15 (2026) 22–34

Contents lists available at ScienceDirect

Deep Brain Stimulation

journal homepage: www.journals.elsevier.com/deep-brain-stimulation

Can intra-operative stimulation data predict post-operative Deep Brain Stimulation outcomes? — A machine learning approach based on probabilistic maps

Vittoria Bucciarelli [a,b,*], Teresa Nordin [c], Marc Stawiski [a], Karin Wårdell [a,c], Jérôme Coste [d], Jean-Jacques Lemaire [d], Philippe Derost [d], Bérengère Debilly [d], Raphael Guzman [e], Simone Hemm [a,c,1], Dorian Vogel [a,1]

[a] Institute for Medical Engineering and Medical Informatics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Muttenz, Switzerland
[b] Department of Biomedical Engineering, University of Basel, Allschwil, Switzerland
[c] Department of Biomedical Engineering, Linköping University, Linköping, Sweden
[d] Université Clermont Auvergne, CNRS, CHU Clermont-Ferrand, Clermont Auvergne INP, Institut Pascal, Clermont-Ferrand, France
[e] Department of Neurosurgery, University Hospital Basel, Basel, Switzerland

ARTICLE INFO

Keywords:
Deep Brain Stimulation
Effect Prediction
Machine Learning

ABSTRACT

Introduction: Deep Brain Stimulation is an established treatment for movement disorders. Data-driven approaches like probabilistic mapping and predictive modeling are being increasingly used to optimize stimulation parameter selection. However, it remains unclear whether intra-operative test data alone can reliably inform such models. Clarifying the predictive value of intra-operative data is essential, as it may influence data collection strategies and surgical and programming protocols.
Objective: This study examined whether post-operative DBS effects can be accurately predicted using intra-operative data alone or if post-operative information is required.
Methods: A dataset comprising 1117 intra-operative and 1553 post-operative stimulation tests from 35 patients (14 Essential Tremor, 21 Parkinson's Disease) was analyzed. Volumes of Tissue Activated (VTAs) were simulated to generate probabilistic maps, from which mapping and target anatomy-related features were extracted to train predictive classification models (low and high improvement, side effects). Model performance was assessed across scenarios linking intra- and post-operative data.
Results: Intra-operative data effectively identified regions associated with optimal stimulation, highlighting their utility in guiding parameter selection (predictive accuracy ~60%). However, the predictive relationship between VTAs, probabilistic maps, and clinical outcomes differed between intra- and post-operative contexts. When trained solely on intra-operative VTAs, the model performed at a near chance level (predictive accuracy ~35%).
Discussion: The study demonstrates that while intra-operative data are useful for identifying optimal target regions, they are insufficient for accurately predicting post-operative DBS outcomes. Incorporating post-operative data remains crucial for reliable and clinically meaningful DBS effect prediction.

Abbreviations: BA, Balanced Accuracy; BF, Bayes Factor; DBS, Deep Brain Stimulation; EF, Electric Field; ET, Essential Tremor; KNN, K-Nearest Neighbors; PD, Parkinson's Disease; PSS, Probabilistic Sweet Spot; PSM, Probabilistic Stimulation Maps; SE, Side Effects; SMOTE, Synthetic Minority Over-sampling Technique; STN, Subthalamic nucleus; Vim, Ventrointermediate nucleus of the thalamus; VTA, Volume of Tissue Activated; XGBoost, Gradient Boosted Decision Tree classifier.

* Corresponding author at: Institute for Medical Engineering and Medical Informatics, School of Life Sciences, University of Applied Sciences and Arts Northwestern Switzerland FHNW, Muttenz, Switzerland.
E-mail address: vittoria.bucciarelli@fhnw.ch (V. Bucciarelli).

[1] These authors share last authorship

https://doi.org/10.1016/j.jdbs.2026.09.003
Received 15 April 2026; Received in revised form 7 September 2026; Accepted 7 September 2026
Available online 8 September 2026

1. Introduction

Deep Brain Stimulation (DBS) is a well-established therapy for movement disorders, particularly Essential Tremor (ET) and Parkinson's Disease (PD) [1]. Its therapeutic effect relies on delivering electrical stimulation to specific anatomical targets in the brain. These targets differ depending on the disease and symptoms being treated, and in some cases also vary between clinical centers [2,3]. The success of DBS depends not only on accurate electrode placement but also on careful adjustment of stimulation parameters. Optimal settings are typically identified during one or more programming sessions, where clinicians test multiple configurations to maximize symptom relief while minimizing side effects. This trial-and-error process is often lengthy and uncomfortable for patients [4]. Data-driven tools have been developed to make programming faster and more efficient. Probabilistic mapping (PSM) identifies brain regions that lead to symptom improvement or worsening in a patient cohort [5,6]. This is achieved by simulating the volumes of tissue activated (VTAs) by specific stimulation settings [7], transferring them into a common anatomical space (e.g. anatomical atlas) [8–10] and applying voxel-wise statistical testing to identify voxels significantly linked to the investigated effect [11]. Effect prediction models build on this knowledge to estimate the likely outcomes of specific stimulation settings. By providing clinicians with informed starting points, they can reduce the number of adjustments needed and improve the overall patient experience [12]. Several models have been proposed to predict the DBS effect, using a variety of input features. Most are based on machine learning [13–22], with a smaller number of studies applying deep learning approaches [23,24]. The features used are derived from stimulation settings [23,25], probabilistic mapping and VTAs [13,16,18–20,22,26–28], demographic and anatomical information [13,29–32], structural [33] and functional [12] connectivity analyses, and electrophysiology measurements from microelectrode recordings (MER) [34] or local field potentials (LFP) [35]. Different combinations of these inputs have been tested, yielding accuracies higher than 70% in both binary [20,28] and three-class [13] classification tasks. However, the VTA or PSM-based studies mentioned above have been trained and validated on data collected post-operatively. In previous work, we have shown that effect prediction approaches can also be successfully applied to intra-operative stimulation data [36]. Building on this, the aim of the present study was to integrate intra-operative and post-operative stimulation tests within a unified DBS effect prediction framework. Specifically, we sought to determine whether the prediction of post-operative outcomes could be informed or enhanced by incorporating intra-operative data, a question of clinical relevance since an accurate intra-operative data-based prediction could potentially reduce the need for extensive post-operative programming. Moreover, understanding the usefulness of intra-operative data may inform the design of surgical protocols, particularly decisions regarding the use of local versus general anesthesia. To this end, we analyzed a patient cohort with extensive intra-operative and post-operative testing, extracted a set of VTA- and probabilistic mapping-based features, and trained an effect prediction model. The model was then evaluated under different scenarios designed to investigate the link and patterns between intra-operative and post-operative testing data.

2. Materials and methods

2.1. Patient and clinical data

2.1.1. Main patient cohort

The study was carried out on a main cohort of 14 patients affected by ET. The patients received DBS targeting in the ventrointermediate nucleus of the thalamus (Vim) for tremor alleviation, at the Department of Neurosurgery of Clermont-Ferrand University Hospital, France between 2008 and 2018. All the patients were implanted bilaterally under local anesthesia with the left brain hemisphere being the first implanted. Two patients were implanted with Abbott SJM 6170 electrodes (diameter 1.27 mm), while the rest were implanted with Medtronic 3389 leads (diameter 1.27 mm). The patients signed a written informed consent for the retrospective analysis of data (Ptolemee Electrophysiologie project: IRB 5921, CE-CIC-GREN-18–03).

2.1.2. Second patient cohort

A second patient cohort affected by a different disorder was included to assess whether the findings from the primary cohort were reproducible and generalizable rather than cohort-specific. This cohort comprised 21 patients with Parkinson's Disease who underwent subthalamic nucleus (STN) DBS at the same clinical institution. Hereafter, this cohort is referred to as the STN cohort, whereas the primary cohort is referred to as the Vim cohort.

2.1.3. Intra-operative stimulation tests

Intra-operative stimulation tests (Table S1) were performed over 14 mm (1 mm step) along 2 parallel trajectories, 2 mm apart, with a microelectrode (Alpha-Omega Engineering, Israel; frequency 130 Hz; pulse width 60 μs; diameter 0.5 mm). The current amplitude was increased from 0.2 mA to 3 mA in steps of 0.2 mA. Symptom improvement with respect to baseline was assessed by the same experienced neurosurgeon (JJL) and neurologist (BD or PD) and categorized according to a 5-point rating scale: 0% (no improvement), 25% (poor improvement), 50% (fair improvement), 75% (good improvement), 100% (complete symptom alleviation) with some intermediate scores if needed. The baseline was determined during the few seconds preceding the stimulation. The minimum stimulation amplitude that produced optimal symptom improvement, as well as the maximum amplitude tolerated before the onset of adverse effects (oculomotor effects, paresthesia, and dysarthria), were documented. This data has also been used in [36–38]. The distributions of intra-operative improvement scores are shown in Fig. 1**A**.

2.1.4. Post-operative stimulation tests

Post-operative tests (Table S1) were performed approximately 3 months after surgery to ensure the absence of lesioning effects due to the procedure. Each electrode contact was individually evaluated by increasing stimulation voltage from 1 V to 5 V in steps of 0.3 V (pulse width 60μs, frequency 130 Hz).

