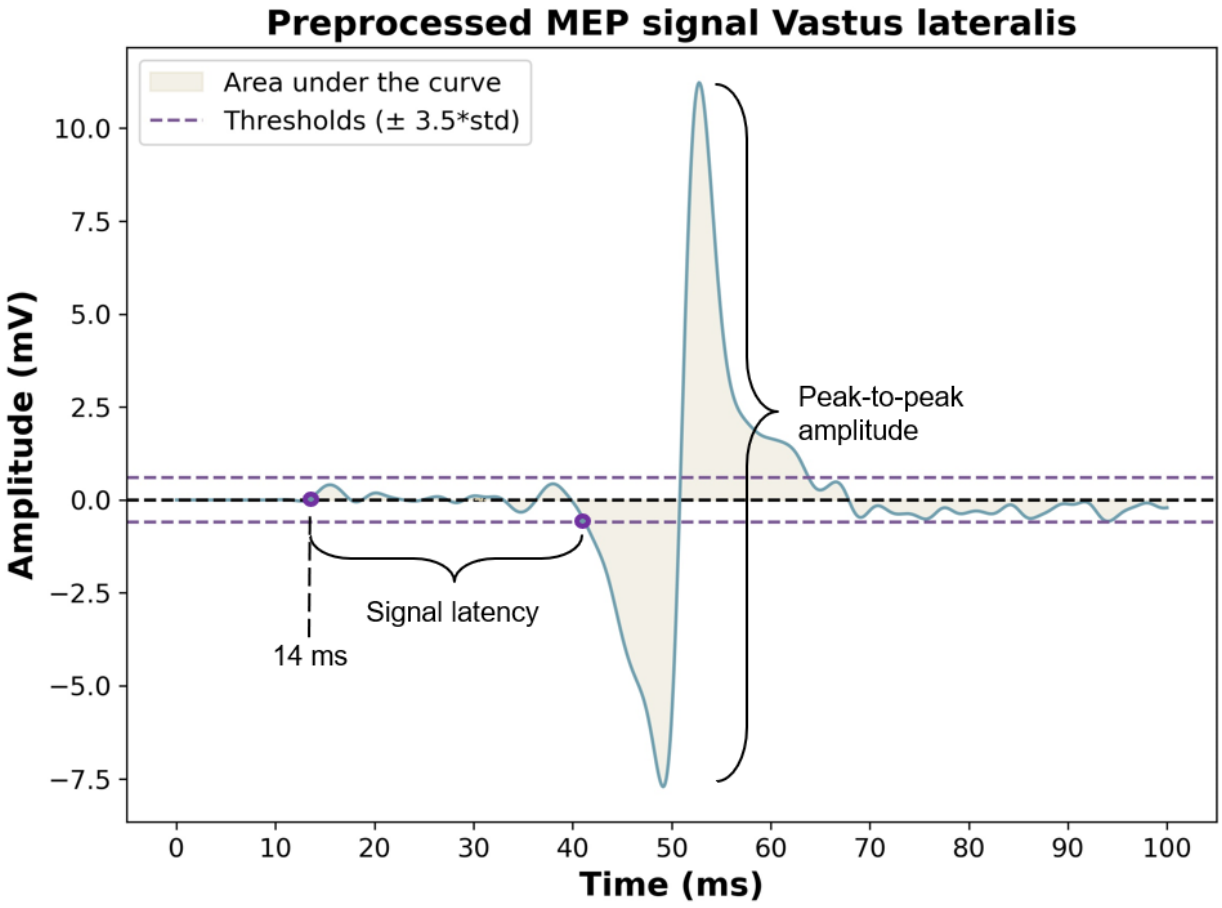

Figure 3: Visualization of the signal properties peak-to-peak amplitude, area under the curve and signal latency. std: standard deviation.

3.4 Data preprocessing for machine learning

3.4.1 Patients with insufficient data excluded

Patients with no MEP or SSEP data were removed from the respective models. Additionally, patients with more than 25% missing feature data were excluded from the model. Lastly, patients were excluded if only one signal remained after elimination of signals before the last measured baseline, since this makes signal difference calculations and subsequent determination of features impossible.

3.4.2 Handling missing values

After patients with insufficient data were removed, the table presenting all features contained four NaN (Not a Number) values. These entries specifically related to the columns corresponding to the latency features, if zero or only one signal exceeded the threshold value (see section 3.3 Feature Selection). To address this, the NaN values in these columns were replaced by the median value of the column. Furthermore, in cases where intraoperative blood pressure values were not available, the median value of the mean intraoperative MAP, derived from all other patients was used as a substitute.

3.4.3 Encoding variables

The process of variable encoding involved transforming features into numeric formats suitable for the effective use by ML algorithms. In the context of ‘absence or presence of preoperative neurological deficits’ as a feature, 'Absence' was converted to '0', while 'Presence’ was converted to '1'. Likewise, in the case of surgical area, 'Below Thoracic 12' was translated to '0', and 'Above Thoracic 12' was translated to '1'. Lastly, the outcome table underwent normalization, wherein 'Neurological stable deficits' was encoded as '0', ‘Neurologically intact’ was encoded as ‘1’ and 'Neurological improvement' was encoded as '2'.

3.4.4 Scaling

Scaling is used in ML to ensure that each feature contributes proportionally to the learning process. This is important as some ML algorithms, like K-nearest neighbors (KNN) assume that all features are centered around 0 or have a similar variance. [37] This prevents features with larger scales from overpowering those with smaller ones, ensuring the classifier learns correctly from all features as expected. Three options were considered for scaling: no scaling, applying a normally distributed scaler (‘StandardScaler’), or applying a robust range matching scaler (‘RobustScaler’). In figure 4, three different representations are plotted using the first two principal components from both the MEP and SSEP dataset. Principal components are new variables obtained from the original variables in a dataset, capturing the most substantial variance. They help reduce dimensionality while retaining the most critical information from the data. [38] This figure illustrates that data distribution clearly increases when no scaling or a StandardScaler is used, as opposed to a RobustScaler, which suggests better interpretability of the data. Since specific models that will be implemented, such as KNN and Support vector machine (SVM) (see section 3.6 Model development), require scaled data for optimal functionality, a StandardScaler is used to standardize the datasets. A StandardScaler uses the Z-score to scale the data.

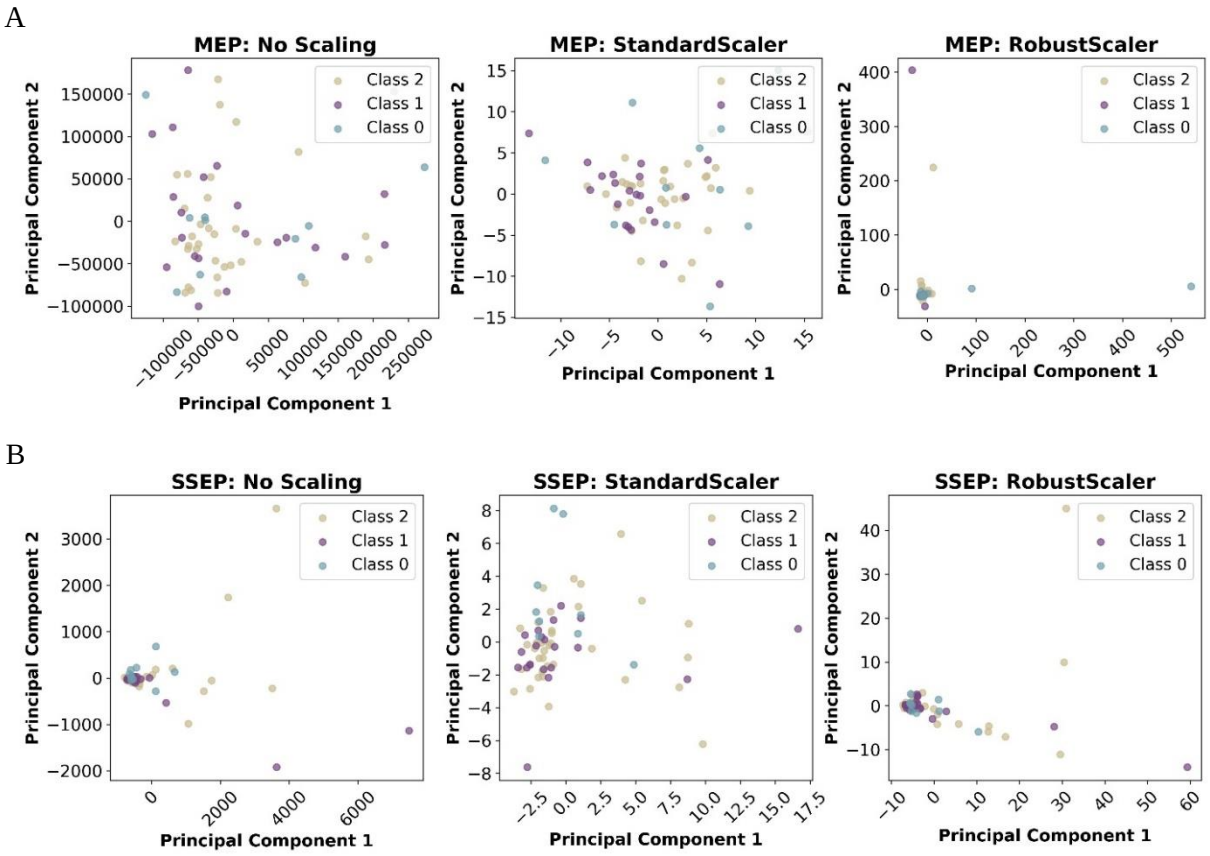

Figure 4: Distribution of different Scalers on A: the MEP dataset and B: the SSEP dataset, across the first two principal components. Left: No Scaling, middle: StandardScaler, right: RobustScaler. MEP: Motor evoked potential, SSEP: Somatosensory evoked potential.

3.5 Splitting the dataset

Cross validation (CV) is a technique used in ML to assess the performance of a model by dividing a dataset into subsets for training and testing. [39] In this process, the dataset is split into multiple segments, or folds, to train the model on a subset of the data while the remaining parts are used for testing. However, normal CV in model testing often leads to overfitting of the training set because models are evaluated on the same dataset used for tuning. [40] Nested CV, as a solution, helps avoid this problem by splitting the process into two levels, allowing hyperparameter tuning while avoiding bias in model evaluation by using different data splits for tuning and testing (see Figure 5). Restricting the hyperparameter search to a subset of the dataset significantly reduces, if not completely eliminates, the potential risk that the search process will overfit the original dataset. [40] The disadvantage of CV is that the computation time increases. In this study, both the outer and the inner CV loops were divided into five folds. For the outer loop, this means splitting the datasets in 80% for training and 20% for each fold of the loop, and for the inner loop, this means 64% for training and 16% for validation for each fold of the inner loop.

