ML-STIM: Machine learning for subthalamic nucleus intraoperative mapping

Journal of Neural Engineering

PURPOSE-LED PUBLISHING™

ML-STIM: Machine learning for subthalamic nucleus intraoperative mapping

To cite this article before publication: Fabrizio Sciscenti *et al* 2025 *J. Neural Eng.* in press https://doi.org/10.1088/1741-2552/adf579

IOP Publishing Journal of Neural Engineering
Journal **XX** (XXXX) XXXXXX https://doi.org/XXXX/XXXX

ML-STIM: Machine Learning for SubThalamic nucleus Intraoperative Mapping

Fabrizio Sciscenti[1,2], Valentina Agostini[1,2], Laura Rizzi[3,4], Michele Lanotte[3,4], and Marco Ghislieri[1,2]

[1] Department of Electronics and Telecommunications, Politecnico di Torino, Turin, 10129, Italy
[2] Polito[BIO]Med Lab, Politecnico di Torino, Turin, 10129, Italy
[3] Department of Neuroscience "Rita Levi Montalcini", University of Turin, Turin, 10126, Italy
[4] AOU Città della Salute e della Scienza di Torino, Turin, 10126, Italy

E-mail: fabrizio.sciscenti@polito.it

Received xxxxxx
Accepted for publication xxxxxx
Published xxxxxx

Abstract

Objective. Deep Brain Stimulation (DBS) of the Subthalamic Nucleus (STN) is effective in alleviating motor symptoms in medication-refractory patients with Parkinson's Disease (PD). Intraoperative identification of the STN relies on MicroElectrode Recordings (MERs), typically analysed by trained operators. However, this approach is time-consuming and subject to variability. For this reason, this study proposes ML-STIM (Machine Learning for SubThalamic nucleus Intraoperative Mapping), a machine learning pipeline designed to automate STN classification from MERs, ensuring high accuracy and real-time performance. ***Approach.*** ML-STIM consists of MERs pre-processing, feature extraction, and classification using a MultiLayer Perceptron (MLP). An adaptive artifact removal algorithm was optimized to balance artifacts identification and STN signal preservation, and the features were selected among those recommended in literature through correlation analysis and ReliefF ranking. The pipeline was trained and validated on a public dataset (Dataset A, 46 patients) and tested on an independent dataset (Dataset B, 36 patients), from a different surgical center, to assess generalizability. Dataset B is made publicly available as well. ***Main results.*** ML-STIM achieved 87.8 ± 1.7% accuracy on Dataset A and 83.8 ± 1.6% accuracy on Dataset B, significantly outperforming a state-of-the-art deep learning model (ResNet-AT, $p < 0.01$). The artifact removal step significantly improved classification specificity ($p < 0.001$). ML-STIM processed raw 10-second recordings in 139.4 ± 2.1 milliseconds, demonstrating real-time feasibility. ***Significance.*** These results confirm ML-STIM as an accurate, interpretable, and computationally efficient solution for intraoperative STN identification in DBS surgeries.

Keywords: deep brain stimulation, STN-DBS, electrode placement, artifacts detection, real-time classification, multilayer perceptron, PD

1. Introduction

Parkinson's Disease (PD) is a progressive neurodegenerative disorder affecting millions of people worldwide [1]. It causes both motor and non-motor symptoms originating from a loss of dopamine in the Substantia Nigra and the Striatum, which are deep brain structures integrated into the Basal Ganglia network responsible for the modulation of sensorimotor and limbic systems [2]. Dopamine replacement therapy, like levodopa (L-DOPA) administration,