For each combination of contact and voltage, symptom severity was evaluated using a 5-point rating scale ranging from 0 to 100, consistent with the one employed during intra-operative assessments. Also in this case, the symptom baseline was determined before starting the stimulation tests. Occurrence of side effects was also noted with the same categories used intra-operatively. Although patients were assessed across multiple symptoms, this study focused exclusively on tremor for the Vim cohort and rigidity for the STN cohort to enable comparison with intra-operative data. To ensure consistency with the intra-operative dataset, only the largest non-effective VTA was included for each electrode contact, excluding smaller VTAs that also failed to elicit improvement. The distributions of post-operative improvement scores are shown in Fig. 1**B.**

2.2. Probabilistic maps generation

Probabilistic maps were generated separately for the Vim and STN cohorts. The probabilistic mapping procedure is described in [39] and [40] and summarized in Fig. 2. Patient-specific brain tissue conductivity models were generated from MR T1 images with the software ELMA[2] [41]. Conductivity values were set to 2.0 S/m for the cerebrospinal fluid, 0.123 S/m for the grey matter and 0.075 S/m for the white matter considering the pulse width of 60 μs and the pulse frequency of 130 Hz.

[2] https://liu.se/en/article/ne-downloads

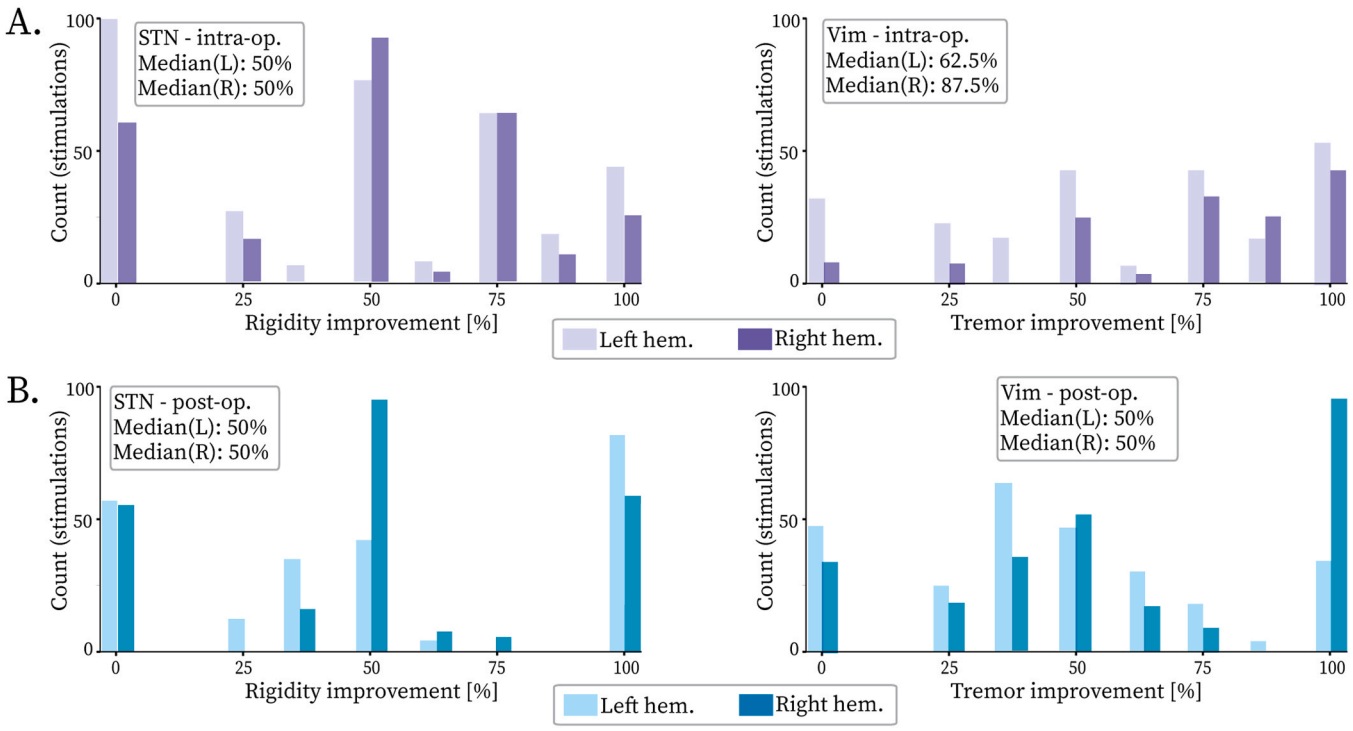

Fig. 1. Improvement scores distribution for intra-operative (A) and post-operative (B) stimulation tests for STN (on the left) and Vim (on the right) cohorts. The scores are further subdivided by brain hemisphere, with the right hemisphere counts in a darker shade and the left hemisphere counts in a lighter shade.

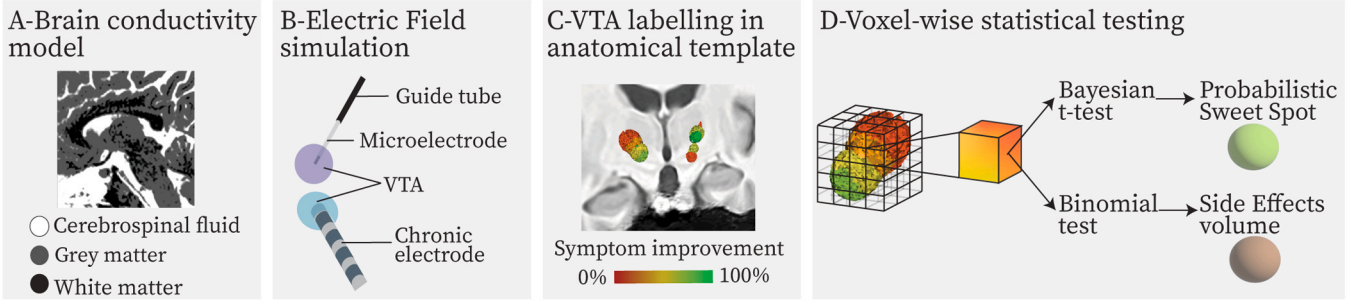

Fig. 2. Graphical representation of the probabilistic mapping workflow. A) Patient-specific brain conductivity model generated by segmentation of the patient's T1 MRI. B) Simulation of the electric field to obtain Volumes of Tissue Activated (VTA). C) Labelling of the VTAs with the associated observed effects and transformation to the cohort-specific anatomical template[47]. D) Statistical testing to determine Sweet Spot or Side Effects volume. Figure modified from [38].

The affine transformation for placement of the intra-operative electrode models in the patient reference space was calculated from the Leksell coordinates in 3D Slicer[3] [42] using the Stereotaxia plugin [43]. Manually annotated post-operative electrode trajectories were obtained after registering the post-operative CT image to the pre-operative T1 MRI in 3D Slicer. The Electric Field (EF) spread corresponding to the tested stimulation parameters was simulated in T1 space using Comsol Multiphysics 5.5 (COMSOL AB, Sweden). VTAs were delineated using an electric field threshold of 0.2 V/mm, which approximates the activation of 3–4 μm myelinated axons. Differences in electrode geometry and stimulation modality were accounted for during the EF simulation [44, 45]. For the intra-operative simulations the guide tubes of the exploration electrodes were set as ground and the active electrode was assigned the current used during the stimulation tests. For the post-operative simulations the electrodes were in monopolar configuration with the active contact acting as voltage source and the outer surface of the simulation domain set to ground. The smallest elements (0.204 mm) of the Finite Element Modelling mesh were located nearby the stimulating contacts in order to capture the strong EF gradients. A more detailed description of the EF simulations can be found in [7,44,46]. The VTAs were then transformed into a cohort-specific anatomical template [9,47] and labelled with their associated DBS effect.

Voxel-wise statistical testing was performed to identify brain regions linked to symptom improvement (Probabilistic Sweet Spot, PSS) or to the occurrence of side effects (SE). A Bayesian *t*-test with normal prior was applied to the VTAs associated with improvement. Voxels with a Bayes Factor (BF)$\geq$ 10 were classified as part of the PSS. A BF$\geq$ 10 indicates strong evidence in favor of the alternative hypothesis [48] H_1: "improvement_voxel$\geq$threshold", meaning that the improvement observed in a given voxel exceeds a predefined threshold. A threshold of 50% was chosen to enable calculation of PSS for both intra-operative and post-operative scenarios, given the medians of the improvement score distributions (Fig. 1). Due to the limited number of tests resulting in side effects, different types of side effects were combined to define a general side effect volume for each cohort. The information about side

[3] https://www.slicer.org/

effects is binary (side effect / no side effect); thus, voxels were tested with a binomial test with False Discovery Rate correction. Voxels with p-value<0.05 were included in the SE volume. Voxels stimulated in fewer than 25% of patients, or less than 10% of the maximum voxel-wise stimulation count, were excluded from both PSS and SE. PSS and SE were computed from intra-operative data only, post-operative data only, or by integrating intra-operative and post-operative data (combined maps).

Furthermore, a stimulation count map (nMap), indicating the frequency with which each voxel was stimulated, and a voxel-wise weighted mean improvement map (wMeanMap) were generated.

2.3. Dataset features and target variable

Twenty-two features were derived from the relationships between the VTA and anatomy (target structure, $n_{features}$=6) and the VTA and probabilistic maps ($n_{features}$=16) (Table 1). A detailed explanation of the features is provided in the Supplementary material – *Section 2*. The features were used to predict the effect caused by the VTA. The subject identifier was included to evaluate the model's capacity for subject-specific learning, particularly in scenarios where intra-operative data from the same individual were available to inform the prediction of post-operative outcomes. Due to the discrete nature of the available DBS effect ratings, the outcome variable (DBS effect) was classified into three categories of clinical interest:

- Class 0: side effects
- Class 1: no or low improvement (<50%)
- Class 2: medium or high improvement (≥50%)

When a stimulation setting simultaneously produced symptom improvement and side effects, precedence was given to the side effect outcome. Thus, such stimulation settings were assigned to the side effect class (Class 0).

2.3.1. Target classes distribution

During intra-operative testing, the most frequent category was class 2 (high improvement), comprising 247 and 377 samples in the Vim and STN samples respectively (Table S2). A comparable distribution was observed in the post-operative assessments, with class 2 again representing the largest proportion of cases (335 for the Vim cohort and 344 for the STN cohort).