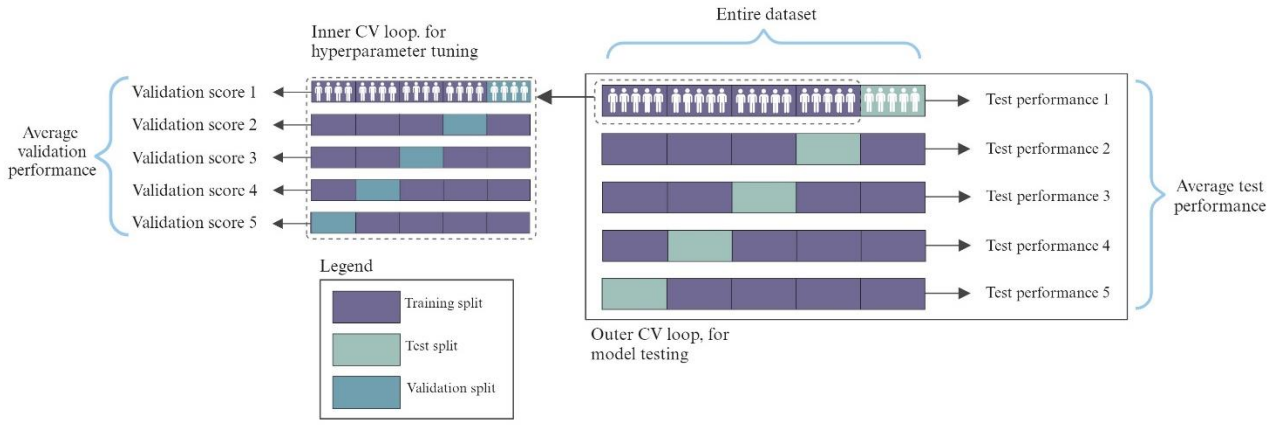

Figure 5: Schematic representation of nested cross-validation. This approach is typically used to optimize and train the model. The final performance of the model on unseen data is obtained by averaging the test performances over the outer cross-validation folds. Created with BioRender.com

3.6 Model development

To predict neurological outcome after spinal surgery using IONM data with ML, we used multiple classification models. These models were all trained using the same train-test split that was created during nested cross-validation (see 3.5 Splitting the dataset). Four different ML classifiers were used to construct the models:

- Support vector machine
- K-nearest neighbors
- Random forest
- Extreme gradient boosting

3.6.1 Support Vector Machine

The primary goal of SVM (Figure 6A) is to find the optimal hyperplane that best separates data points into different classes. [41] SVM works by maximizing the distance between the hyperplane and the closest data points of each class, and proves to be a suitable classifier to test on our dataset, as it remains effective even when the number of dimensions exceeds the number of samples. [37]

3.6.2 K-nearest neighbors

The classification of new datapoints in KNN (Figure 6B) is determined by the classes of their nearest data point, aiding in informed decisions about its specific class. The parameter 'k' indicates the number of neighbors taken into account in this classification process. For small datasets, KNN has a relatively short computational time. Additionally, this classifier is of interest for our dataset, as it exhibited the best performance in the study conducted by Jamaludin et al. [28]

3.6.3 Random forest

Random forest (RF) (Figure 6C) is an ensemble learning method that combines multiple decision trees to generate more accurate predictions. [42] Every decision tree is created by randomly selecting data from the available dataset. Nodes within these trees divide the dataset into smaller subgroups based on feature values categorized as high or low. The overall prediction of a RF is achieved by combining the predictions of each decision tree. RF serves as an effective classifier when the number of variables exceeds the number of samples. [43] In addition, the simplicity of computing feature importance plots in this algorithm is important for analyzing key features in the outcome prediction.

3.6.4 Extreme gradient boosting

Extreme gradient boosting (XGBoost) (Figure 6D), an implementation of gradient boosting, constructs an ensemble model by combining several decision trees. Unlike a random forest, training takes place iteratively, successively adding new trees while taking into account the errors of previous trees. [44] This iterative process allows XGBoost to focus on correcting the shortcomings of previous models.

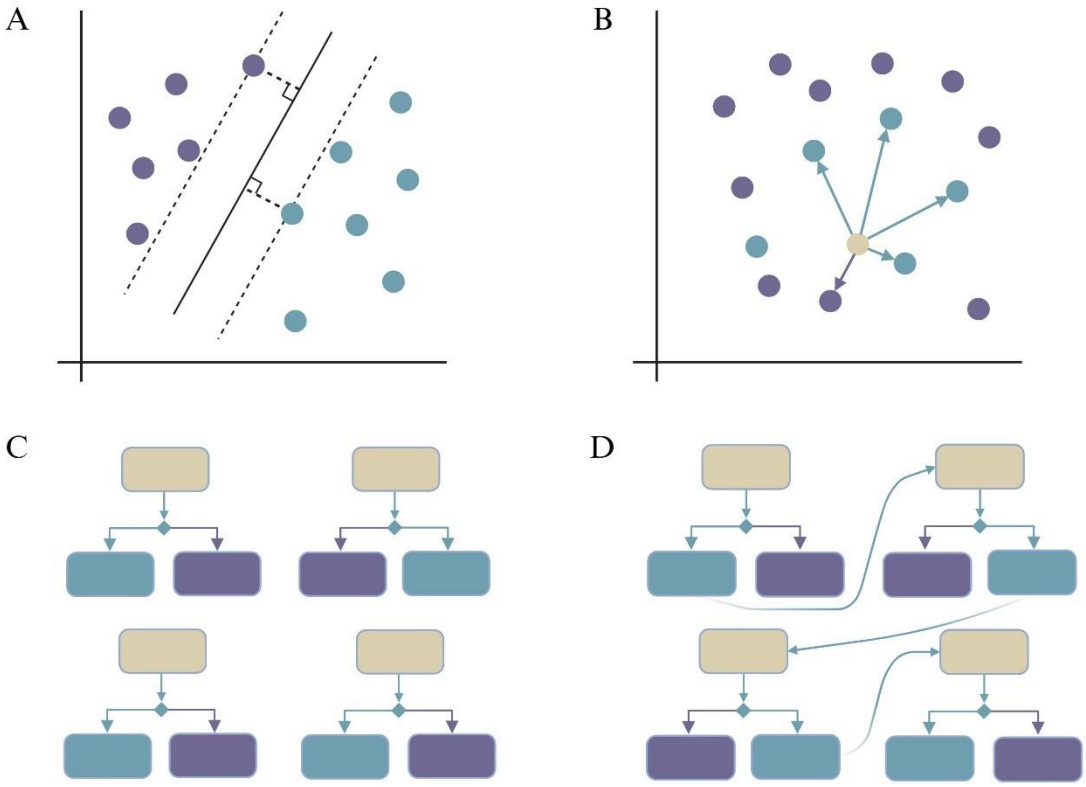

Figure 6: Schematic representation of the used classification techniques. A. Support vector machine; B. K-nearest neighbors; C. Random forest; D. Extreme gradient boosting. Created with BioRender.com

3.7 Model Evaluation

After constructing the classification models using the training data, these models were applied to the test set to evaluate their performance. A series of evaluation metrics was used to assess the models, including sensitivity (recall), specificity, accuracy, and when applicable, precision. Table 2 shows these metrics and their corresponding formulas. In addition, to visually assess the models, receiver operating characteristic (ROC) plots were generated, alongside precision-recall plots, and when applicable, feature importance plots. The precision-recall plot proves valuable in scenarios with imbalanced data within a multiclass problem. In this case, accuracy and specificity are often high, resulting in a high AUC score. The precision-recall plot gives a good representation of performance for such unbalanced data because true negatives (TN) are not included in the calculations. To visually assess all models in one plot, micro-averaged ROC curves and precision-recall curves were generated. These curves are a summarized performance measure in statistics, often used in the context of classification problems. It adjusts for differences in class size and ensures that the largest class has greater significance, which is preferable in scenarios with imbalanced data. [37]

Table 2: Performance metrics and their corresponding formulas. TP: true positive rate, FN: false negative rate, TN: true negative rate, FP: false positive rate.

Performance Metrics	**Corresponding formula**
Sensitivity (recall)	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Accuracy	$\frac{TN + TN}{TP + FN + FP + TN}$
Precision	$\frac{TP}{TP + FP}$

4 Results

4.1 Data

Initially, data was obtained from 69 patients, who underwent surgery between October 2019 and August 2023. The age distribution of the patients ranged from 19 to 87 years (median: 58, standard deviation: 16.6). Of these patients, 48 showed preoperative neurological deficits such as loss of strength or hypesthesia, while the remaining 21 reported no preoperative neurological deficits. 39 patients exhibited neurological improvement, 20 patients were neurologically intact, and 10 patients showed neurological stable deficits, of whom three experienced slight neurological deterioration during or after the surgery in the study period. After patients with insufficient data were removed, the SSEP model included 65 patients, the MEP model included 64 patients, and the MEP-SSEP model included of 62 patients. (see Figure 7) In total, 67 patients were included in one of the three ML models. (see Table 3).

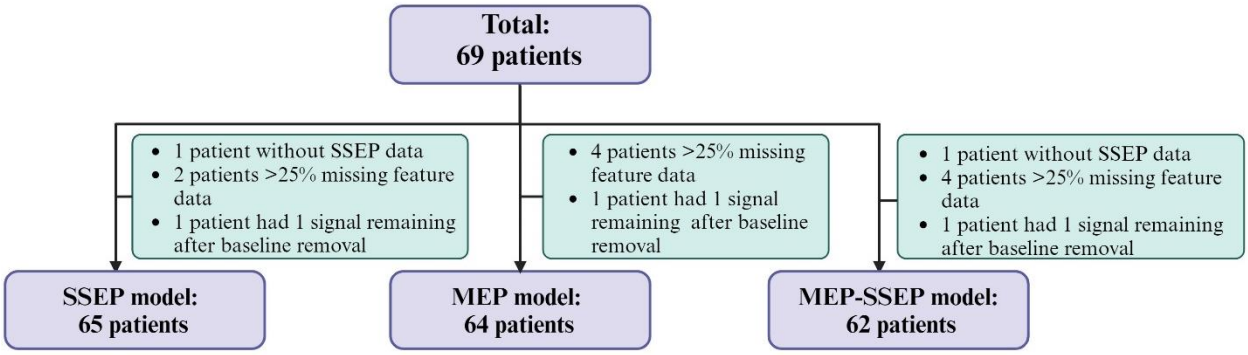

Figure 7: Flowchart of the patients included for each modality with their exclusion criteria. MEP: motor evoked potential, SSEP: somatosensory evoked potential. Created with BioRender.com

Table 3 Patient demographics and clinical data of all 67 patients included in at least one of the three machine learning models. MAP: Mean arterial pressure, C: Cervical, T: Thoracic, L: Lumbar, S: Sacral.