is the most common medical approach to treat PD. Still, despite its effectiveness in the early stages of the disease, as it progresses, L-DOPA starts provoking non-negligible side effects, such as dyskinesia and motor fluctuations [3]. For medication-refractory cases, high-frequency Deep Brain Stimulation (DBS) is the most effective surgical intervention [4]. It has been proven successful in alleviating PD motor symptoms, such as tremors, rigidity, and bradykinesia [5–7]. DBS neurosurgery involves the placement of bilateral electrodes in specific brain structures through stereotactic surgery. The electrodes are connected by wires to a programmable Implantable Pulse Generator (IPG), implanted subcutaneously on the chest wall [4, 8]. The IPG works by delivering high-frequency (>100 Hz) electrical impulses to specific brain areas, such as the SubThalamic Nucleus (STN), to alleviate motor symptoms. In addition, it allows reducing the therapeutic dosage of L-DOPA, thereby ameliorating the overall quality of life of patients suffering from PD [5]. During the surgery, a stereotactic frame is attached to the patient's skull to define a system of coordinates to precisely guide the placement of the DBS electrode. Then, the patient undergoes MRI or CT scans, with the frame in place, to help the surgeon create a 3D map of the brain and plan a secure lead trajectory [8, 9]. In this context, MicroElectrode Recordings (MERs) offer a significant improvement for DBS targeting. In clinical practice, MERs are routinely acquired during the surgical procedure for intra-operative mapping [10, 11]. Trained operators rely on MERs, in addition to microstimulation, to choose the best placement before the final DBS lead implantation by visually inspecting signals. However, this approach is time-consuming and heavily dependent on the operator's expertise, potentially leading to variability in outcomes.

1.1 Related works

Studies about targeting the STN from MERs date back to the early 2000s. In the literature, researchers proposed several strategies to automatically identify the STN from MERs by extracting relevant features for Machine Learning (ML) algorithms or rather using Deep Learning (DL) approaches to classify signals based on their intrinsic properties. Studies about MER classification differ for signal processing, classification models employed, extracted features, data normalization, and validation strategy [12, 13]. Specifically, signal processing strategy strongly depends on the type of signals analysed, which could be Multi-Unit Activities (MUAs) or Local Field Potentials (LFPs). In general, LFP signals are low-pass filtered with cut-off frequencies below 500 Hz, while MUAs are high-pass filtered with cut-off frequencies greater than 200 Hz [11]. Furthermore, MERs are often corrupted by artifacts because of the small electrode size and the low voltage of the source signal (< 100 μV), which make signals susceptible to mechanical shifts and electromagnetic interference [14–18]. In the literature, several adaptive threshold-based methods for artifact removal have been proposed. These methods often divide each recording into short, consecutive segments, calculate the energy content or magnitude of each segment, and then classify epochs as either clear or corrupted using a hard threshold [19–21]. More advanced methods include extracting spectral features from signal epochs and training ML classifiers to discriminate between clear and corrupted epochs [14], or employing DL neural networks that take as input either the raw signal epoch [17, 22] or an image derived from the signal epoch (such as the scalogram from the continuous wavelet transform) [15]. Regarding STN targeting, the most commonly used ML classifiers include Support Vector Machines (SVMs) [23], k-Nearest Neighbors (kNN) [24, 25], Random Forests of decision trees (RF) [21, 26, 27], Hidden Markov Models (HMMs) [20, 28], and Artificial Neural Networks (ANNs) [29]. More recently, an increasing number of studies have been using DL algorithms for MER classification due to their ability to automatically extract meaningful features from signals without requiring manual feature engineering and selection. However, despite their effectiveness, these models often lack explainability, making it difficult to interpret their decision-making process. The most popular DL approaches for MER classification are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) [30–34].

As highlighted in **Table 1**, many of the previously published studies suffer from limitations that hinder their generalizability and clinical applicability. In particular, the main limitations include: (*i*) the lack of explainability in DL approaches, which makes it difficult to interpret the decision-making process; (*ii*) the lack of freely available datasets, limiting the standardized training and testing of new methods; (*iii*) the small number of patients typically involved, with single-center studies including only 5 to 50 patients and only a few multi-center studies analyzing up to 100 patients [13]; and (*iv*) the differences in acquisition systems' characteristics across studies, such as the MER sampling frequency and the number of simultaneously recorded electrodes (e.g., five electrodes arranged in a Ben's Gun [11]), which further complicate cross-study comparisons and model reproducibility.

1.1 Aim of the study

This study aims to introduce a comprehensive Machine Learning pipeline for SubThalamic nucleus Intraoperative Mapping (ML-STIM), including processing, feature engineering, and classification. ML-STIM is specifically designed to bridge the key gaps still present in the field by ensuring high classification accuracy, enhanced

Table 1. Recent related works.