Table 1
Features used for DBS outcome prediction. The target structure was the patient-specific target structure segmented by the neurosurgeon on the patient's pre-operative MR image. Relative distances R, S and A are, respectively, the distances in right-left, superior-inferior and anterior-posterior orientation between the centroid of the VTA and the centroid of either the target structure, the PSS or the SE. Abbreviations: VTA=Volume of Tissue Activated; PSS=Probabilistic Sweet Spot; SE=Side Effects; BF=Bayes Factor: Dice=Dice coefficient; Hausdorff=Hausdorff distance; nMap=stimulation frequency map; wMeanMap=weighted mean improvement map.

Category	Features list
Anatomy	VTA-target structure overlap Dice, VTA-target structure centroid distance, VTA-target structure Hausdorff, VTA-target structure relative distance (R, A, S)
Probabilistic Sweet Spot (PSS)	VTA-PSS overlap Dice, mean(BF) in VTA-PSS overlap, mean(BF) in VTA, VTA-PSS centroid distance, VTA-PSS relative distance (R, A, S)
Side Effects (SE)	VTA-SE overlap Dice, VTA-SE centroid distance, VTA-SE relative distance (R, A, S)
Other maps	max(nMap) inside VTA, mean(wMeanMap) in VTA, max (wMeanMap) in VTA, VTA volume
Other	Patient_id, session (intra-operative/post-operative), hemisphere

2.4. Prediction pipeline

The prediction pipeline is illustrated in Fig. 3.

2.4.1. Feature preprocessing

Prior to model training, numerical features were first imputed to handle missing values, if any, using a K-Nearest Neighbors (KNN) imputer with $k = 5$. This approach estimates missing values based on the five nearest neighbors in feature space, preserving the underlying structure of the data. Notably, in our datasets, missing values were observed only for the anatomical features of a single patient in the Vim cohort, for whom the corresponding anatomical label file was unavailable. Subsequently, all numerical features were standardized using z-score normalization (mean=0 and variance=1) to ensure comparable scales across features. Categorical variables were encoded using one-hot encoding [49] to convert them into a numerical format suitable for machine learning. In addition, patient identifiers were encoded numerically to include patient-level information as a feature, accounting for potential intra-patient correlations.

2.4.2. Class balancing

To mitigate group imbalance in the three-class classification task, Synthetic Minority Over-sampling Technique (SMOTE) [50] was applied to the training set. SMOTE generates synthetic examples for minority classes by interpolating existing samples in feature space, allowing the classifier to better learn decision boundaries for underrepresented groups.

2.4.3. Prediction model

A Gradient Boosted Decision Tree classifier (XGBoost) [51] was trained on the computed and preprocessed features. XGBoost was selected due to its robustness to heterogeneous feature distributions, its ability to model non-linear relationships, and suitability for multi-class classification problems. The classifier was trained using the multi-class logarithmic loss (mlogloss) as the objective function. After training, the model generated predicted probabilities for each class. Class predictions were obtained by selecting the class with the highest predicted probability.

The prediction was performed separately on the Vim and STN cohorts, due to the difference in evaluated symptoms (tremor for Vim, rigidity for STN).

The analysis was conducted in Python 3.12 on a 32-core AMD 2990WX 3.80 GHz workstation with 128 GB of RAM. The gradient boosting model was implemented using the XGBoost library, and additional machine learning tools, including data preprocessing, model evaluation, and feature importance analysis, were performed using scikit-learn (sklearn[4]).

2.5. Model validation and evaluation

2.5.1. Validation scenarios

To assess whether incorporating intra-operative data could improve predictions of post-operative outcomes, model performance was evaluated across seven scenarios, each addressing a specific question. The scenarios are graphically represented and summarized in Table 2.

- Scenario 1: Prediction performance of intra-operative data on intra-operative data: how good is the prediction performance on intra-operative data?
- Scenario 2: Prediction performance of post-operative data on post-operative data: how good is the prediction performance on post-operative data?

[4] https://scikit-learn.org/stable/

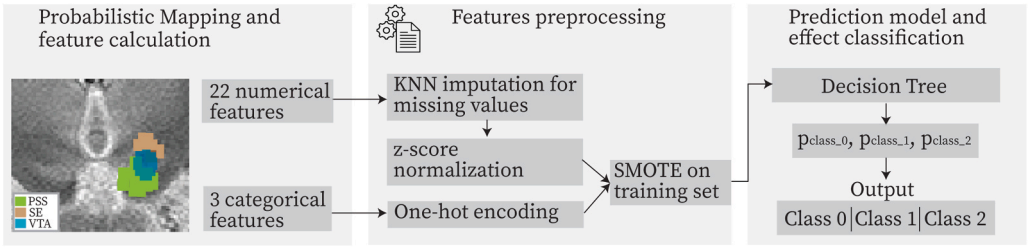

Fig. 3. DBS effects prediction pipeline. After defining Probabilistic Sweet Spots and Side Effect areas with the probabilistic mapping workflow, features are calculated for each VTA. Numerical features undergo K-Nearest Neighbors (KNN) imputation and z-score normalization while categorical features are one-hot encoded. The training dataset is oversampled with the Synthetic Minority Over-sampling Technique (SMOTE) to mitigate class imbalance. The features are fed into a Gradient Boosted Decision Tree classifier (XGBoost) model that outputs a probability (p) for each of the 3 output classes. The predicted class is the one with the highest probability.

Table 2
Summary of usage of intra-operative and post-operative VTAs and probabilistic maps in the seven validation scenarios. T: training; V: validation; -: element not used. Circles with thick borders: VTAs; circles without border: probabilistic maps. Intra-operative: purple; post-operative: blue. The maps with a color fading from blue to purple in Scenarios 6 and 7 indicate the combined maps calculated by integrating intra-operative and post-operative data. Scenarios requiring the usage of a leave-one-out strategy to avoid data leakage are marked with the people group symbol.

Scenario	Intra-op. VTAs	Intra-op. maps	Post-op. VTAs	Post-op. maps	Graphical representation (Intra-op. Post-op.)
1	T, V	T, V	-	-	Training Validation
2	-	-	T, V	T, V	Training Validation
3	T	T, V	V	-	Training Validation
4	-	T, V	T, V	-	Training Validation
5	T	T	V	V	Training Validation
6	T	T, V	T, V	T, V	Training Validation
7	-	T, V	T, V	T, V	Training Validation

- Scenario 3: Generalization from intra-operative to post-operative outcomes: can intra-operative tests and probabilistic maps generalize to predict post-operative outcomes?
- Scenario 4: Predictive utility of intra-operative probabilistic maps: can probabilistic maps derived from intra-operative data reliably predict post-operative outcomes?
- Scenario 5: Transferability of learned relationships across contexts: are the relationships between VTAs/maps and outcomes learned intra-operatively transferable to the post-operative context?
- Scenario 6: Integration of intra-operative and post-operative data: does combining intra-operative and post-operative data for training improve the prediction performance?
- Scenario 7: Predictive utility of combined probabilistic maps: can probabilistic maps derived from both intra- and post-operative data be used to predict post-operative outcomes?

For leave-one-out validation scenarios, probabilistic maps were recomputed n times, where n is the number of patients in the cohort, to prevent data leakage between the training and validation sets. In each fold, all stimulation recordings belonging to the held-out patient were excluded from both probabilistic map generation and predictive model training, ensuring that the validation was performed on a completely unseen patient. Scenarios 1 and 2 served as reference benchmarks, as the model is expected to perform best on uniform datasets.

2.5.2. *Model evaluation*

Model performance was assessed using:

- Balanced accuracy (BA) (Eq. 1) which measures how well a model correctly classifies classes, range [0, 1]:

$$BA = \frac{1}{C}\sum_{i=1}^{C}\frac{TP_i}{TP_i + FN_i} \quad (1)$$

- Precision (Eq. 2) which indicates how many predicted positives are true positives, range [0, 1]:

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

- Recall (Eq. 3) which indicates how many of the actual positive cases the model correctly identifies, range [0, 1]:

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

Where C: number of classes (i.e. 3 in this case), TP: true positives, FP: false positives and FN: false negatives.

Precision and recall were first calculated independently for each class using a one-vs-rest strategy. In this setting, each class is treated as the positive class while the remaining classes are treated as negative. To obtain a single performance measure across all classes, we used weighted averaging, where each class-specific metric is weighted by the number of true instances of that class, obtaining weighted precision and weighted recall.

The proportions of correctly and incorrectly classified samples for each class were presented using confusion matrices. Pairwise comparisons of the metrics values between scenarios were performed using the Wilcoxon signed-rank test. Since multiple pairwise comparisons were conducted across the seven scenarios, p-values were adjusted using the Holm correction [52] to control the family-wise error rate.

2.6. Feature analysis

A feature analysis was conducted to examine potential differences in feature contributions between intra-operative and post-operative contexts. Permutation importance and SHAP (SHapley Additive exPlanations) values were employed for this purpose. In particular, permutation importance was applied to rank the features and to examine whether the most influential features were consistent between the intra-operative and post-operative contexts. Feature importance values quantify the relative contribution of each feature to the model's predictions. These values are often scaled between 0 and 1 such that the most influential feature attains the highest value. In practice, the ranking of features provides more meaningful information than the absolute magnitude of the importance values. Features exhibiting very low importance (e.g., below 0.01) are generally considered negligible and may be candidates for exclusion from the model [53,54]. SHAP values were then analyzed to determine the contribution of each feature to class prediction and to assess whether their effects were consistent across the two DBS phases.