Patient	Age (Years)	Preoperative neurological deficits	Mean intraoperative MAP (mmHg)	Reason for surgery	Three months postoperative neurological outcome
1	85	Yes	78.39	Extirpation extramedullary tumor C3-C5	Neurological improvement
2	42	No	94.97	Extramedullary tumor L4	Neurologically intact
3	65	Yes	81.81	Ll-L2 intramedullary tumor	Neurological improvement
4	42	No	85.10	Tumor L2 intradural	Neurologically intact
5	45	Yes	69.32	Untethering (laminectomy L5)	Neurological improvement
6	43	Yes	99.11	Intradural extramedullary tumor T4	Neurological stable deficits
7	46	No	99.30	Resection intradural tumor C7	Neurologically intact
8	70	Yes	76.67	Thoracic tumor	Neurological improvement
9	69	No	72.24	Intradural extramedullary tumor T3	Neurological improvement
10	76	Yes	88.85	Tumor T11-T12	Neurological improvement
11	81	No	79.59	Tumor extirpation L4-L5	Neurological stable deficits
12	60	Yes	98.24	Extirpation intradural tumor T1-T2	Neurological improvement
13	54	Yes	87.25	Tumor extirpation C3-C4	Neurological improvement
14	68	Yes	81.71	Resection intradural tumor T4-5	Neurological stable deficits
15	50	Yes	91.93	Resection intradural tumor L1-L2	Neurological improvement
16	57	Yes	107.78	Resection intradural tumor T7-T8	Neurological improvement
17	72	Yes	73.30	Fixation T10-S1	Neurological improvement
18	41	Yes	81.81	Kyphosis correction T2-10	Neurological improvement
19	64	No	78.95	Tumor extirpation L2	Neurologically intact
20	51	Yes	78.98	Biopsy spinal cord tumor C5-6	Neurological improvement
21	35	No	70.45	Scoliosis correction T1-L3	Neurologically intact
22	77	Yes	80.26	Extramedullary intradural tumor T4	Neurological improvement

Patient	Age (Years)	Preoperative neurological deficits	Mean intraoperative MAP (mmHg)	Reason for surgery	Three months postoperative neurological outcome
23	58	Yes	72.85	Extirpation intraspinal osteophyte T12	Neurological stable deficits
24	54	Yes	77.34	Cystic lung metastasis L3-L4	Neurological improvement
25	45	Yes	88.67	Ependymoma T8-T10	Neurological stable deficits
26	79	No	93.73	Vertebral Fusion Surgery T8-L3	Neurologically intact
27	26	No	76.36	Extramedullary tumor T3-T4	Neurologically intact
28	68	Yes	92.06	Vertebral Fusion Surgery T10-S1	Neurological improvement
29	53	No	69.05	Intradural tumor L1-L2 & sacral	Neurologically intact
30	19	Yes	77.76	Resection schwannoma T11	Neurological improvement
31	50	No	91.44	Resection intradural tumor L4	Neurologically intact
32	68	Yes	82.55	Intradural tumor Th10	Neurological improvement
33	77	Yes	87.22	Resection intradural tumor L5-S1	Neurological improvement
34	59	Yes	87.37	Scoliosis correction Th6-S1	Neurological improvement
35	25	No	72.70	Resection intradural tumor C7	Neurologically intact
36	26	No	85.73	Resection intradural tumor Th12	Neurologically intact
37	24	Yes	85.12	Unilateral cordotomy low thoracic	Neurological improvement
38	61	Yes	88.15	Scoliosis correction Th12-L4	Neurological improvement
39	38	Yes	70.40	Untethering lipoma cauda equina	Neurological improvement
40	37	No	72.16	Intramedullary cyst C4-C7	Neurologically intact
41	69	No	85.61	Intradural tumor C7-Th1	Neurologically intact
42	73	Yes	92.47	Biopsy intramedullary tumor T5-Th7	Neurological stable deficits
43	66	Yes	83.16	Intradural tumor conus medullaris	Neurological improvement
44	75	No	81.81	Extirpation tumor Th11-12	Neurologically intact
45	53	Yes	87.15	Corpectomy C5 and C6	Neurological improvement
46	75	Yes	87.175	Tumor extirpation Th5	Neurological improvement
47	74	No	82.54	Tumor extirpation Th9-Th10	Neurologically intact
48	69	Yes	92.18	Tumor extirpation Th1-Th2	Neurological improvement
49	24	Yes	83.64	Tumor extirpation L2-L3	Neurological improvement
50	41	No	65.07	Tumor L3-L4	Neurologically intact
51	56	Yes	74.97	Laminectomy C3-C7	Neurological stable deficits
52	57	Yes	76.79	Laminectomy and lateral fixation C4–C6	Neurological improvement
53	87	Yes	102.00	Extirpation meningioma Th10-Th11	Neurological improvement
54	53	Yes	80.54	Decompression Th2-Th5 & L3-L5	Neurological stable deficits
55	69	No	78.66	Scoliosis correction Th11-ilium	Neurologically intact
56	77	No	71.43	Resection intradural tumor C6-C7	Neurological improvement
57	66	Yes	78.73	Extirpation tumor Th9-Th10	Neurological improvement
58	59	Yes	76.38	Scoliosis correction Th10-S1	Neurological improvement
59	49	Yes	66.21	Resection conus medullaris Th11-Th12	Neurological stable deficits
60	54	No	79.48	Thoracic kyphosis	Neurologically intact
61	61	No	101.52	Scoliosis correction Th10-ilium	Neurologically intact
62	85	Yes	83.65	Extirpation meningiomas Th5 and Th8	Neurological improvement
63	84	Yes	135.92	Laminectomy C7-Th1	Neurological improvement
64	50	Yes	71.64	Untethering spinal cord	Neurological improvement
65	39	Yes	61.66	Untethering spinal cord	Neurological improvement
66	58	Yes	89.75	Removal of hemangioblastoma Th8-9	Neurological stable deficits
67	71	Yes	92.88	Ependymoma C3	Neurological improvement

4.2 Overall classifier results

Figure 8 lists all the classifier performances applied to the MEP, SSEP, and combined MEP-SSEP datasets. In addition, Figure 9 shows the micro-averaged ROC curves and the micro-averaged precision-recall curves for the four classifiers in the combined MEP-SSEP model. See Appendix B for an overview of the parameter grids used for hyperparameter tuning to obtain these outcomes. Micro-averaged ROC- and precision-recall curves regarding the separate MEP and SSEP models can be found in Appendix C. Both Figure 8 and Figure 9 illustrate that the XGBoost classifier outperformed the other classifiers on all evaluated aspects. The combined MEP-SSEP model using XGBoost achieved a sensitivity of 70.4%, specificity of 88.3%, accuracy of 87.1%, and precision of 73.7% across the three outcome classes. For XGBoost, the micro-averaged ROC curve showed an AUC of 0.92, and the micro-averaged precision-recall curve showed an average precision (AP) of 0.86.

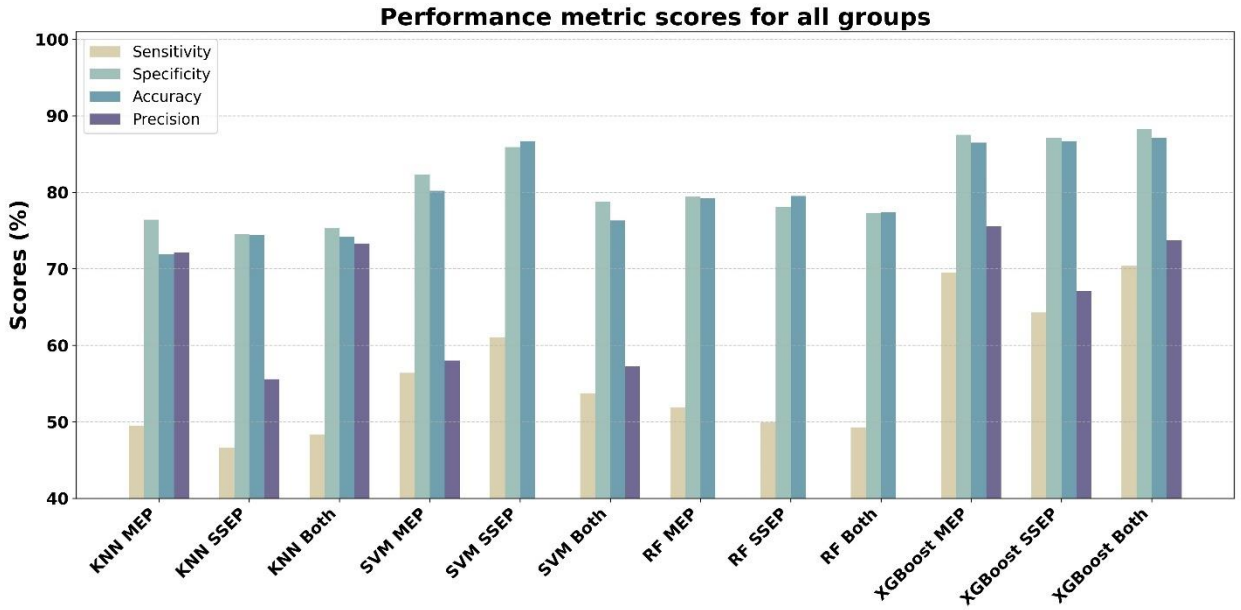

Figure 8: Classifier performance metrics for each modality MEP, SSEP and Both (MEP & SSEP). Scores are percentages and are a combined score over the three outcome classes 'Neurological stable deficits', 'Neurologically intact' and 'Neurological improvement'. KNN: K-nearest neighbors, SVM: support vector machine, RF: random forest, XGBoost: extreme gradient boosting, MEP: motor evoked potential, SSEP: somatosensory evoked potential. Note: Some models couldn't calculate the precision due to division by zero.