Authors	Year	N pts.	Signal	Method	Performance	Limitations
Hosny *et al.* [23]	2021	21	MUA	SVM	Sens = 96.7 % Spec = 92.2 %	No computational time reported.
Coelli *et al.* [27]	2021	13	MUA	RF	Sens = 81.0 % Spec = 95.8 %	Although spike-dependent features are used, there is no spike-sorting algorithm able to work online.
Ciecierski *et al.* [32]	2024	46	MUA, LFP	CNN	Sens = 88.9 % Spec = 84.7 %	Recordings from the same patient both in training and testing sets.
Maged *et al.* [33]	2024	21	MUA, LFP	CNN	Sens = 95.9 % Spec = 98.4 %	Patient-dependent normalization.

Performance metrics are self-reported by the original authors and may not be directly comparable, as they are based on different datasets and validation strategies. N pts.: number of patients enrolled; MUA: Multi-Unit Activity; LFP: Local Field Potential; CNN: Convolutional Neural Network; RF: Random Forest of decision trees; SVM: Support Vector Machine; Sens: Sensitivity; Spec: Specificity.

interpretability, and real-time application. To promote transparency and reproducibility, ML-STIM is trained and validated on the publicly available Dataset by Ciecierski *et al.* [32, 35], and its performance is compared against their approach based on a Residual Neural Network with attention in the temporal domain (ResNet-AT) [32]. Then, to assess generalizability, both methods (ML-STIM and ResNet-AT) are tested on a completely unseen Dataset acquired with a different acquisition system, further assessing their robustness across varying acquisition conditions. The test dataset and the newly proposed Python algorithm are made freely available on Zenodo (https://doi.org/10.5281/zenodo.14894226; accessed on 29 July 2025) and GitHub (https://github.com/Biolab-PoliTO/ML-STIM; accessed on 29 July 2025), respectively.

2. Materials and methods

ML-STIM is a freely available Python (version 3.12) toolbox that includes: (*i*) a processing step aimed at isolating MER signals from unwanted noise, motion artifacts, and electromagnetic artifacts; (*ii*) a feature extraction step to elicit relevant information from MERs; and (*iii*) a classification step through a MultiLayer Perceptron (MLP) for STN targeting. To select the best MLP architecture, the publicly available dataset by Ciecierski *et al.* [32, 35], hereafter referred to as 'Dataset A', was used. Moreover, to test the effectiveness of the artifact removal strategy on STN targeting, ML-STIM is first compared to its variant without the intermediate artifact removal step (ML-STIM$_{noAR}$). Then, the version (either ML-STIM or ML-STIM$_{noAR}$) that demonstrates superior performance is compared to Ciecierski's method ResNet-AT [32] to test the performance of the proposed algorithm against a representative state-of-the-art approach. **Figure 1** schematically illustrates the procedure.

Finally, to test the pipeline generalizability to different scenarios and acquisition conditions, ML-STIM and ResNet-AT testing is performed on an entirely unseen dataset, hereafter referred to as 'Dataset B'.

2.1 Dataset A: training and validation set

Dataset A contains data collected during 46 DBS surgeries performed by the same surgical team. MERs were acquired simultaneously from 3 electrodes (i.e., medial, anterior, and central electrodes) at a sampling rate of 24 kHz. Data were collected bilaterally at depths ranging from -10 mm to +8 mm relative to the estimated STN location (EDT: Estimated Distance from Target). The number of available recordings per hemisphere is 50.5 ± 5.7 (average ± standard deviation), and every recording is labelled as either inside (IN-STN) or outside the STN (OUT-STN). Overall, the dataset is made of 3210 OUT-STN recordings (duration: 9.4 ± 1.1 s) and 1440 IN-STN recordings (duration: 9.0 ± 1.3 s). Dataset A, including all electrodes and hemispheres, is randomly split into training (80%), validation (10%), and test (10%) sets in a stratified way. These partitions were inherited from the publicly available dataset, as provided by the original authors. Specifically, 80% of IN-STN recordings and 80% of OUT-STN recordings are assigned to the training set; half of the remaining IN-STN and OUT-STN are equally distributed in the validation and test sets. Recordings were divided into 1-second-long overlapping segments, with an overlap of 50% (i.e., 500 ms) for data augmentation.

For further details about signal acquisitions and dataset composition, please refer to Ciecierski *et al.* [32, 35].