3. Results

To maintain clarity and conciseness, the main text primarily focuses on results from the Vim cohort. This cohort was used for model development and optimization, including the evaluation of model configurations and input features, whereas the STN cohort was included as an additional cohort to investigate whether the identified relationships and modelling approach extended to a different DBS target and disease context. Additional STN analyses are reported in the Supplementary Material, with key findings summarized in the main text.

3.1. Model performances

Model performance on the Vim cohort is summarized in Fig. 4 and Table 3. The highest balanced accuracy (BA=0.56) was obtained in Scenario 1, which relied exclusively on intra-operative data. This scenario was also associated with the highest precision (0.63) and recall (0.55). Slightly lower performance was observed in Scenario 2 (BA=0.53), where only post-operative data were used; in this case, classification performance for Class 2 was notably poor. Precision (0.56) and recall (0.53) were reduced with respect to Scenario 1. Comparable results to Scenario 2 were achieved in Scenarios 4 and 7 (BA=0.50 and 0.52, respectively), both of which employed intra-operative or combined maps but relied on post-operative VTAs for training and validation. Notably, using only intra-operative maps in Scenario 4 improved the classification of Class 2, however at the expense of reduced accuracy for Class 0.

The lowest performance was observed in Scenarios 3 and 6 (BA 0.39 and 0.43, respectively), which partially or exclusively used intra-operative data for training and post-operative data for testing. These scenarios were characterized by particularly low recall (0.27 and 0.32) indicating limited sensitivity. Scenario 5, which also relied solely on intra-operative data during training, achieved a higher BA of 0.51 with intermediate precision (0.52) and recall (0.44); however, its confusion matrices revealed misclassification patterns similar to those in Scenarios 3 and 6, with predictions predominantly assigned to Classes 0 and 2.

The patterns observed in the model performance across the six scenarios were largely comparable between the Vim and STN (Table 3, Fig. 5) cohorts.

Metrics distributions across the different validation scenarios are shown in Fig. 6. For the Vim cohort, pairwise statistical comparisons (Table 4) demonstrated that Scenario 3 performed significantly worse than Scenarios 1, 2, 4, and 7 in terms of balanced accuracy, precision, and recall, while Scenario 6 showed significantly lower balanced accuracy than Scenarios 1, 2, and 7. In contrast, no statistically significant differences were observed among Scenarios 1, 2, 4, 5, and 7 for balanced accuracy, indicating comparable predictive performance within this group. Similar trends were observed for the STN cohort, with Scenario 1 consistently providing the best performance and significantly outperforming most other validation scenarios, whereas the lowest performances were observed in Scenarios 3, 5, and 6.

Intra-operative and post-operative scenarios also exhibited slight differences in the locations of the hotspots within the nMap (representing areas of high VTA density) and the wMeanMap (representing areas associated with maximal improvement), when examined in their anatomical context (Figures S1 to S4). Examples of PSS and SE obtained in the same leave-one-out iteration are presented in Figure S5.

3.2. Feature analysis

Feature importance analysis showed that VTA volumes and their overlaps with probabilistic maps of therapeutic and adverse effects, as well as with the target structure, were the most influential predictors. Their relative contributions, however, varied across different contexts.

3.2.1. Feature ranking with permutation importance

In the Vim cohort (Table 5) VTA volume emerged as the most influential feature in both intra-operative and post-operative contexts with importance ~0.25. In intra-operative data, spatial relationships such as VTA-PSS centroid distance (importance = 0.075) and VTA-target structure Hausdorff distance (importance = 0.038) also contributed moderately. In post-operative data, VTA-PSS overlap Dice (importance = 0.14) emerged as the second most important feature, followed by other overlap- and distance-based measures.

Also in the STN cohort (Table S3), VTA volume was the most influential feature in both intra- and post-operative context. Intra-operatively, model predictions were primarily driven by target-

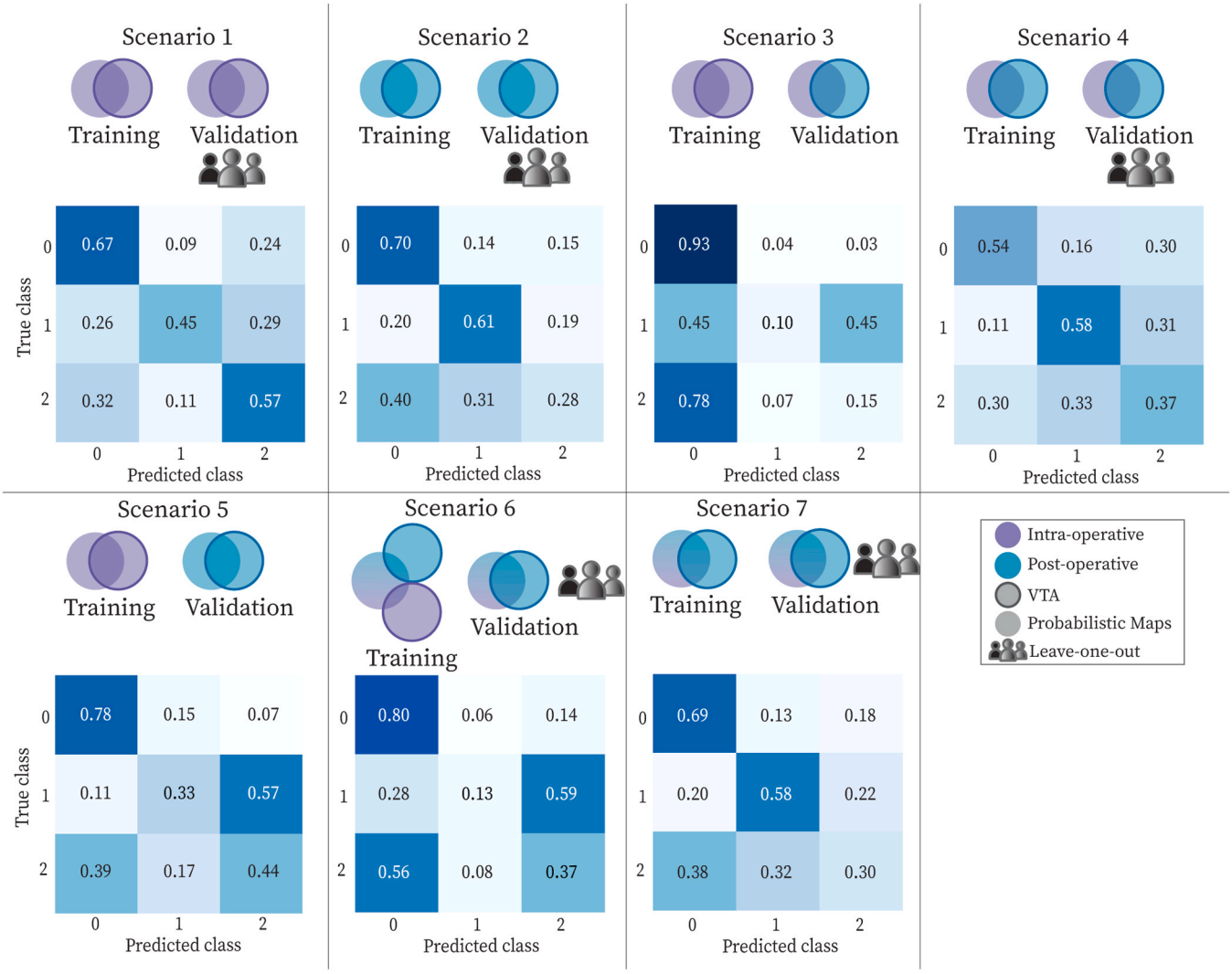

Fig. 4. Model performance in the different validation scenarios. The confusion matrices show the ratios of classified and misclassified samples for each class for scenarios 1–7 in the Vim cohort.

Table 3
Balanced accuracy (BA), precision and recall scores for each scenario for the Vim and STN cohorts. For each metric the first line shows the mean value and the second line the lower and upper values of the 95% confidence interval (CI).

Scenario	*1*	*2*	*3*	*4*	*5*	*6*	*7*
Vim							
BA	0.56	0.53	0.39	0.50	0.51	0.43	0.52
BA 95% CI	0.47–0.62	0.46–0.60	0.37–0.42	0.45–0.56	0.45–0.57	0.39–0.47	0.46–0.59
Precision	0.63	0.56	0.43	0.54	0.52	0.48	0.55
Precision 95% CI	0.54–0.73	0.47–0.68	0.32–0.58	0.46–0.66	0.41–0.65	0.39–0.61	0.47–0.66
Recall	0.55	0.53	0.27	0.51	0.44	0.32	0.51
Recall 95% CI	0.48–0.63	0.47–0.58	0.21–0.32	0.46–0.56	0.38–0.51	0.25–0.39	0.46–0.57
STN							
BA	0.68	0.47	0.35	0.47	0.39	0.39	0.43
BA 95% CI	0.62–0.77	0.43–0.52	0.28–0.40	0.43–0.52	0.33–0.46	0.35–0.43	0.39–0.48
Precision	0.75	0.54	0.38	0.55	0.41	0.46	0.51
Precision 95% CI	0.69–0.82	0.45–0.64	0.29–0.49	0.47–0.64	0.28–0.56	0.35–0.59	0.41–0.61
Recall	0.73	0.53	0.36	0.54	0.31	0.31	0.50
Recall 95% CI	0.67–0.80	0.46–0.60	0.31–0.41	0.48–0.61	0.26–0.36	0.26–0.37	0.43–0.58

structure metrics, including VTA–target centroid (importance = 0.046) and Hausdorff distances (importance = 0.029), followed by PSS- and SE-based features. Post-operatively, this pattern was reversed, with PSS- and SE-related metrics contributing more strongly to prediction.

The feature contributions to the prediction of target classes highlighted by SHAP values are reported in Figure S6 and Figure S7, and the main results are summarized in Table 6.