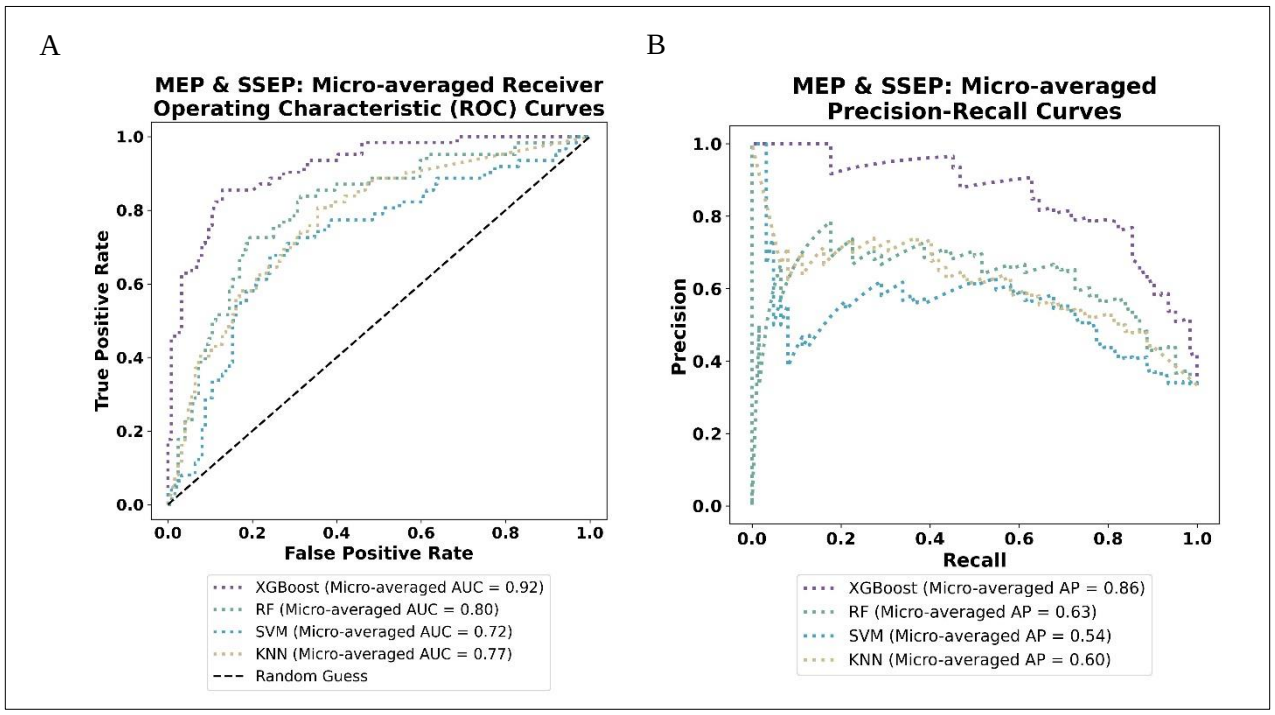

Figure 9: Micro averaged curves regarding the combined MEP-SSEP model A: Micro-averaged Receiver operating characteristics (ROC) curves for all classifiers. B: Micro-averaged precision-recall curves for all classifiers. XGBoost: extreme gradient boosting, RF: random forest, SVM: support vector machine, KNN: K-nearest neighbors, MEP: motor evoked potential, SSEP: somatosensory evoked potential.

4.3 Individual modalities results

Given the superior performance of XGBoost, a closer look on the results of this classifier will follow. Figure 10 shows ROC curves for each individual class, plotted for the combined MEP-SSEP, MEP, and SSEP models. The highest scores are observed in the combined MEP-SSEP model, with an area under the curve of 0.70 for 'Neurological stable deficits,' 0.98 for 'Neurologically intact,' and 0.87 for 'Neurological improvement'. Figure 11 shows the respective precision-recall curves for each individual modality for each outcome class. ‘Neurological stable deficits’ was best predicted by the MEP model (AP = 0.39), whereas ‘Neurologically intact’ and ‘Neurological improvement’ were best predicted by the combined MEP-SSEP model, with APs of 0.97 and 0.86 respectively. To verify that the model does not overfit, a technique was used in which the outcomes are chosen at random. (App D)

To later discuss the contributions comparing the individual MEP and SSEP models, the scores of these individual models are also considered. The individual MEP model achieved an AUC of 0.67 for neurologically stable deficits, slightly higher than 0.64 of the SSEP model, while both MEP and SSEP achieved an AUC of 0.94 for neurologically intact. (see Figure 10) In predicting neurological improvement, the MEP model scored an AUC of 0.86, slightly outperforming the SSEP model (AUC = 0.81). Examining the precision-recall curves, the MEP model exhibits higher APs: 0.39 for predicting neurological stable deficits compared to SSEP's 0.27 (see Figure 11). In predicting neurologically intact, MEP and SSEP achieve 0.95 and 0.94 respectively, while for neurological improvement, MEP scores 0.83, exceeding SSEP's 0.77.

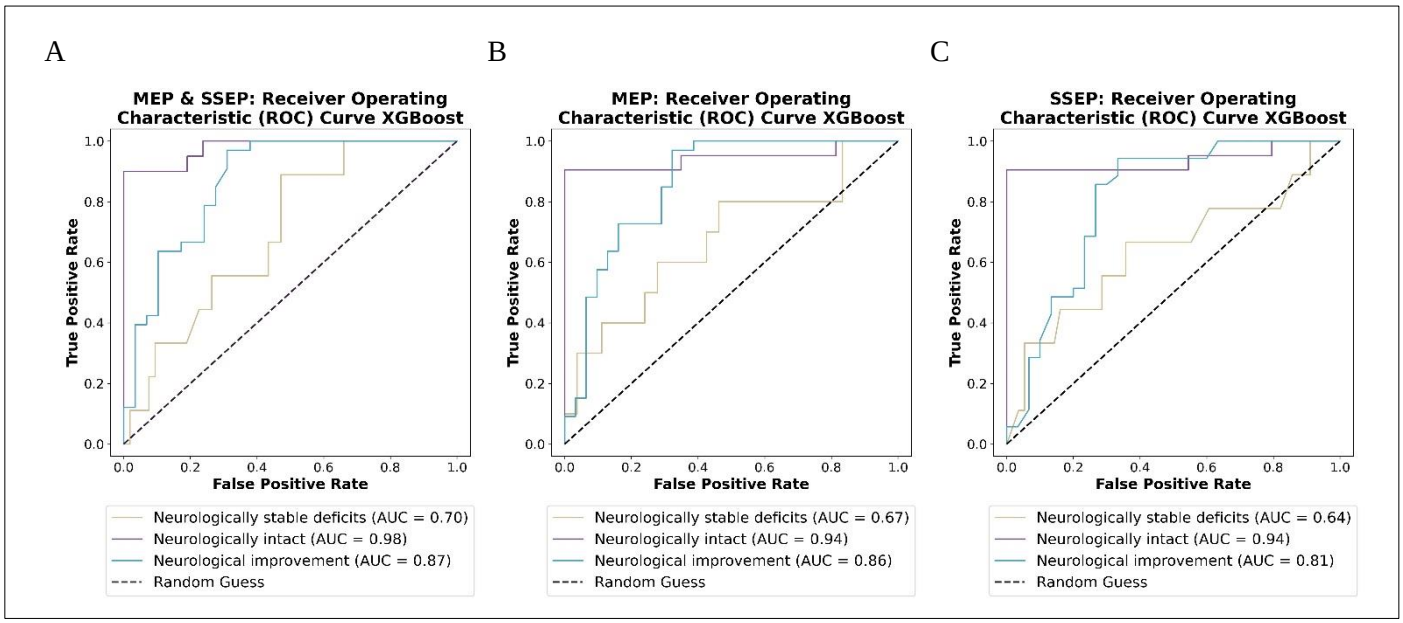

Figure 10: ROC curves per individual outcome class by the XGBoost classifier for all separate models. A: MEP & SSEP, B: MEP, C: SSEP. MEP: motor evoked potential, SSEP: somatosensory evoked potential, XGBoost: extreme gradient boosting.

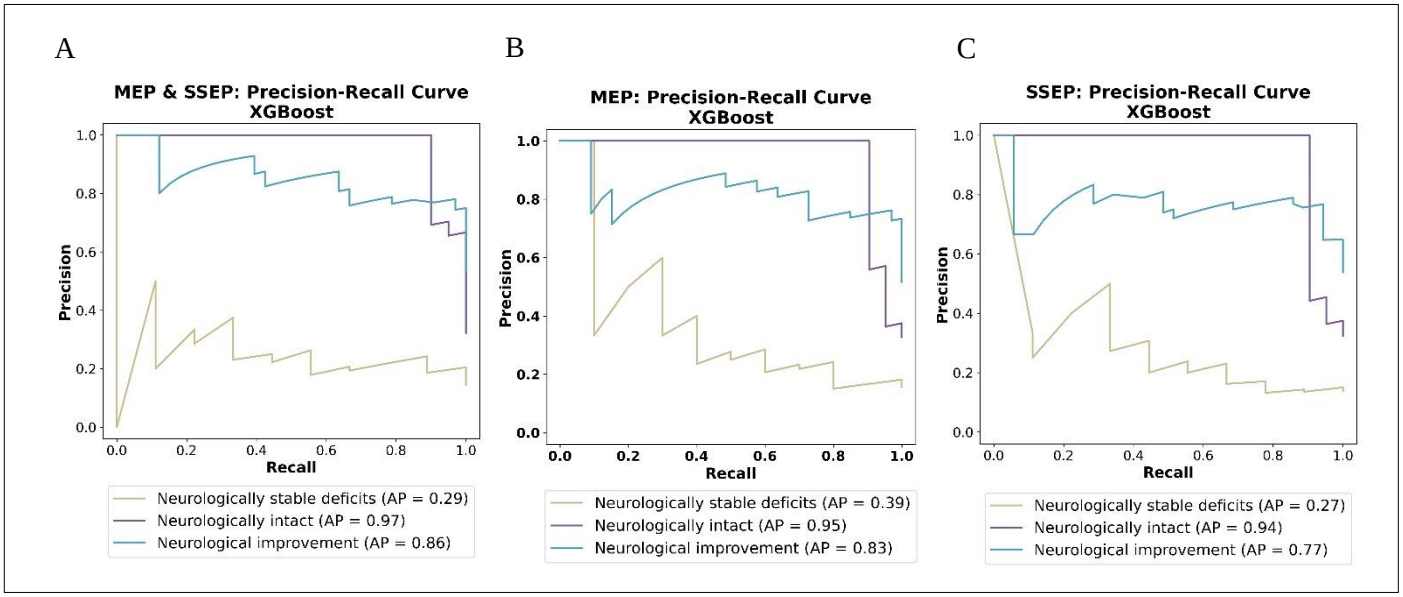

Figure 11: Precision-recall curves per individual outcome class by the XGBoost classifier for all separate models. A: MEP & SSEP, B: MEP, C: SSEP. MEP: motor evoked potential, SSEP: somatosensory evoked potential, XGBoost: extreme gradient boosting.