2.2 Dataset B: test set

Dataset B contains data collected from 36 patients suffering from PD (age: 58.4 ± 7.8 years; disease duration: 11.3 ± 4.2 years; 25 males; 11 females). PD patients were enrolled at the Stereotactic and Functional Neurosurgery Unit of the University of Turin (Turin, Italy) among those eligible for STN-DBS neurosurgery. Subjects were selected according to the Core Assessment Program for Surgical Interventional

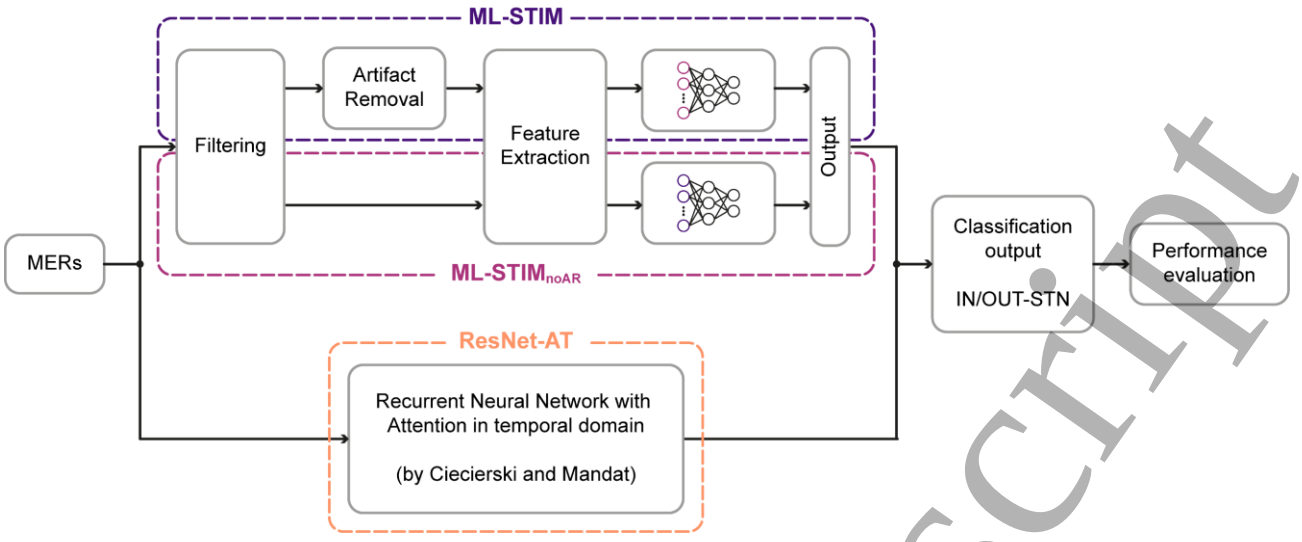

Figure 1. Block-diagram of the procedure followed to train and validate ML-STIM (Machine Learning for SubThalamic nucleus Intraoperative Mapping) for classifying MicroElectrode Recordings (MERs) as INside or OUTside the SubThalamic Nucleus (STN), compared to its variant without an artifact removal step (ML-STIM$_{noAR}$), and the Deep Learning approach by Ciecierski *et al.* [32], ResNet-AT (Residual Neural Network with Attention in the Temporal domain).

Therapies in Parkinson's disease (CAPSIT-PD) [36] and STN-DBS bilateral surgery was performed following the procedure previously described elsewhere [37]. MERs were acquired through the NeuroSet acquisition system (NeuroStar, Tubingen, Germany) from a single electrode at a sampling frequency of 20 kHz and high-pass filtered at 200 Hz upon recording. For all PD patients, MERs were acquired bilaterally at different depths ranging from -10 mm to +7 mm relative to the EDT. On average, the number of available recordings per hemisphere is 11.7 ± 1.1, and every recording is labelled as either inside (IN-STN) or outside the STN (OUT-STN). Overall, the dataset is made of 406 OUT-STN recordings (duration: 32.6 ± 29.9 s) and 285 IN-STN recordings (duration: 61.2 ± 9.9 s). Just like it was done for Dataset A, Dataset B was divided into 1-second epochs with a 50% overlap. To ensure consistency in sampling frequency, Dataset B was resampled to 24 kHz using the FFT-based method *resample* from Python's library Scipy (release 1.15.3).

2.3 MER pre-processing

Prior to STN targeting, MERs classified with the newly proposed approach ML-STIM, were processed as represented in **Figure 1**. Subcortical signals were filtered, motion artifacts were removed, and time- and frequency-domain features were extracted using 1-second sliding window, as detailed below.