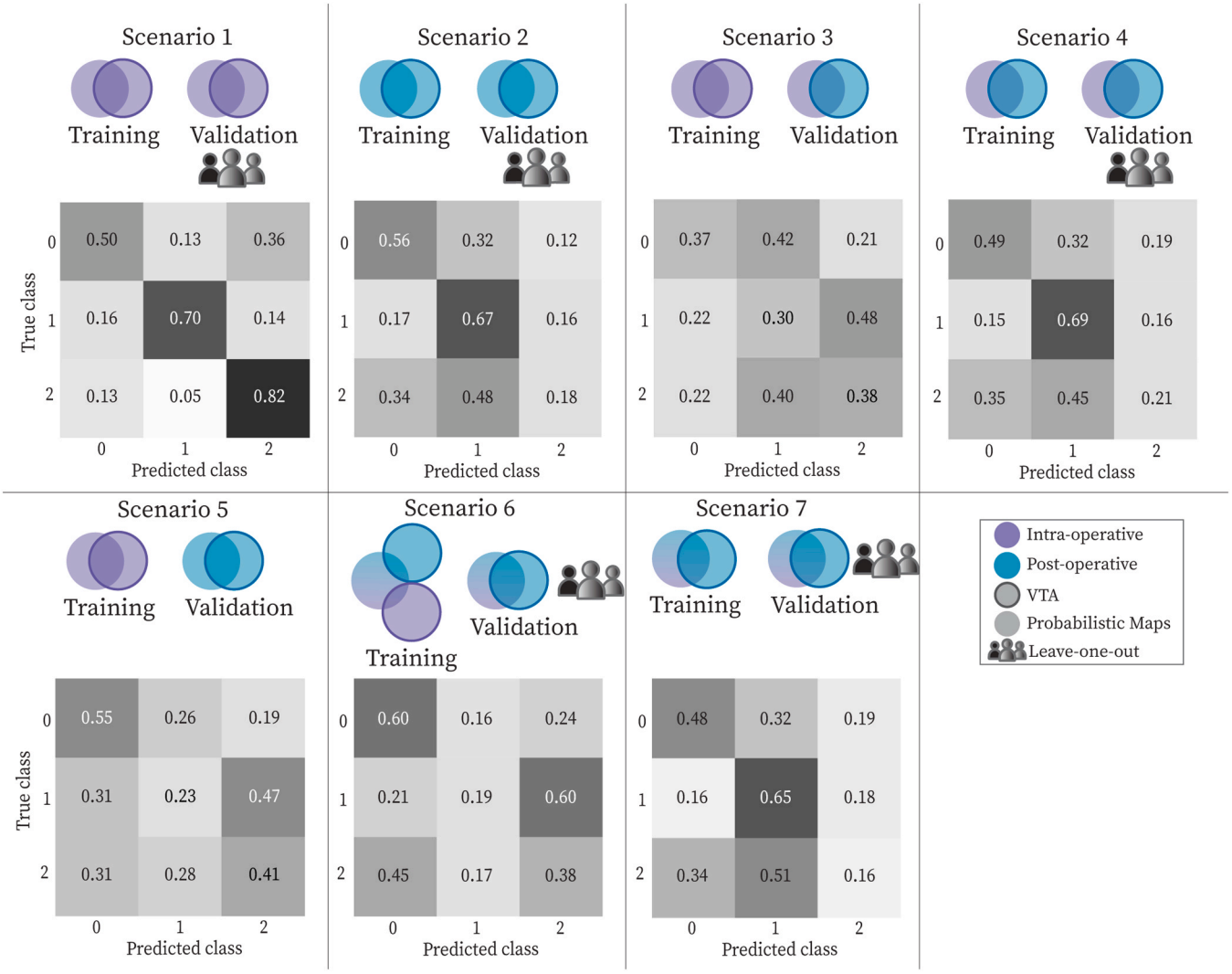

Fig. 5. Model performance in the different validation scenarios. The confusion matrices show the ratios of classified and misclassified samples for each class for scenarios 1–7 in the STN cohort.

4. Discussion

In this study, we examined the extent to which post-operative DBS effects can be predicted using intra-operative data alone or in combination with post-operative information. Clarifying the predictive value of intra-operative data has important clinical implications, as it may enable earlier estimation of stimulation outcomes and provide guidance for surgical workflow decisions. Our findings indicate that intra-operative data contribute meaningfully to probabilistic mapping; however, the relationship between VTAs and maps differs between the intra- and post-operative settings, underscoring the necessity of including post-operative data in the training of predictive models. VTA volumes and their spatial relationships with probabilistic maps of therapeutic and adverse effects were the most influential predictors.

4.1. Within-phase prediction Scenarios

Overall, model performance was higher in the intra-operative setting compared to the post-operative one, with the largest performance decline observed in the STN cohort. Post-operative predictions of high improvement were less accurate despite the class having the most samples, suggesting limited discriminative power of the features. Conversely, in the intra-operative context, side effects were more difficult to predict. This may be explained by both the smaller sample size of this class and by the fact that certain stimulations can simultaneously induce clinical improvement and adverse effects. The balanced accuracy of approximately 60% and 70% observed on intra-operative data (Scenario 1), in the Vim and STN cohorts respectively, is broadly consistent with the performance reported in previous studies on stimulation effect prediction [13,20,28]. However, direct comparisons should be interpreted with caution because these studies addressed different prediction tasks. In particular, they focused on binary classification of symptom improvement or the presence of a stimulation effect, whereas our framework considered three outcome classes, explicitly distinguishing therapeutic effects, limited or absent improvement, and stimulation-induced side effects. This represents a more challenging classification problem and therefore performance metrics are not directly comparable across studies.

4.2. Mixed-phase prediction Scenarios

Results from Scenarios 4 and 7 show that intra-operative stimulation maps yield predictive performance comparable to that of post-operative maps, when both training and testing are performed on post-operative VTAs. Interestingly, using only intra-operative maps helped to balance performance across the 3 classes for the Vim cohort. In contrast, training only on intra-operative VTAs proved ineffective, as shown in Scenarios 3 and 5: the model was unable to generate reliable predictions, and the

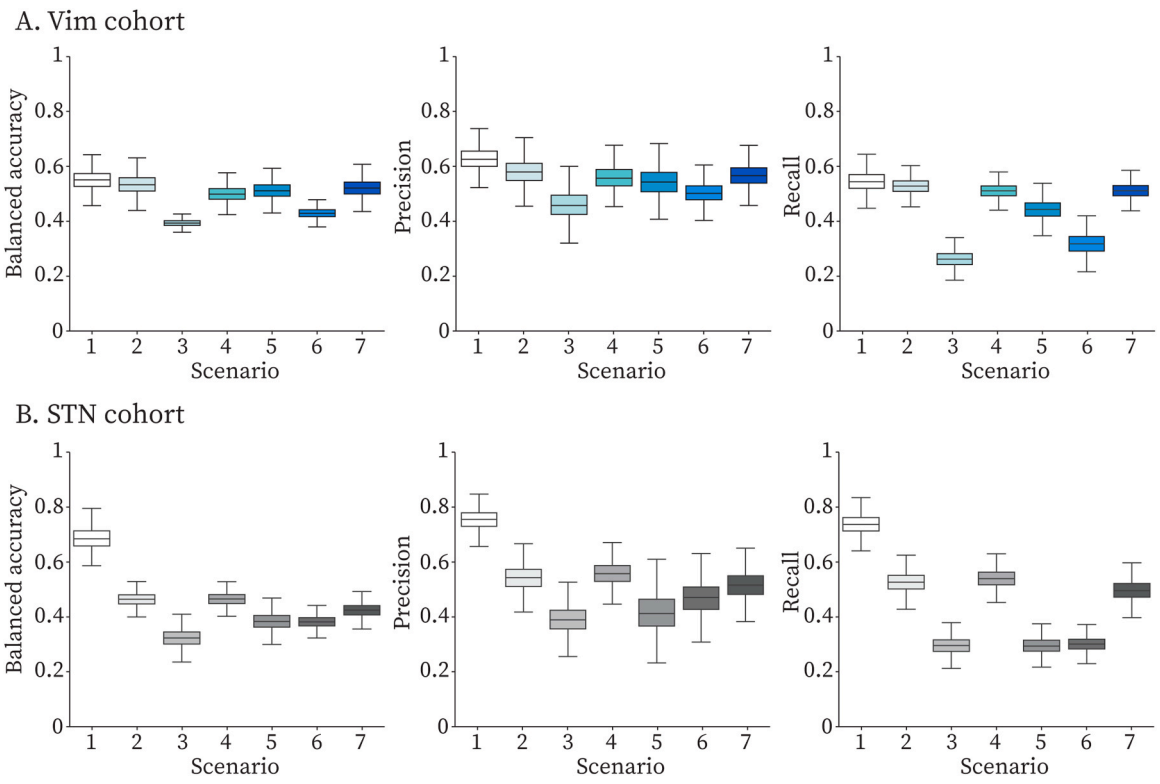

Fig. 6. Distributions of balanced accuracy, precision, and recall. The distributions are shown for the seven validation scenarios for the Vim (A, top row) and STN (B, bottom row) cohorts. The boxplots indicate the median and the interquartile range.

Table 4
Pairwise statistical comparison of predictive performance between the seven validation scenarios. Statistical significance was assessed using paired Wilcoxon signed-rank tests on balanced accuracy (BA), precision, and recall, with p-values adjusted for multiple comparisons using the Holm correction. Significant differences after correction are indicated by * ($p<0.05$) and ** ($p<0.01$). The dash indicates the cells belonging to the diagonal of the matrix.