To assess the associated key predicting features, the top seven most contributing features were plotted in Figure 12 for the three different modalities. In all three models, the presence or absence of preoperative neurological deficits is the most prominent feature, contributing about 29%, 16% and 33% in the predictions of the MEP-SSEP model, the MEP model and the SSEP model, respectively. This is followed by the MEP feature 'latency of the last measured TA signal relative to baseline', contributing 13.5% and 7% for the MEP-SSEP model and the MEP model, respectively. The second most predictive feature in the SSEP model is the 'percentage AUC increase in a signal from baseline for the right leg', contributing for 9%.

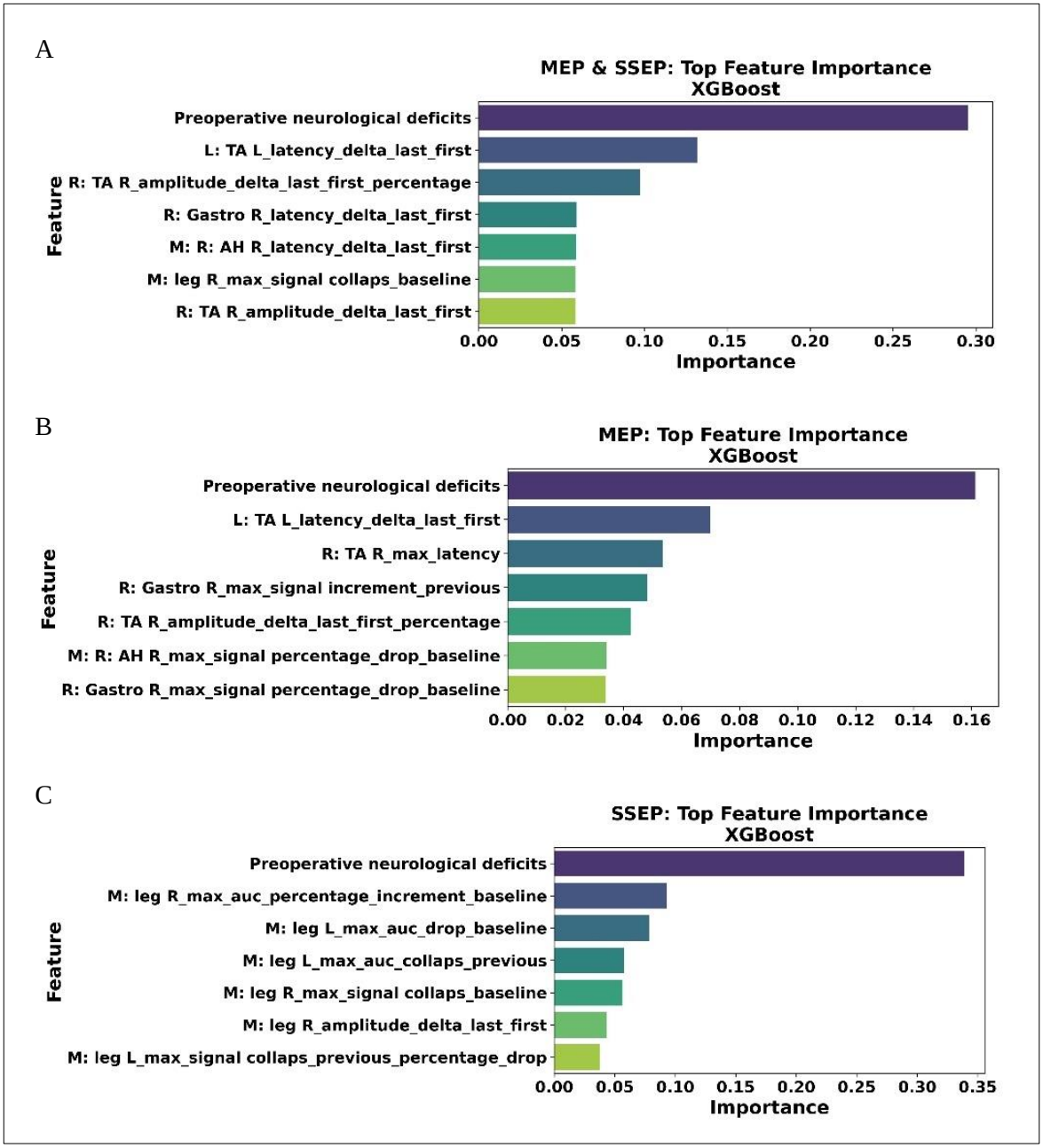

Figure 12: Feature importance plots for the three models. The seven features holding the highest predictive value are plotted. A: MEP & SSEP, B: MEP, C: SSEP. Additional features age, surgical area and mean intraoperative MAP were not incorporated in this figure because of their negligible values; For the MEP-SSEP, MEP and SSEP model respectively: Age 0%, 0%, 0%, surgical area 0%, 0%, 1.5%, mean intraoperative MAP: 0.6%, 0%, 0%. MEP: motor evoked potential, SSEP: somatosensory evoked potential, XGBoost: extreme gradient boosting.

5 Discussion

In this study, ML was used to predict neurological outcomes three months postoperatively, using both IONM features and additional patient-related features. The aim was to correctly classify patients into the three groups: 'Neurological stable deficits', 'Neurologically intact' and 'Neurological improvement', and analyze its key predicting features. Four ML classifiers were constructed, with XGBoost showing superior performance in all performance metrics. Notably, the combined MEP-SSEP model exhibited the highest scores for sensitivity: 70.4%, specificity: 88.3% and accuracy: 87.1%, while the MEP model exhibited the highest performance for precision: 75.6%. Overall, the model shows a high predictive performance for predicting neurological outcomes within the three distinct groups. On average, the combined MEP-SSEP model yields the most favourable outcome scores: AUC = 0.70, 0.98, 0.87 and AP = 0.29, 0.97, 0.86 for 'Neurological stable deficits', 'Neurologically intact' and 'Neurological improvement', respectively. These findings not only highlight the ability to predict neurological outcomes, but also shed light on key predicting features and identify those with less significance.

The one similar study, conducted by Jamaludin et al. [28] focused on the outcome measures 'Neurological improvement' and 'No improvement', when predicting neurological outcomes after spinal surgery using IONM data. Their best-performing model, employing a KNN classifier, showed relatively high sensitivity (87.5%) but low specificity (33.3%). Comparison of these findings with our results shows higher specificity (88.3%) but lower sensitivity (70.4%) in our study. The higher mean scores in our study can be attributed to the integration of a larger set of IONM features along with their specific calculations, the inclusion of additional features, and the targeted training of multiple algorithms using nested cross-validation. Therefore, it is important for future research to implement these applications. Nevertheless, it is important to remember that higher sensitivity is preferred over higher specificity in this type of study. This preference arises from the goal of correctly classifying patients into the positive group (true positives) rather than correct classification of the negative group (true negatives). Hence, future research should explore methods to enhance sensitivity without excessively compromising specificity.

5.1 Individual MEP and SSEP contributions

Previous research indicates a superior predictive value for MEP signals compared to SSEP signals. For example, Antkowiak et al. (2022) [29] showed higher sensitivity (92.3%) and specificity (81.8%) for MEPs as opposed to SSEPs (50% sensitivity, 81% specificity), while comparing significant IONM alerts with a patient's postoperative neurological outcome. Consistent with these trends were studies conducted by Cannizzaro et al. (2022) [17] and Dauleac et al. (2022). [9] In our study, the MEP model outperformed the SSEP model slightly with higher sensitivity (69.5% vs. 64.3%), specificity (87.5% vs. 87.1%) and precision (75.6% vs. 67.1%), while the SSEP model performs slightly better on accuracy scores (SSEP: 86.7% vs. MEP: 86.5%). Looking at the specific contributions per individual class predictions, MEP outperformed SSEP slightly on both AUC's and APs. Though MEP has higher average scores, which is in line with the existing literature, the similarity of these values highlights the importance of including SSEP signals in the intraoperative assessment. Despite challenges in interpreting real-time SSEP signals, their individual importance in predicting neurological outcomes three months postoperatively is nearly equal to MEP importances. However, the feature importance plots show that of the top seven features in the combined MEP-SSEP model, five relate to MEP, while only one relates to SSEP. Thus, when both modalities are accessible, MEP features have greater predictive value.