2.3.1 Filtering. MERs are band-pass filtered (IIR, Butterworth, 6th order) between 200 Hz and 5 kHz, to keep only the signal band of interest [29, 38], focusing on the frequency range associated with MUAs. Additionally, a notch recursive filter (IIR, 2nd order, 3-dB bandwidth 4 Hz) is applied to suppress higher harmonics of the 50 Hz line frequency [14]. Finally, the filtered recordings undergo the artifact removal step.

2.3.2 Artifact removal. The primary objective of this step is to accurately identify and eliminate artifacts from subcortical recordings (i.e., MERs). It works by scanning the input signal using a sliding window of length l (the first key parameter of the artifact removal algorithm) with steps of $l/2$ samples (i.e., 50% overlap). Once identified, artifacted segments are discarded, and the remaining clean portions of the signal are concatenated to form a continuous, artifact-free signal. To distinguish between corrupted and uncorrupted signal epochs, it computes the amplitude variance s_n^2 of the window x_n as defined in **Equation 1**:

$$s_n^2 = \frac{1}{l} \cdot \sum_{j=1}^{l} (x_n[j] - \mu_n)^2 \quad (1)$$

where $x_n[j]$ represents the j^{th} element of the window x_n, and μ_n its mean value. Then, the algorithm compares the amplitude variance s_n^2 to the variance computed from the latest "stable" signal epoch x_p, referred to as s_p^2, as detailed in **Equation 2**:

$$s_n^2 < th \cdot s_p^2 \quad (2)$$

where th represents a hard threshold that serves as the second key parameter of the artifact removal algorithm [16]. The variance s_p^2 is firstly initialized to a very large value (ideally infinite), and then it is iteratively updated to s_n^2

computed from the last signal epoch x_n that satisfies **Equation 2**. For real-time applications, the algorithm would operate on signal segments of l samples. Chunks passing the artifact detection criteria (**Equation 2**) would be concatenated until one second of clean signal is obtained. The choice of l and th is pivotal for achieving an accurate and real-time implementation of the artifact removal algorithm. Since MERs acquired from the STN often exhibit bursting activity (i.e., short bursts of spiking neurons interleaved by silent periods), the parameters l and th are chosen to optimize artifact removal while preserving authentic STN signals. This ensures that these distinctive bursts are not mistakenly identified as artifacts. To do so, MERs are labelled as either non-STN (null class), artifacts (class 1), or STN activity (class 0), and the algorithm's parameters are chosen to minimize the objective function $O(l, th)$ defined in **Equation 3**:

$$O(l, th) = \sqrt{2b \cdot v_{stn}^2 + 2(b-1) \cdot (v_{art} - 1)^2} + \gamma(l) \quad (3)$$

where v_{stn} is the false positive rate (i.e., STN activity misclassified as artifact) and v_{art} is the true positive rate (i.e., artifacts correctly detected). The parameter b is the relative weight of v_{stn}, empirically set to 0.8. Additionally, $\gamma(l)$ is a penalty term defined to avoid long computational times, as defined in **Equation 4**:

$$\gamma(l) = \exp\left(\frac{l - l_{max}}{\tau}\right) \quad (4)$$

where l_{max} is fixed at 1 second and τ is set equal to 500 milliseconds. The artifact removal algorithm was optimized using manually labelled artifacts from Dataset B as ground truth. Three independent operators (i.e., a neurosurgeon with extensive MER analysis experience and two biomedical engineers) manually segmented artifacts based on visual inspection of MER time series and spectrogram heatmaps obtained through Short-Time Fourier Transform (STFT). MERs were analysed in 200-ms windows. Any discrepancies in artifact segmentation were discussed until a consensus was reached. **Figure 2** illustrates the manual annotation provided to each rater along with a representative example demonstrating the application of the artifact removal step.

Bayesian Optimization was used to determine the optimal parameters, th = 1.33 and l = 100 ms, that minimize the objective function $O(l, th)$ defined in **Equation 3**. The selection of these optimal parameters resulted in a false positive rate (v_{stn}) of 0.18 and a true positive rate (v_{art}) of 0.54 when compared against manually labeled artifacts. The true positive rate (v_{art}) of 0.54 reflects the conservative nature of the artifact detection strategy, which is designed to minimize false positives to avoid excessive STN signal loss. **Figure 3.A** visually represents the objective function $O(l, th)$, with the indication of the optimal algorithm parameters highlighted by a blue asterisk.