	BA							Precision							Recall						
	1	2	3	4	5	6	7	1	2	3	4	5	6	7	1	2	3	4	5	6	7
Vim																					
1	-		*			*		-		*					-		*				
2		-	**			*			-	*						-	*				
3	*	**	-	*			**	*	*	-	*			*	*	*	-	*	*		*
4			*	-						*	-						*	-			
5					-							-					*		-		
6	*	*				-	*						-							-	
7			**			*	-			*				-			*				-
STN																					
1	-	*	**	**	**	**	**	-	*	**	*	**	**	**	-	**	**	**	**	**	**
2	*	-	**		**	**	*	*	-	**		**	*		**	-	**		**	**	
3	**	**	-	**		*	**	**	**	-	*			*	**	**	-	**			
4	**		**	-	*	*		*		*	-	*			**		**	-	**	**	
5	**	**		*	-		*	**	**		*	-		*	**	**		**	-		
6	**	**	*	*		-	*	**	*				-		**	**		**		-	
7	**	*	**		*	*	-	**		*		*		-	**						-

inclusion of intra-operative VTAs in combined models (Scenario 6) appeared to introduce noise. These findings suggest that the relationship between VTAs, maps, and clinical outcomes differs between the intra- and post-operative contexts, a notion further supported by the observation that intra-operative and post-operative nMap and wMeanMap do not perfectly align when examined within the same anatomical slices. Importantly, these results were robust to the choice of the VTA activation threshold. A sensitivity analysis using electric field thresholds of 0.15 V/mm and 0.25 V/mm, in addition to the nominal threshold of 0.2 V/mm, yielded the same qualitative patterns across Scenarios 3–5 (Figures S8, S9). The limited transferability of models trained on intra-operative VTAs and the improved performance obtained when training on post-operative VTAs were consistently observed across thresholds, indicating that they were not driven by the specific threshold used for VTA generation. Discrepancies between intra-operative and post-operative scenarios can be due to differences in electrode characteristics and stimulation modality. In particular, the substantially different electrode diameters used during intra-operative (0.5 mm) and post-operative (1.27 mm) stimulation affect the spatial extent and shape of the elicited electric fields and, consequently, the resulting VTA simulations. In addition, intra-operative stimulation was performed under current-controlled conditions, whereas post-operative stimulation used voltage-controlled settings. These stimulation modes define the electric field in different ways: current-controlled stimulation specifies the

Table 5
Feature importance analysis results for the Vim cohort. The table shows the first 5 most important features (higher value implies higher influence of the feature on the prediction) for the model when applied to only intra-operative (Scenario 1) or only post-operative (Scenario 2) stimulation tests.

Intra-operative		Post-operative	
Feature	Importance	Feature	Importance
VTA volume	0.26	VTA volume	0.25
VTA-PSS centroid distance	0.075	VTA-PSS overlap Dice	0.14
mean(BF) in VTA-PSS overlap	0.062	VTA-SE relative distance A	0.026
VTA-target structure Hausdorff	0.038	VTA-SE overlap Dice	0.014
VTA-SE overlap Dice	0.010	VTA-target structure Hausdorff	0.011

injected current, while voltage-controlled stimulation specifies the applied voltage, with the resulting current depending on the electrode–tissue interface and impedance [44]. However, differences in electrode geometry and stimulation modality were explicitly taken into account in the modeling and simulation framework to ensure that the resulting VTAs accurately reflected the corresponding stimulation conditions. In addition to these technical differences, the two phases also differ in the clinical context of outcome assessment. Intra-operative testing evaluates acute stimulation effects over short time intervals in an awake patient. These measurements may be influenced by transient effects caused by electrode insertion, including the microlesional effect, where mechanical disruption of the tissue during electrode advancement can temporarily modify the clinical symptoms independently of stimulation [55,56]. Moreover, despite ensuring resolution of observable or reported side effects before successive stimulation trials and assessing symptoms prior to stimulation, repeated stimulation may still introduce variability related to incomplete recovery, patient fatigue, and stress. In contrast, post-operative assessments performed several months after implantation reflect a more stable physiological state, after resolution of acute insertion-related effects. Thus, the relationship between VTA properties and clinical outcome may differ between phases because intra-operative and post-operative assessments capture partially different physiological conditions. Taken together, these results suggest that training on post-operative VTAs is essential for model-based outcome prediction. Nonetheless, intra-operative data remain highly valuable for the generation of probabilistic maps. These could enable the construction of reliable representations of therapeutic and adverse effects volumes prior to the first post-operative programming session.

4.3. Feature analysis

Probabilistic mapping features and VTA volume emerged in general as the primary drivers of model predictions. However, feature relevance differed between intra-operative and post-operative contexts and between the Vim and STN cohorts. The contribution of VTA volume likely reflects both physiological and methodological factors. From a physiological perspective, VTA volume represents the spatial extent of tissue activation and may therefore provide relevant information about the likelihood of engaging therapeutic or non-therapeutic regions. However, its importance may also be influenced by the data collection procedure, particularly in the intra-operative setting, where stimulation amplitudes are often increased until a clinical effect is observed and testing may preferentially focus on amplitudes close to the therapeutic threshold. This introduces a coupling between stimulation efficacy and VTA size, potentially amplifying the predictive contribution of VTA-related features and limiting the generalizability of models trained on such data. Future studies could reduce this bias by systematically sampling stimulation responses across a broader range of clinically relevant stimulation amplitudes and by using quantitative methods to assess stimulation effects (e.g., objective measurements of tremor or rigidity [57]). Such approaches may provide a more detailed characterization of the relationship between VTA properties and clinical outcomes. In the Vim cohort, intra-operative outcomes were more influenced by spatial distances, while post-operative outcomes primarily relied on overlap measures. In contrast, the STN cohort shifted from target-structure-driven intra-operative predictions to PSS/SE-driven post-operative predictions. Intra-operative models showed clearer and more distinct feature–class relationships, whereas post-operative models exhibited weaker and more ambiguous associations reflecting reduced predictive performance. These results further confirm the observations from the model performance analyses: the relationships between features and outcomes differ between intra-operative and post-operative contexts, likely explaining why models trained on intra-operative VTAs fail to reliably predict post-operative outcomes.

Table 6
Summary of the relationships between features and predictions of class 0, 1 and 2 for Vim and STN cohort and intra-operative and post-operative stimulation tests as observed in the SHAP plots. The table shows whether the prediction of a certain class is associated with a high (orange) or low (light blue) value of the feature. White spaces indicate that a precise contribution of the feature was not visible in the SHAP plots. Abbreviations: op.=operative; H=high; L=low.

Features Feature value associated with class prediction (Low / Unclear / High)	**Vim**						**STN**					
Phase	**Intra-op.**			**Post-op.**			**Intra-op.**			**Post-op.**		
Class	0	1	2	0	1	2	0	1	2	0	1	2
VTA volume	H	H	L	H	L		H	H	L	H	L	
VTA-PSS centroid distance	H		L	H			H	L			L	
Mean(BF) in VTA-PSS overlap	H	L	H		L	H	L	L	H		L	
VTA-PSS overlap Dice	H	L	H		L	H	H	L	H	L	L	H
VTA-SE overlap Dice	H	L	L	H	L		H	L	L	H	L	H
VTA-SE relative distance A	L	H	H	L	H		L	H	L	L	H	L
VTA-target Hausdorff	H	H	L	H	H	L	L		L	L	L	H
VTA-target centroid distance	L	L					L	H	L	H	L	
VTA-target overlap Dice	H				H		H	L	L	L	L	H
VTA-target relative R.	H		L		H		H	L	H		L	H

4.4. Clinical implications and translational considerations

The predictive performances obtained in this study should be interpreted in the context of their intended application. No established threshold exists for the clinical utility of machine learning models in DBS, and the required level of performance depends on whether the model is intended to support or replace clinical decision-making. The present models are not sufficiently accurate to be used as standalone tools for surgical targeting or DBS programming. Rather, they should be viewed as decision-support systems that provide complementary information to clinicians. In this context, predictive performance exceeding chance level (0.33 accuracy in a 3-class classification task) may still be valuable, particularly when integrated with clinical expertise. However, clinical validation in a specific application will be required to establish its clinical utility.

An additional consideration is that classification performance depends on the threshold used to convert clinical ratings into discrete outcome classes. In the present study, we repeated the analysis using two alternative thresholds (25% and 75% improvement), and the relative ranking of the seven validation scenarios remained unchanged (Figures S10 to S13). These findings indicate that the observed performance patterns are robust to reasonable variations in class definition, although the choice of threshold remains an additional source of variability in classification-based DBS prediction models. In future applications, class boundaries could be further refined based as additional clinical evidence becomes available.

From a clinical perspective, Scenarios 3 and 5 represent the most relevant validation settings because they evaluate whether intra-operative observations can predict post-operative stimulation outcomes. One potential motivation for this approach is to reduce the dependence on extensive post-operative clinical testing by exploiting information collected during surgery. However, the reduced performance observed in these scenarios suggests that intra-operative stimulation responses alone cannot be directly translated into chronic stimulation effects. Therefore, post-operative stimulation testing, such as monopolar review, remains important for informing patient-specific predictive models. Nevertheless, as larger datasets become available, a promising future direction would be to pre-train the model on a large cohort and subsequently fine-tune it using a limited number of post-operative stimulation tests from a new patient. Such approach could reduce the amount of individualized clinical testing required while preserving patient-specific adaptation. In the present study, we also investigated whether patient-specific information substantially contributed to prediction performance by repeating the analyses without patient identifiers as input features. Since the resulting performances remained comparable (Figures S14, S15), the models appear to capture generalizable relationships between stimulation features and clinical outcomes rather than relying on patient-specific memorization.