5.2 Best predicting features

The presence or absence of preoperative neurological deficits is the most important feature in all three models, contributing about 29%, 16% and 33% in the predictions of the MEP-SSEP model, the MEP model and the SSEP model, respectively. Nevertheless, it is important to note that patients without preoperative neurological deficits consistently belong to the 'Neurologically intact' outcome group, as none of them showed neurological deterioration. This clarifies why the models show strong performance in this group when the preoperative neurological deficits feature is taken into account. To assess the exact impact of this feature, a sensitivity analysis was performed by removing this feature for the combined MEP-SSEP model (App E). The findings show that the model retains its predictive ability in terms of AUCs, showing 0.71, 0.75, and 0.54 for 'Neurological stable deficits', 'Neurologically intact' and 'Neurological improvement', respectively. However, a significant performance reduction was observed for the 'Neurologically intact' and 'Neurological improving' groups. Moreover, the calculated APs indicate unreliable predictions without this characteristic, with scores of 0.26, 0.57 and 0.58 for 'Neurological stable deficits', 'Neurologically intact' and 'Neurological improvement', respectively. Given the class imbalance within our model leading to high specificity, the precision-recall curve emerges as a crucial performance measure. (see section 3.3 Model evaluation) Consequently, the reliability of the model depends on the inclusion of this feature, and predictions based solely on IONM features cannot be considered reliable. Therefore, accurate neurological assessment until shortly before surgery is crucial for predicting the patient's neurological status three months postoperatively. The importance of this feature in our model is in line with the existing literature. [8, 17, 34-36] For example, Cannizzaro et al. [17] stated that the most important predictor associated with an improved score on the modified McCormick scale was the patients' preoperative neurological status ($p<0.001$), and Wang et al. [8] showed that the patients with preoperative spinal deficits have a higher decreasing percentage in IONM amplitude than those patients without. Similar trends were shown by Merali et al. [34], Zhang et al. [36] and Shimizu et al. [35]

The results reveal additional notable findings regarding the latency features. Currently, physician assistants at Haga teaching hospital rely primarily on signal collapse and signal increment to interpret MEP and SSEP signals intraoperatively. Signal latencies are not included in this assessment. However, the feature importance plots show that the most prominent feature following preoperative neurological deficits is a latency feature. For the combined MEP-SSEP model it contributes 13.5% to the outcomes, and in the MEP model approximately 7%. Furthermore, in the best performing model, the combined MEP-SSEP model, not only the second-best feature, but three out of the top five features happen to be latency features. This highlights the importance of including signal latency in the visual intraoperative assessment, alongside signal collapse and increment, to provide a more comprehensive evaluation of the signals. The current literature barely addresses the relationship between MEP signal latencies and postoperative neurological outcomes as amplitude changes seem to have a greater predictive value. [45] Nevertheless, our study shows different results, highlighting the need to explore the neurological prediction of latency features in future research.

5.3 Least predicting features

Besides the most important features, specific features hold no significant weight in predicting the three months postoperative neurological status. Age holds a significance of 0% in all three models. Likewise, the surgical area, higher or lower than T12, hold a significance of 0% in the MEP-SSEP and the MEP model, and a significance of 1.5% in the SSEP model. This suggests that the decision to use IONM should be irrespective of surgical area or age. However, it should be noted that the surgical area feature is strongly normalized as above or below T12. In further research, when a larger group of patients is applicable, it might be useful to modify this feature to a specific vertebra to investigate its inclusion. Furthermore, the mean intraoperative MAP also does not figure prominently in the results. In the individual MEP and SSEP

model, this feature has an importance of 0%, and in the combined MEP-SSEP model it holds a significance of 0.6%. This contradicts existing literature, which suggests that this feature has a high predictive value for postoperative neurological deficits. For instance, the study by Agarwal et al. [32] suggests that none of the 87 patients with a Mean intraoperative MAP higher than 96.3 mmHg showed neurological improvement after emergency decompressive surgery for acute spinal cord injury. However, in our study we still see neurological improvements up to the highest measured mean intraoperative MAP of 135.92 mmHg. In addition, the patient with the lowest measured mean intraoperative MAP (61.66 mmHg) showed neurological improvement. In current practice, if IONM signals collapse with a low or high MAP, surgery is usually postponed until the MAP stabilizes. However, our findings suggest that excessively high or low MAP may not be a direct indicator of poor neurological outcomes three months postoperative. Thus, in the future, consideration may be given to resume the surgery sooner when signals deviate in combination with blood pressure anomalies.

5.4 Further findings

Independent of the machine-learning models used, an important observation emerged during this project. All patients included in the study without preoperative deficits had no neurological worsening postoperatively. In contrast, 10 out of the 48 patients with preoperative neurological deficits showed no neurological improvement (20.8%), of whom three deteriorated (6.3%). Therefore, for patients who wish to postpone surgery until actual neurological symptoms are experienced, such as some patients with a benign spinal cord tumor, may be strongly advised to have the surgery before developing these symptoms. This significantly reduces their chances of maintaining or developing postoperative neurological deficits. Furthermore, it is essential to reflect on the fact that only three patients in our study showed neurological deterioration. Ideally, we would also include a group of patients who were not monitored with IONM to compare their outcomes and determine the rate of deterioration, but this was beyond the scope of this study.

5.5 Limitations

It would have been valuable for this study to include a separate outcome group with patients experiencing neurological worsening, but due to the insufficient number of representations in our patient group, this was not feasible. A multicenter study could have overcome this limitation by including a larger group of patients. This, in turn, would improve the training of the ML algorithm, likely leading to better results. Furthermore, examining possible additional features is essential, especially considering the two different types of signal collapse observed during surgery: sudden signal collapse due to compression and in addition for MEP signals, progressive signal collapse due to ischemia. [3] For this type of signal collapse, it would be desirable to include signal features that can be measured over time. Perhaps the algorithm could make important predictions based on these temporal features. However, to include these signal features, an approximately consistent number of signals must be measured for each patient to effectively run the ML algorithm. This poses a practical challenge because of variations in operative time, diverse critical moments during surgery and variations in the duration of anaesthetic clearance. In addition, in our study, only four muscle groups were measured in each patient, facilitating their inclusion in this type of ML, regardless of whether this was an affected muscle or not. It is possible that our model would show greater predictive value of IONM features if the affected muscle was included in the ML model for each patient. However, conducting similar studies with this recommendation presents challenges because this specific muscle group must be measured in each patient individually to train the ML algorithm, which is often not the case in practice. Combining muscle groups for the ML model is undesirable since the nerve roots of most muscles already split in the spinal cord. [30] One last limitation is that we assessed the neurological status three months postoperative. However, analyzing data over a longer postoperative follow-up period would provide more comprehensive and conclusive results.

5.6 Future implementation

Before a real-time predictive model can be deployed in the operating room, several steps must be completed. After optimizing the ML model, performing extensive training on a large number of patients, incorporating additional features and specific affected muscles, an intuitive interface must be designed that allows the medical staff to seamlessly enter patient data. This interface must integrate effectively with the ML model. In addition, the NIM-Eclipse must load real-time data into the model. Next, the model must be thoroughly tested and validated on patient case studies in various scenarios in a controlled environment, evaluating performance characteristics and responsiveness in real time. Compliance with local regulations and ethical guidelines regarding the use of ML models in medical settings is crucial and requires approval under the Medical Device Regulation (MDR). Once approved, the model can be integrated into the operating room. Training sessions need to be provided for the medical staff on how to use the system. Also, feedback needs to be collected from physicians on the performance and usability of the model. The ML model needs to be trained regularly with updated data to improve performances to changing patient conditions. Finally, conduct post-implementation evaluations by consistently comparing model outcomes with patients' actual neurological status to identify incorrectly predicted cases or areas in need of improvement.

6. Conclusion

A reliable prediction of neurological outcomes three months postoperatively can be made combining MEP and SSEP IONM features, provided that the patient's preoperative status is accurately documented and included in the prediction. Though either MEP or SSEP features alone offer predictive value, MEP features show superior predictive values compared to SSEP features when both modalities are accessible, with latency emerging as a prominent predictive IONM feature.

7. References

1. Olmsted, Z.T., et al., *Direct Wave Intraoperative Neuromonitoring for Spinal Tumor Resection: A Focused Review.* World Neurosurg X, 2023. **17**: p. 100139.
2. Tamaki, T. and T. Yamane, *Proceedings: Clinical utilization of the evoked spinal cord action potential in spine and spinal cord surgery.* Electroencephalogr Clin Neurophysiol, 1975. **39**(5): p. 539.
3. Charalampidis, A., et al., *The Use of Intraoperative Neurophysiological Monitoring in Spine Surgery.* Global Spine J, 2020. **10**(1 Suppl): p. 104S-114S.
4. Decruz, J., et al., *Neuromonitoring in Cervical Spine Surgery: When Is a Signal Drop Clinically Significant?* Asian Spine J, 2021. **15**(3): p. 317-323.
5. Nguyen, V., et al., *EMG, MEP, and SSEP in the Intraoperative Neurophysiologic Monitoring of Lumbar Surgeries.* Neurology, 2019. **92**(15).
6. Park, J.H. and S.J. Hyun, *Intraoperative neurophysiological monitoring in spinal surgery.* World J Clin Cases, 2015. **3**(9): p. 765-73.
7. Zha, X., et al., *A Deep Learning Model for Automated Classification of Intraoperative Continuous EMG.* IEEE Trans Med Robot Bionics, 2021. **3**(1): p. 44-52.
8. Wang, S., et al., *Frequent neuromonitoring loss during the completion of vertebral column resections in severe spinal deformity surgery.* Spine J, 2017. **17**(1): p. 76-80.
9. Dauleac, C., et al., *Predictors of functional outcome after spinal cord surgery: Relevance of intraoperative neurophysiological monitoring combined with preoperative neurophysiological and MRI assessments.* Neurophysiol Clin, 2022. **52**(3): p. 242-251.
10. Reddy, R.P., et al., *What is the predictive value of intraoperative somatosensory evoked potential monitoring for postoperative neurological deficit in cervical spine surgery?-a meta-analysis.* Spine J, 2021. **21**(4): p. 555-570.
11. Park. D and Kim, I., *Application of Machine Learning in the Field of Intraoperative Neurophysiological Monitoring: A Narrative Review.* MDPI, 2022.
12. Yang, J.L., et al., *A proposed classification system for guiding surgical strategy in cases of severe spinal deformity based on spinal cord function.* Eur Spine J, 2016. **25**(6): p. 1821-9.
13. Appel, S., et al., *Effect of Intra- and Extraoperative Factors on the Efficacy of Intraoperative Neuromonitoring During Cervical Spine Surgery.* World Neurosurg, 2019. **123**: p. e646-e651.
14. Qiu, J., et al., *Intra-operative neurophysiological monitoring in patients with dystrophic neurofibromatosis type 1 scoliosis.* Somatosens Mot Res, 2021. **38**(2): p. 95-100.
15. Sielatycki, J.A., et al., *A novel MRI-based classification of spinal cord shape and CSF presence at the curve apex to assess risk of intraoperative neuromonitoring data loss with thoracic spinal deformity correction.* Spine Deform, 2020. **8**(4): p. 655-661.
16. Sun, S.P., et al., *Integration of MRI and somatosensory evoked potentials facilitate diagnosis of spinal cord compression.* Sci Rep, 2023. **13**(1): p. 7861.
17. Cannizzaro, D., et al., *Intramedullary spinal cord tumors: the value of intraoperative neurophysiological monitoring in a series of 57 cases from two Italian centers.* J Neurosurg Sci, 2022. **66**(5): p. 447-455.
18. Park, D., et al., *Usefulness of Intraoperative Neurophysiological Monitoring During the Clipping of Unruptured Intracranial Aneurysm: Diagnostic Efficacy and Detailed Protocol.* Front Surg, 2021. **8**: p. 631053.
19. Kim, S.H., S.H. Lee, and D.A. Shin, *Could Machine Learning Better Predict Postoperative C5 Palsy of Cervical Ossification of the Posterior Longitudinal Ligament?* Clin Spine Surg, 2022. **35**(5): p. E419-e425.
20. Javaid, M., *Significance of machine learning in healthcare: Features, pillars and applications.* ScienceDirect, 2022.
21. Senders, J.T., et al., *Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review.* World Neurosurg, 2018. **109**: p. 476-486.e1.