To evaluate the impact of the artifact removal step on STN targeting accuracy, the proposed pipeline (ML-STIM) is compared to its variant without this intermediate artifact removal step, referred to as ML-STIM$_{noAR}$.

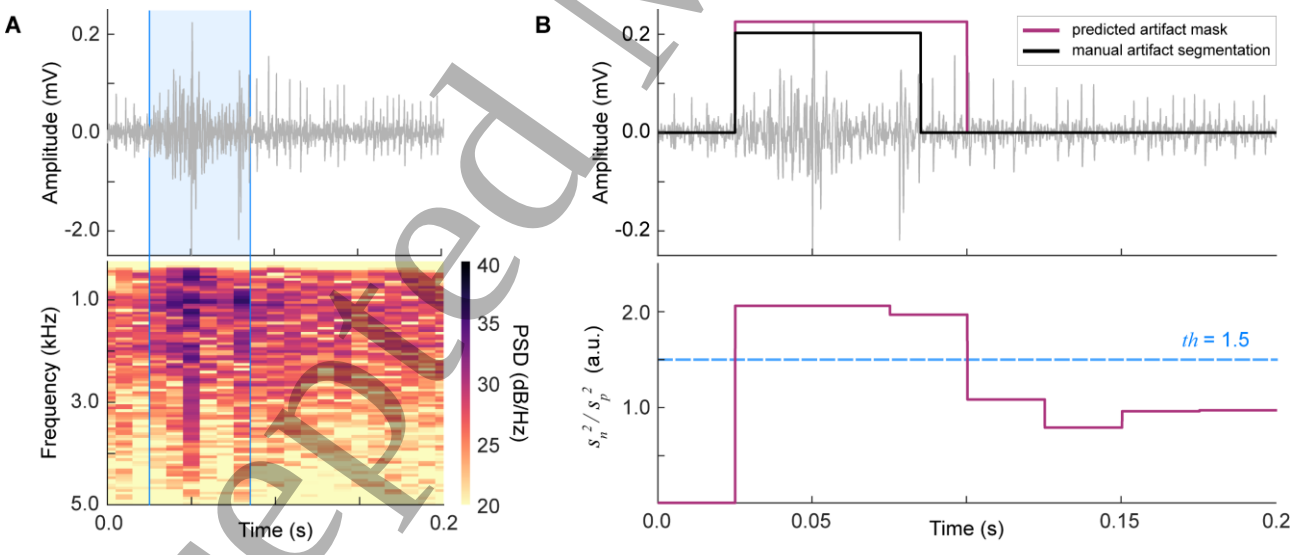

Figure 2. Artifact removal step. (Panel A) Artifact manual annotation set-up: 200-milliseconds segment of MicroElectrode Recording (MER) (top) acquired from the SubThalamic Nucleus (STN), and its corresponding Short-Time Fourier Transform (bottom). Manually labelled artifacts are highlighted in blue. (Panel B) Magnified MER with the superimposition of the manual artifact annotation and an example of the predicted artifact mask by setting th = 1.5 (threshold) and l = 50 ms (window length) (top); the bottom panel shows the ratio between amplitude variances of consecutive overlapping signal windows: segments for which s_n^2/s_p^2 exceeds th are segmented as artifacts.