The findings emerging from Scenarios 3 and 5 also have implications for the ongoing discussion regarding awake versus asleep DBS surgery. One of the advantages of awake DBS is the possibility of performing intra-operative clinical testing, albeit at the cost of longer surgical procedures and increased patient burden. Our results indicate that, within the proposed VTA-based framework, intra-operative stimulation responses alone do not provide sufficient information to reliably predict chronic post-operative outcomes. Although this does not diminish the value of intra-operative testing for target verification and the identification of stimulation-induced side effects, it suggests that its contribution to predicting long-term stimulation efficacy may be more limited than previously assumed.

Overall, this study demonstrates that VTA-based machine learning models can capture clinically relevant information within homogeneous stimulation contexts, while also highlighting the challenges associated with transferring these models across different surgical phases. Ultimately, the clinical value of such models is unlikely to rely on fully automated prediction, but rather on their ability to integrate complex stimulation-response relationships and provide quantitative support during DBS programming. Future studies with larger, multicenter datasets and prospective validation will be required to determine how these models can be incorporated into clinical workflows.

4.5. Strengths and limitations

Given the highly patient-specific nature of DBS effects, a limitation of the present study is the relatively small number of patients. However, this is partially compensated for by the large number of stimulations per patient. Although each patient contributed a large number of stimulation settings, these observations are not statistically independent because they originate from the same individual. To account for this dependency, all analyses were validated using leave-one-patient-out cross-validation, ensuring that all stimulation settings from the validation patient were excluded from both probabilistic map generation and predictive model training. Consequently, the reported performance reflects generalizability to previously unseen patients. Nevertheless, the effective sample size is driven primarily by the number of patients rather than by the total number of stimulation settings, and larger patient cohorts will be important to further improve the robustness and generalizability of these models. Symptom evaluation was subjective, even though it was consistently performed by the same experienced clinician, which reduces inter-rater variability [58]. Nevertheless, clinical ratings remain affected by measurement uncertainty. To quantify the potential impact of this variability on the available labels, we estimated the within-subject standard error of measurement (SEM) (see Supplementary material – *Section 3*). The resulting SEM values indicate residual variability of approximately 15–27 percentage points across cohorts and phases, suggesting that neighboring improvement categories cannot always be distinguished with high precision [59]. Notably, the SEM for the post-operative STN cohort was 26.6 percentage points, exceeding the 25-point width of a single rating category, indicating that adjacent categories in this cohort and DBS phase may be particularly difficult to distinguish from measurement noise alone. Therefore, label uncertainty represents an inherent limitation of the present classification task and may contribute to an upper bound on the predictive performance achievable by any model trained on these outcomes.

Additional limitations stem from the features used for prediction. Connectivity data, which have been shown to enhance predictive performance in previous studies [26] were not incorporated and could help explain currently unexplained variance. Similarly, clinical and demographic variables, such as age and gender, have been reported to improve predictions or even inform models based solely on these features [13,14,16,19,22,30,31]. In a previous study [36], we included these variables, but they did not contribute meaningfully to the model's predictions. This may reflect the relative homogeneity of our cohort; such features could prove more informative in larger and more diverse patient populations, where inter-individual variability is greater. The study is further limited by the treatment of side effects, which were not categorized individually due to the small number of records per category. Instead, a general side effect region was defined, capturing clinically relevant adverse outcomes but likely obscuring distinctions between effects arising from stimulation of different anatomical regions. A key strength of the study is the availability of both intra-operative and post-operative data for the same patients. Although another study [60] also included both types of data, their focus was on predicting the minimum amplitude threshold required to elicit effects across phases, and their models were not based on VTAs or PSMs. The inclusion of VTAs and probabilistic maps in our approach provides unique insights into how stimulation parameters relate to therapeutic effects across different evaluation contexts. However, differences between intra-operative and post-operative assessments should also be considered when interpreting cross-phase comparisons. Intra-operative testing is performed in an acute setting and may be affected by transient factors such as the microlesional effect, tissue perturbation during mapping,

and residual effects from previous stimulation trials. In addition, differences in electrode configuration and stimulation modality may influence the resulting VTAs. These factors may contribute to differences in the VTA–clinical outcome relationship between phases. Nevertheless, the availability of paired intra-operative and post-operative data from the same patients represents a unique opportunity to investigate the generalizability of VTA-based models and to better understand the factors limiting their transfer across clinical settings.

5. Conclusions

In this study, we evaluated the potential of predicting post-operative DBS effects using intra-operative data alone or in combination with post-operative information. Our findings highlight that predictive modeling of DBS outcomes benefits from both intra- and post-operative data. Intra-operative data are informative for identifying regions associated with optimal effects when stimulated. These regions are robust and can inform DBS parameter selection in both intra-operative and post-operative contexts. Conversely, the relationships between VTAs, probabilistic maps, and clinical outcomes differ between intra- and post-operative settings. Therefore, post-operative screening data remain essential for training models intended for automated DBS programming and cannot be fully substituted by intra-operative testing alone. Taken together, this suggests that intra-operative data are particularly valuable for elucidating mechanisms of action, whereas post-operative data provide the most direct guidance for programming and clinical optimization. Feature analysis revealed that probabilistic mapping-related measures play a major role in driving predictions and should be incorporated into effect prediction models. Yet, their contributions vary between intra- and post-operative scenarios, further emphasizing the necessity of including post-operative data in DBS effect prediction models.

Declaration of Competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was financially supported by the Swiss National Science Foundation (Grant No. 205320–207491) and the Parkinson Research Foundation at Linköping University.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jdbs.2026.09.003.

References

[1] Wårdell K, et al. Deep brain stimulation: emerging tools for simulation, data analysis, and visualization. Front Neurosci 2022;16:834026. https://doi.org/10.3389/fnins.2022.834026.
[2] Chiken S, Nambu A. Mechanism of deep brain stimulation. Neuroscientist 2016;22(3):313–22. https://doi.org/10.1177/1073858415581986.
[3] Lozano AM, et al. Deep brain stimulation: current challenges and future directions. Nat Rev Neurol 2019;15(3). https://doi.org/10.1038/s41582-018-0128-2.
[4] Brinke TR, Odekerken VJJ, Dijk JM, van den Munckhof P, Schuurman PR, de Bie RMA. Directional deep brain stimulation: first experiences in centers across the globe. Brain Stimul 2018;11(4):949–50. https://doi.org/10.1016/j.brs.2018.04.008.
[5] Nguyen TAK, et al. Directional stimulation of subthalamic nucleus sweet spot predicts clinical efficacy: proof of concept. Brain Stimul 2019;12(5):1127–34. https://doi.org/10.1016/j.brs.2019.05.001.
[6] Coblentz A, et al. Mapping efficacious deep brain stimulation for pediatric dystonia. J Neurosurg Pediatr 2021;27(3):346–56. https://doi.org/10.3171/2020.7.PEDS20322.
[7] Åström M, Zrinzo LU, Tisch S, Tripoliti E, Hariz MI, Wårdell K. Method for patient-specific finite element modeling and simulation of deep brain stimulation. Med Biol Eng Comput 2009;47(1):21–8. https://doi.org/10.1007/s11517-008-0411-2.
[8] Fonnov V, Evans AC, McKinstry RC, Almli CR, Collins DL. Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage 2009;47:S102. https://doi.org/10.1016/S1053-8119(09)70884-5.
[9] Vogel D, Wårdell K, Coste J, Lemaire J-J, Hemm S. Atlas Optimization for Deep Brain Stimulation. In: Jarm T, Cvetkoska A, Mahnič-Kalamiza S, Miklavcic D, editors. 8th European Medical and Biological Engineering Conference. Cham: Springer International Publishing; 2021. p. 130–42. https://doi.org/10.1007/978-3-030-64610-3_16.
[10] Lemaire J-J, et al. An MRI deep brain adult template with an advanced atlas-based tool for diffusion tensor imaging analysis. Sci Data 2024;11(1):1189. https://doi.org/10.1038/s41597-024-04053-x.
[11] Eisenstein SA, et al. Functional anatomy of subthalamic nucleus stimulation in Parkinson disease: STN DBS Location and PD. Ann Neurol 2014;76(2):279–95. https://doi.org/10.1002/ana.24204.
[12] Boutet A, et al. Predicting optimal deep brain stimulation parameters for Parkinson's disease using functional MRI and machine learning. Nat Commun 2021;12(1). https://doi.org/10.1038/s41467-021-23311-9.
[13] Shamir RR, Dolber T, Noecker AM, Walter BL, McIntyre CC. Machine learning approach to optimizing combined stimulation and medication therapies for Parkinson's disease. Brain Stimul 2015;8(6):1025–32. https://doi.org/10.1016/j.brs.2015.06.003.
[14] Mosley PE, Smith D, Coyne T, Silburn P, Breakspear M, Perry A. The site of stimulation moderates neuropsychiatric symptoms after subthalamic deep brain stimulation for Parkinson's disease. NeuroImage Clin 2018;18:996–1006. https://doi.org/10.1016/j.nicl.2018.03.009.
[15] Oliveira FHM, Machado ARP, Andrade AO. On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with Parkinson's disease. Comput Math Methods Med 2018;2018(1):8019232. https://doi.org/10.1155/2018/8019232.
[16] Reich MM, et al. Probabilistic mapping of the antidystonic effect of pallidal neurostimulation: a multicentre imaging study. Brain 2019;142(5):1386–98. https://doi.org/10.1093/brain/awz046.
[17] Shah SA, Brown P, Gimeno H, Lin J-P, McClelland VM. Application of machine learning using decision trees for prognosis of deep brain stimulation of globus pallidus internus for children with dystonia. Front Neurol 2020;11. https://doi.org/10.3389/fneur.2020.00825.
[18] Boutet A, et al. Sign-specific stimulation 'hot' and 'cold' spots in Parkinson's disease validated with machine learning. Brain Commun 2021;3(2):fcab027. https://doi.org/10.1093/braincomms/fcab027.
[19] Chen Y, et al. Predict initial subthalamic nucleus stimulation outcome in Parkinson's disease with brain morphology. CNS Neurosci Ther 2022;28(5):667–76. https://doi.org/10.1111/cns.13797.
[20] Nowacki A, et al. Probabilistic mapping reveals optimal stimulation site in essential tremor. Ann Neurol 2022;91(5):602–12. https://doi.org/10.1002/ana.26324.
[21] Kleinholdermann U, Bacara B, Timmermann L, Pedrosa DJ. Prediction of movement ratings and deep brain stimulation parameters in idiopathic Parkinson's Disease. Neuromodulation Technol Neural Interface 2023;26(2):356–63. https://doi.org/10.1016/j.neurom.2022.09.010.
[22] Wu Y, et al. Prediction of subthalamic stimulation efficacy on isolated dystonia via support vector regression. Heliyon 2024;10(10):e31475. https://doi.org/10.1016/j.heliyon.2024.e31475.
[23] Baumgarten C, Zhao Y, Sauleau P, Malrain C, Jannin P, Haegelen C. Improvement of pyramidal tract side effect prediction using a data-driven method in subthalamic stimulation. IEEE Trans Biomed Eng 2017;64(9):2134–41. https://doi.org/10.1109/TBME.2016.2638018.
[24] Peralta M, Haegelen C, Jannin P, Baxter JSH. PassFlow: a multimodal workflow for predicting deep brain stimulation outcomes. Int J Comput Assist Radiol Surg 2021;16(8):1361–70. https://doi.org/10.1007/s11548-021-02435-9.
[25] Roediger J, Dembek TA, Wenzel G, Butenko K, Kühn AA, Horn A. StimFit-A data-driven algorithm for automated deep brain stimulation programming. Mov Disord 2021;n/a. https://doi.org/10.1002/mds.28878.
[26] Al-Fatly B, Ewert S, Kübler D, Kroneberg D, Horn A, Kühn AA. Connectivity profile of thalamic deep brain stimulation to effectively treat essential tremor. Brain 2019;142(10):3086–98. https://doi.org/10.1093/brain/awz236.
[27] Dembek TA, et al. Probabilistic sweet spots predict motor outcome for deep brain stimulation in Parkinson disease. Ann Neurol 2019;86(4):527–38. https://doi.org/10.1002/ana.25567.
[28] Segura-Amil A, et al. Programming of subthalamic nucleus deep brain stimulation with hyperdirect pathway and corticospinal tract-guided parameter suggestions. Hum Brain Mapp 2023. https://doi.org/10.1002/hbm.26390.
[29] Abboud H, et al. Predictors of functional and quality of life outcomes following deep brain stimulation surgery in Parkinson's disease patients: disease, patient, and surgical factors. Park Dis 2017;2017(1):5609163. https://doi.org/10.1155/2017/5609163.
[30] Farrokhi F, et al. Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms. World Neurosurg 2020;134:e325–38. https://doi.org/10.1016/j.wneu.2019.10.063.
[31] Habets JGV, et al. Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease—proof of principle in a retrospective cohort. PeerJ 2020;8:e10317. https://doi.org/10.7717/peerj.10317.
[32] Ferrea E, Negahbani F, Cebi I, Weiss D, Gharabaghi A. Machine learning explains response variability of deep brain stimulation on Parkinson's disease quality of life. Npj Digit Med 2024;7(1):1–11. https://doi.org/10.1038/s41746-024-01253-y.