22. Joshi, R.S., et al., *State-of-the-art reviews predictive modeling in adult spinal deformity: applications of advanced analytics.* Spine Deform, 2021. **9**(5): p. 1223-1239.
23. Karabacak, M. and K. Margetis, *A Machine Learning-Based Online Prediction Tool for Predicting Short-Term Postoperative Outcomes Following Spinal Tumor Resections.* Cancers (Basel), 2023. **15**(3).
24. Katsos, K., et al., *Current Applications of Machine Learning for Spinal Cord Tumors.* Life (Basel), 2023. **13**(2).
25. Kortus, T., *Automated Robust Interpretation of Intraoperative Electrophysiological Signals – A Bayesian Deep Learning Approach.* Current Directions in Biomedical Engineering, 2021.
26. Qiao, N., et al., *Deep Learning for Automatically Visual Evoked Potential Classification During Surgical Decompression of Sellar Region Tumors.* Transl Vis Sci Technol, 2019. **8**(6): p. 21.
27. Verdonck, M., et al., *Exploratory Outlier Detection for Acceleromyographic Neuromuscular Monitoring: Machine Learning Approach.* J Med Internet Res, 2021. **23**(6): p. e25913.
28. Jamaludin, M.R., et al., *Machine Learning Application of Transcranial Motor-Evoked Potential to Predict Positive Functional Outcomes of Patients.* Comput Intell Neurosci, 2022. **2022**: p. 2801663.
29. Antkowiak, L., et al., *Relevance of intraoperative motor evoked potentials and D-wave monitoring for the resection of intramedullary spinal cord tumors in children.* Neurosurg Rev, 2022. **45**(4): p. 2723-2731.
30. Zhu, L., H.D. Lin, and A.M. Chen, *Accurate segmental motor innervation of human lower-extremity skeletal muscles.* Acta Neurochir (Wien), 2015. **157**(1): p. 123-8.
31. Van Rossum, G. and F.L. Drake, *Python 3 Reference Manual*. 2009, CreateSpace100 Enterprise Way, Suite A200Scotts ValleyCA.
32. Agarwal, N., et al., *Decision tree-based machine learning analysis of intraoperative vasopressor use to optimize neurological improvement in acute spinal cord injury.* Neurosurg Focus, 2022. **52**(4): p. E9.
33. Chou, A., et al., *Expert-augmented automated machine learning optimizes hemodynamic predictors of spinal cord injury outcome.* PLoS One, 2022. **17**(4): p. e0265254.
34. Merali, Z.G., et al., *Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy.* PLoS One, 2019. **14**(4): p. e0215133.
35. Shimizu, T., et al., *Efficacy of a machine learning-based approach in predicting neurological prognosis of cervical spinal cord injury patients following urgent surgery within 24 h after injury.* J Clin Neurosci, 2023. **107**: p. 150-156.
36. Zhang, H., et al., *Predictive Risk Factors of Poor Preliminary Postoperative Outcome for Thoracic Ossification of the Ligamentum Flavum.* Orthop Surg, 2021. **13**(2): p. 408-416.
37. Pedregosa, F., et al., *Scikit-learn: Machine Learning in {P}ython.* Journal of Machine Learning Research, 2011. **12**: p. 2825--2830.
38. Jolliffe, I. and J. Cadima, *Principal component analysis: a review and recent developments.* The Royal Society, 2016.
39. Gunasegaran, T. and Y. Cheah, *Evolutionary cross validation*, in *2017 8th International Conference on Information Technology (ICIT)*. 2017. p. 89-95.
40. Wainer, J. and G. Cawley, *Nested cross-validation when selecting classifiers is overzealous for most practical applications.* Expert Systems with Applications, 2021. **182**.
41. Awad, M. and R. Khanna, *Support Vector Machines for Classification*. Efficient Learning Machines. 2015: Apress, Berkeley, CA.
42. Ali, J., *Random Forests and Decision Trees* International Journal of Computer Science, 2012.
43. Janitza, S. and E. Celik, *A computationally fast variable importance test for random forests for high-dimensional data.* 2018. **12**(4): p. 885-915.
44. Bentéjac, C., Csörgő, A. & Martínez-Muñoz, G., *A comparative analysis of gradient boosting algorithms.* 2020. **54**: p. 1937-1967.

45. Liu, T., et al., *Variability of somatosensory evoked potential and motor evoked potential change criteria in thoracic spinal decompression surgery based on preoperative motor status.* Spine J, 2023.

7. Supplementary materials

Appendix A. IONM warning criteria

Table A-I: Warning criteria for the percentage signal collapse indicated by intraoperative neuromonitoring (IONM) evoked potential values, extracted from various articles and presented along their respective journals. SSEP: somatosensory evoked potential, MEP: motor evoked potential, D-waves: Direct waves.

Article	Journal	SSEP (%)	MEP (%)	D-waves (%)	SSEP Latency (%)	MEP latency (%)
Park et al.(2015)[6]	World Journal of Clinical Cases	> 50	> 50	< 50	> 10	> 10
Olmsted et al. (2023)[1]	World Neurosurgery			< 50		
Appel et al. (2019)[13]	World Neurosurgery	> 50	> 80		> 10	
Cannizzaro et al. (2022)[17]	Journal of Neurosurgical Sciences		100	20 < x < 50		
Qiu et al.(2021)[14]	Somatosensory & Motor Research	> 50	> 80		> 10	
Sielatycki et al. (2020)[15]	Spine Deformity Journal	> 50			> 10	
Wang et al. (2017)[8]	European Spine Journal	> 50	> 80			
Dauleac et al. (2022)[9]	Journal of clinical neurophysiology	> 50	>50	< 50	> 10	
Yang et al. (2016)[12]	European Spine Journal	> 50			> 10	
Sun et al. (2023)[16]	Scientific reports - Nature	> 50			> 10	

Appendix B. Parameter grids

Table B-I: Parameter grid used for hyperparameter tuning for each classifier.

Classifier	Parameter grid
Random forest	'model__n_estimators': [80, 100, 150, 200], 'model__max_depth': [None, 5], 'model__min_samples_split': [10, 15], 'model__min_samples_leaf': [3, 4]
K-nearest neighbors	'model__n_neighbors': [4, 5], 'model__weights': ['uniform', 'distance'], 'model__p': [1]
Support vector machine	'model__C': [100, 1000], 'model__kernel': ['rbf', 'sigmoid'], 'model__probability': [True]
XGBoost	'model__learning_rate': [0.005, 0.01] 'model__n_estimators': [60, 80]

Appendix C. Micro-averaged ROC- and precision-recall curves MEP and SSEP

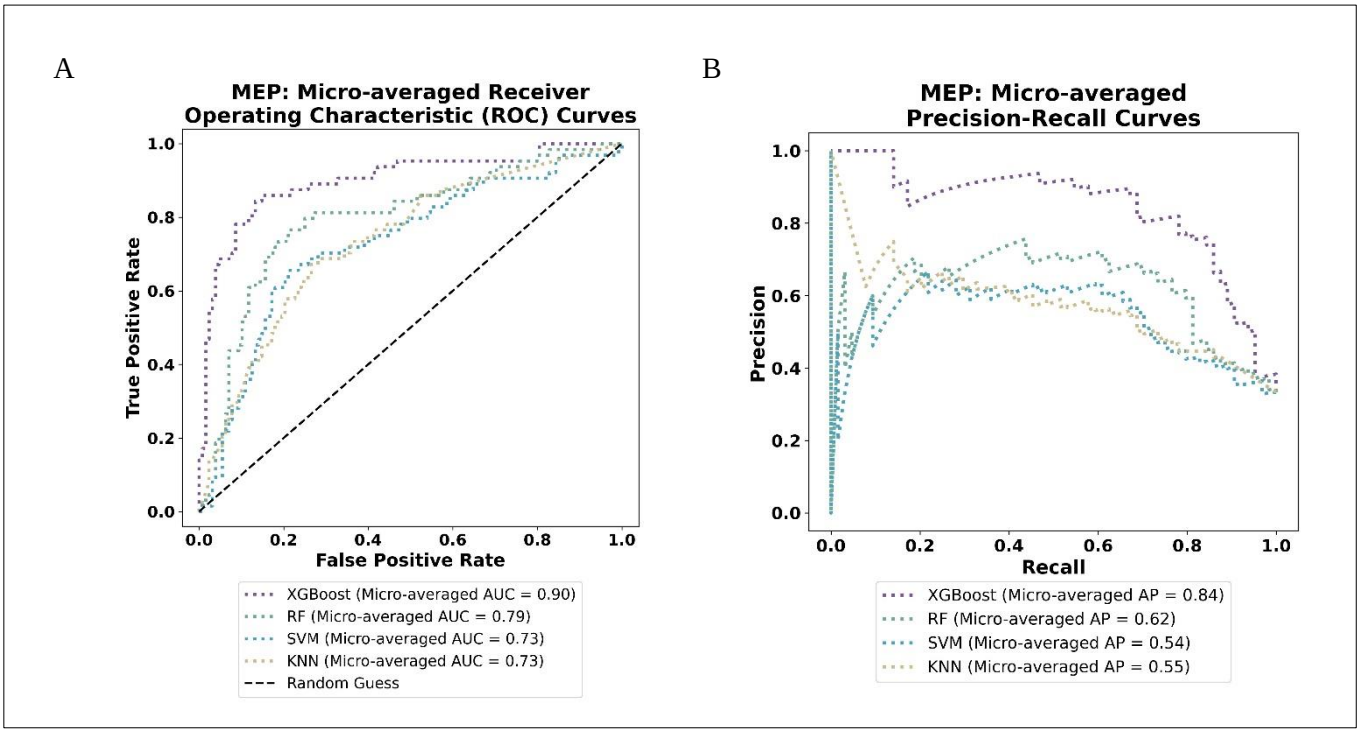

Figure C-I: Micro averaged curves regarding the MEP model. A: Micro-averaged receiver operating characteristic (ROC) curves for all classifiers. B: Micro-averaged precision-recall curves for all classifiers. XGBoost: extreme gradient boosting, RF: random forest, SVM: support vector machine, KNN: K-nearest neighbors.