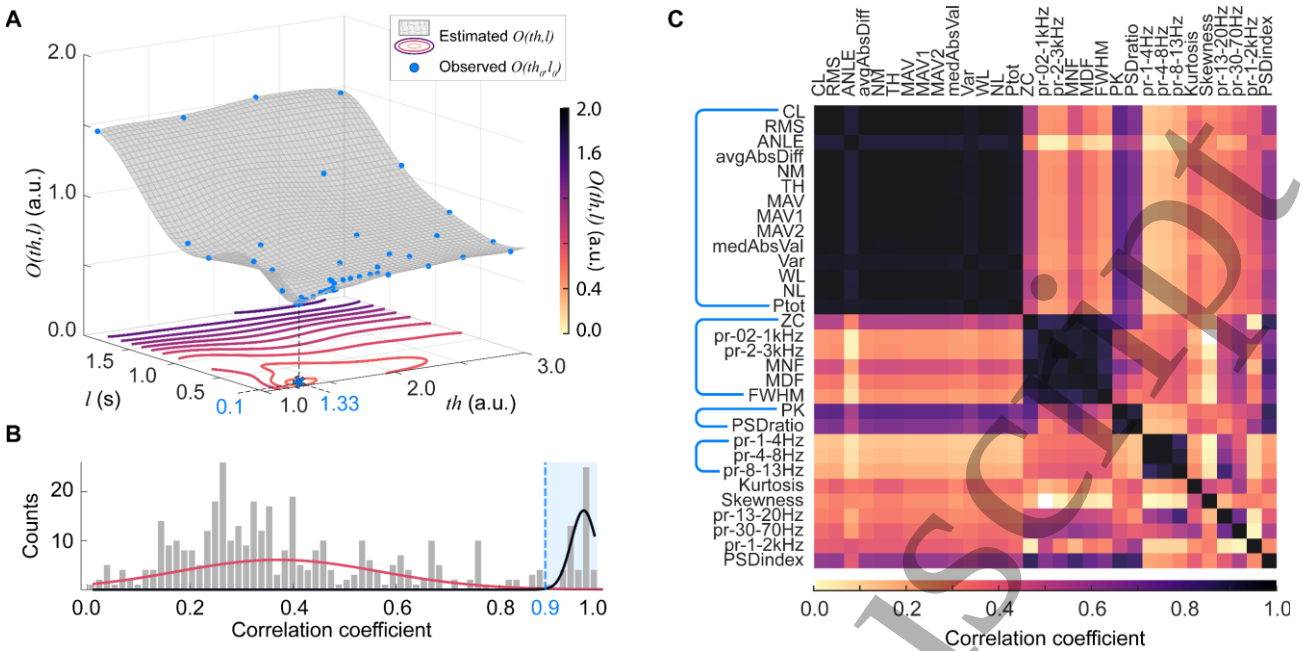

Figure 3. Pipeline optimization results. (Panel A) Approximation of the objective function $O(th, l)$ via Bayesian Optimization. The function minimum is indicated with a blue asterisk corresponding to the optimal algorithm parameters: th = 1.33 (threshold) and l = 100 ms (window length). (Panel B) Distribution of the absolute correlation coefficients (grey bars), with superimposed the two Gaussian components (orange and black solid lines) fitted on the distribution. The intersection point (ξ=0.9) between the Gaussian components is indicated with a blue dashed vertical line. (Panel C) Feature-Feature correlation matrix where the element (i,j) represents the absolute correlation coefficient between the i^{th} and j^{th} features. Features correlated for more than 0.9 are grouped with blue arcs.

2.4 Feature extraction and selection

After pre-processing, signals were segmented into 1-second-long epochs with 50% overlap and 31 descriptive features, selected based on literature recommendations [12, 16, 20, 21, 23–29, 39–44], were extracted from each epoch. The extracted features can be distinguished into time- and frequency-domain features. Time-domain features include: Curve Length (CL), Root Mean Square (RMS), Average Non-Linear Energy ($ANLE$), Zero Crossings (ZC), Average Absolute Difference ($avgAbsDiff$), Noise Mode (NM) computed as three times the standard deviation, Threshold (TH), Mean Absolute Value (MAV), Mean Absolute Value computed from the signal segment whose first and last quarters are weighted 0.5 and linearly from 0 to 1 ($MAV1$ and $MAV2$, respectively), Median Absolute Value ($medAbsVal$), amplitude Variance (Var), Waveform Length (WL), Noise Level (NL) defined as the mode of the signal's envelope, Peak Count (PK), amplitude Kurtosis ($Kurtosis$), and amplitude Skewness ($Skewness$). Frequency-domain features were extracted through Welch's method (Hann's 2400-sample window, 50% overlap, and NFFT equal to 2400 samples). Features from bands lower than 200 Hz were extracted from the PSD of the rectified signal [45]. The following frequency-domain features were extracted from each 1-second-long epoch: total Power ($Ptot$), relative frequency in the bands 1-4 Hz (pr_1_4Hz), 4-8 Hz (pr_4_8Hz), 8-13 Hz (pr_8_13Hz), 13-30 Hz (pr_13_30Hz), 30-70 Hz (pr_30_70Hz), 200-1000 Hz (pr_02_1kHz), 1-2 kHz (pr_1_2kHz), and 2-3 kHz (pr_2_3kHz), Mean Frequency (MNF), Median Frequency (MDF), PSD's Full-Width at Half Maximum ($FWHM$), $PSDindex$ [39], and the ratio between the power-band in the range 3-4 kHz and the power-band in the range 1.6-2.2 kHz ($PSDratio$) [20].