[33] Rajamani N, et al. Deep brain stimulation of symptom-specific networks in Parkinson's disease. Nat Commun 2024;15(1):4662. https://doi.org/10.1038/s41467-024-48731-1.
[34] Kostoglou K, Michmizos KP, Stathis P, Sakas D, Nikita KS, Mitsis GD. Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings. IEEE Trans Biomed Eng 2017;64(5):1123–30. https://doi.org/10.1109/TBME.2016.2591827.
[35] Shah A, et al. Combining multimodal biomarkers to guide deep brain stimulation programming in Parkinson disease. Neuromodulation Technol Neural Interface 2023;26(2):320–32. https://doi.org/10.1016/j.neurom.2022.01.017.
[36] Bucciarelli V, et al. Predicting deep brain stimulation outcomes using intra-operative stimulation test data. Curr Dir Biomed Eng 2025;11(1):350–3. https://doi.org/10.1515/cdbme-2025-0189.
[37] Bucciarelli V, et al. Influence of statistical approaches on probabilistic sweet spots computation in Deep Brain Stimulation for severe Essential Tremor. NeuroImage Clin 2025;47:103820. https://doi.org/10.1016/j.nicl.2025.103820.
[38] Bucciarelli V, et al. Towards robust probabilistic maps in deep brain stimulation: exploring the impact of patient number, stimulation counts, and statistical approaches. Front Comput Neurosci 2026;19. https://doi.org/10.3389/fncom.2025.1699192.
[39] Bucciarelli V, et al. Influence of statistical approaches on probabilistic sweet spots computation in deep brain stimulation for severe essential tremor. NeuroImage Clin 2025;47:103820. https://doi.org/10.1016/j.nicl.2025.103820.
[40] Vogel D, et al. Probabilistic stimulation mapping from intra-operative thalamic deep brain stimulation data in essential tremor. J Neural Eng 2024;21(3):036017. https://doi.org/10.1088/1741-2552/ad4742.
[41] Johansson JD, Alonso F, Wardell K. Patient-Specific Simulations of Deep Brain Stimulation Electric Field with Aid of In-house Software ELMA. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Int. Conf., vol; 2019. p. 5212–6. https://doi.org/10.1109/EMBC.2019.8856307.
[42] Fedorov A, et al. 3D Slicer as an image computing platform for the Quantitative Imaging Network. Magn Reson Imaging 2012;30(9):1323–41. https://doi.org/10.1016/j.mri.2012.05.001.
[43] Beninca J, Zemma E, Lovey D, Vera L, Ibáñez M. Programa para la planificación de cirugías estereotácticas. NeuroTarget 2017;11(4). https://doi.org/10.47924/neurotarget2017137.
[44] Alonso F, Latorre M, Göransson N, Zsigmond P, Wårdell K. Investigation into deep brain stimulation lead designs: a patient-specific simulation study. Brain Sci 2016;6(3):39. https://doi.org/10.3390/brainsci6030039.
[45] Åström M, Diczfalusy E, Martens H, Wårdell K. Relationship between neural activation and electric field distribution during deep brain stimulation. IEEE Trans Biomed Eng 2015;62(2):664–72. https://doi.org/10.1109/TBME.2014.2363494.
[46] Hemm-Ode S, et al. Patient-specific electric field simulations and acceleration measurements for objective analysis of intraoperative stimulation tests in the thalamus. Front Hum Neurosci 2016;10(577). https://doi.org/10.3389/fnhum.2016.00577.
[47] Vogel D, Shah A, Coste J, Lemaire J-J, Wårdell K, Hemm S. Anatomical brain structures normalization for deep brain stimulation in movement disorders. NeuroImage Clin p 2020:102271. https://doi.org/10.1016/j.nicl.2020.102271.
[48] Kruschke JK, Liddell TM. The bayesian new statistics: hypothesis testing, estimation, meta-analysis, and power analysis from a bayesian perspective. Psychon Bull Rev 2018;25(1):178–206. https://doi.org/10.3758/s13423-016-1221-4.
[49] Cerda P, Varoquaux G, Kégl B. Similarity encoding for learning with dirty categorical variables. Mach Learn 2018;107(8–10):1477–94. https://doi.org/10.1007/s10994-018-5724-2.
[50] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002;16:321–57. https://doi.org/10.1613/jair.953.
[51] Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, in KDD '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 785–94. https://doi.org/10.1145/2939672.2939785.
[52] Holm S. A simple sequentially rejective multiple test procedure. Scand J Stat 1979;6(2):65–70. https://www.jstor.org/stable/4615733?seq=1.
[53] Wang H, Liang Q, Hancock JT, Khoshgoftaar TM. Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. J Big Data 2024;11(1):44. https://doi.org/10.1186/s40537-024-00905-w.
[54] Athisayamani S, Singh TS, AR, Hwang J-Y, Joshi GP. A novel double machine learning approach for detecting early breast cancer using advanced feature selection and dimensionality reduction techniques. Sci Rep 2025;15(1):22971. https://doi.org/10.1038/s41598-025-06426-7.
[55] Hamani C, et al. Insertional effect following electrode implantation: an underreported but important phenomenon. Brain Commun 2024;6(3):fcae093. https://doi.org/10.1093/braincomms/fcae093.
[56] Mestre TA, Lang AE, Okun MS. Factors influencing the outcome of deep brain stimulation: placebo, nocebo, lessebo, and lesion effects. Mov Disord 2016;31(3):290–8. https://doi.org/10.1002/mds.26500.
[57] Shah A, et al. Stimulation maps: visualization of results of quantitative intraoperative testing for deep brain stimulation surgery. Med Biol Eng Comput 2020;58(4):771–84. https://doi.org/10.1007/s11517-020-02130-y.
[58] Hariz G-M, Fredricks A, Stenmark-Persson R, Hariz M, Forsgren L, Blomstedt P. Blinded versus unblinded evaluations of motor scores in patients with Parkinson's disease randomized to deep brain stimulation or best medical therapy. Mov Disord Clin Pract 2021;8(2):286–7. https://doi.org/10.1002/mdc3.13141.
[59] Shah A, et al. Intraoperative acceleration measurements to quantify improvement in tremor during deep brain stimulation surgery. Med Biol Eng Comput 2017;55(5):845–58. https://doi.org/10.1007/s11517-016-1559-9.
[60] Sammartino F, Rege R, Krishna V. Reliability of intraoperative testing during deep brain stimulation surgery. Neuromodulation Technol Neural Interface 2020;23(4):525–9. https://doi.org/10.1111/ner.13081.