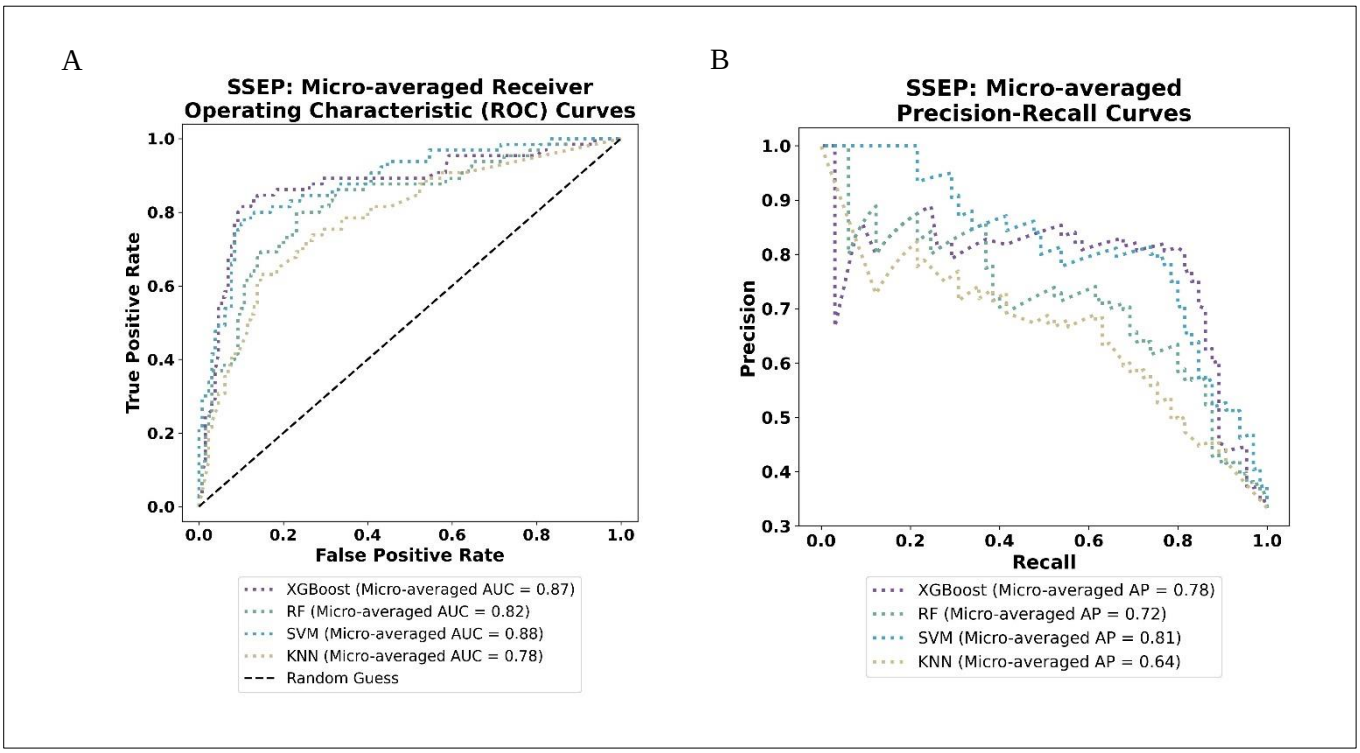

Figure C-2: Micro averaged curves regarding the SSEP model. A: Micro-averaged receiver operating characteristic (ROC) curves for all classifiers. B: Micro-averaged precision-recall curves for all classifiers. XGBoost: extreme gradient boosting, RF: random forest, SVM: support vector machine, KNN: K-nearest neighbors.

Appendix D. MEP-SSEP random outcome ROC curves

To assess the performance of the model and ensure that it does not overfit, we used a technique in which the outcomes are chosen at random. By choosing the outcomes randomly, the model should find no relation between the features and the outcome, which should reflect in the ROC-AUC curve with a value around ±50%, as can be seen in Figure D-I.

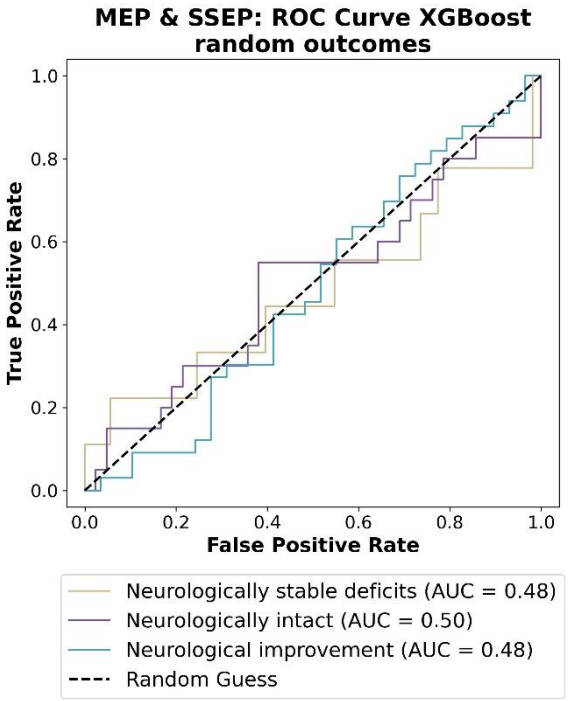

Figure D-I: ROC curves per individual class regarding the MEP-SSEP model for XGBoost. All outcomes are randomized which results in outcomes ±50%. MEP: Motor evoked potential, SSEP: Somatosensory evoked potential, XGBoost: extreme gradient boosting.

Appendix E. Sensitivity analysis MEP-SSEP model

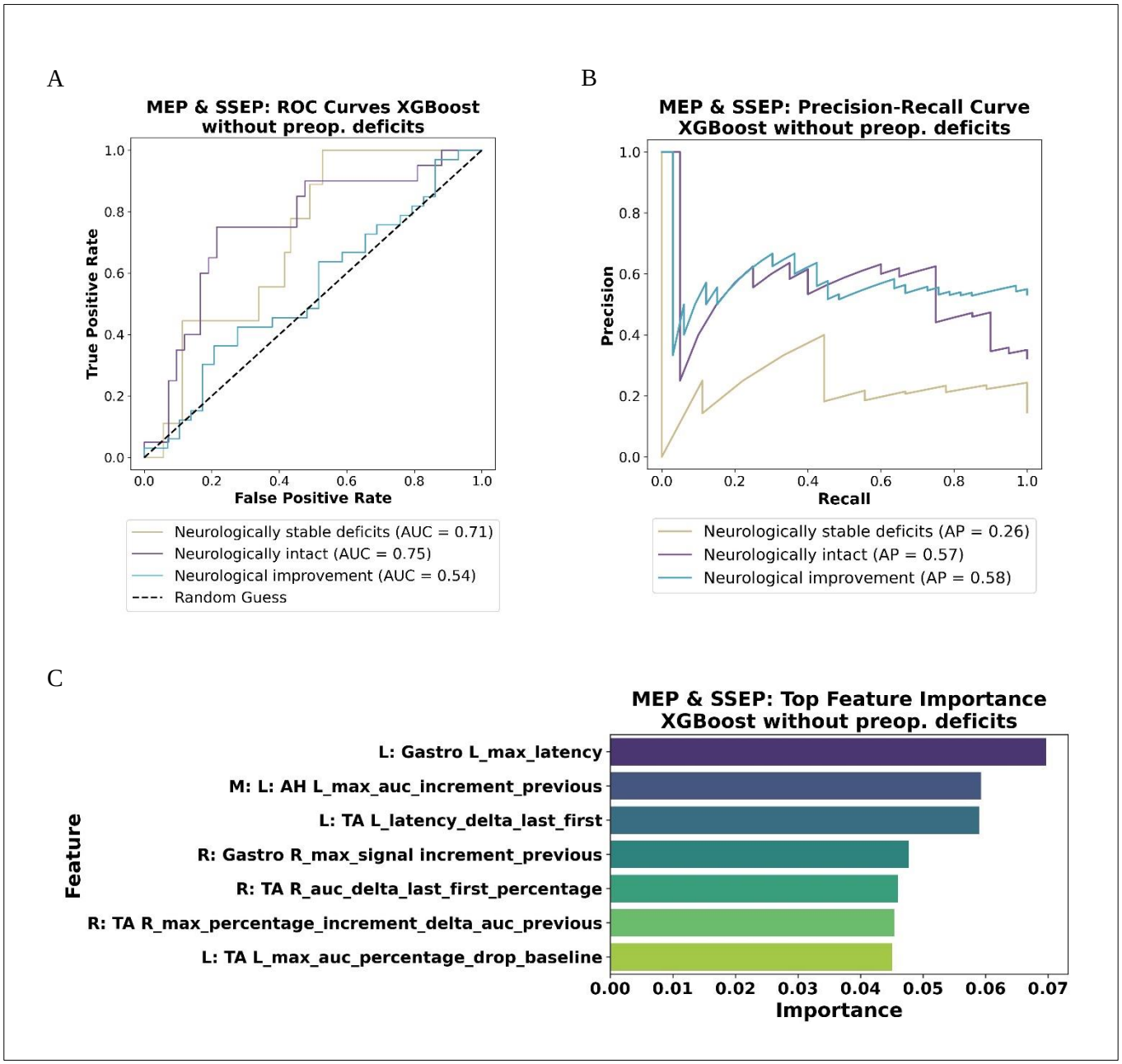

Figure E-I: Performance plots of the MEP-SSEP model trained with the XGBoost classifier without the feature preoperative neurological deficits. A: ROC curves for each individual class, B: precision-recall curves for each individual class, C: Feature importance plot with the top seven predictive features. ROC: receiving operator characteristic, MEP: motor evoked potential, SSEP: somatosensory evoked potential, XGBoost: extreme gradient boosting