Further details about the time- and frequency-domain features are provided in **Table S.1** of the supplementary material.

To select a subset of relevant features, feature-feature correlation was evaluated computing the feature correlation matrix, where diagonal elements (self-correlation) equal 1 and off-diagonal elements range between -1 and +1 indicating the strength and direction of linear correlations between features [46]. Since the objective is to assess the presence of correlation regardless of its direction, the absolute value of the correlation coefficient was considered. **Figure 3.C** shows a representative correlation matrix, displaying the absolute correlation coefficients for each features pair by using a color map. Strongly correlated features were identified by setting a threshold (ξ) to the absolute correlation coefficients. The threshold ξ was determined by identifying the intersection point between the two Gaussian curves of a Gaussian Mixture Model (GMM) fitted on the absolute correlation coefficient

distribution as represented in **Figure 3.B** [47]. In this study, based on the correlation coefficient distribution, the threshold ξ was defined equal to 0.9.

Strongly correlated features were then clustered into subgroups and further processed to select only the most representative feature from each subgroup and reduce the feature set to be used for STN targeting. To do so, the Bhattacharyya coefficient (BC) was computed for every feature within a subgroup. The BC measures the discriminative power of a feature by evaluating the overlap between its distributions in the two classes (i.e., inside and outside the STN). Thus, for each subgroup, only the feature with the lowest BC value (i.e., highest discriminative power) was retained for the following analyses.

Finally, both weakly correlated features ($< \xi$) and the most discriminative features obtained from each subgroup are ranked exploiting the ReliefF ranking algorithm, which uses a weighting criterion based on the discriminant capability of each feature [48]. Prior to classification, all features are min-max scaled using the extreme values (1st and 99th percentiles) computed from the training partition of Dataset A. The resulting normalized feature set is then used as input to the STN classifier.

2.5 STN classifier design

In this study, a MultiLayer Perceptron (MLP) was implemented for classifying STN signals from MERs. MLP is the simplest configuration of a sequential Artificial Neural Network (ANN), being composed of an input layer, one or more hidden layers, and an output layer. To define the best MLP architecture, Dataset A was divided into 3 different sets according to the study by Ciecierski *et al*. [32, 35]: training set (80%), validation set (10%), and test set (10%). Training and validation sets were used to train the MLP models, while the test set was used to assess classifier performance. Different configurations were tested to identify the MLP architecture with the highest performance, including the number of input neurons, number of hidden layers, and number of hidden units per layer.

In particular, the optimization method described in **Algorithm 1** was implemented, testing different number of input features from the top 10 features obtained through the

Algorithm 1. MLP-architecture optimization algorithm.

Input data: $T_s \in \mathbb{R}^{N,f}$, $V_s \in \mathbb{R}^{M,f}$
Input parameters: C_N , C_L , L_{max} , M_{max} , $toll$
Output: A_{best} ;
Initializations:
  $l \leftarrow 1$, $m \leftarrow 2$, $c_n \leftarrow 0$, $c_l \leftarrow 0$ *# counters*
  $A \leftarrow (f, m, 1)$ *# architecture with f input neurons, and $l = 1$ hidden layer with $m = 2$ neurons*
  $A^l_{best} \leftarrow [\,]$, $A_{best} \leftarrow [\,]$ *# best architectures*
  $p_{best} \leftarrow 0$, $p^l_{best} \leftarrow 0$ *# best performances*
  $F_l \leftarrow true$, $F_n \leftarrow true$ *# flags*
While F_l **is** $true$ **and** $l \leq L_{max}$ **do**:
  $p^l_{best} \leftarrow 0$, $A^l_{best} \leftarrow [\,]$, $F_n \leftarrow true$, $c_n \leftarrow 0$
  $m \leftarrow A[2]$ *# initialize m as the neurons in the last added (first) hidden layer*
  While F_n **is** $true$ **and** $m \leq M_{max}$ **do**:
    If len$(A) < l$ **do**: $A \leftarrow$ insert$(m, position = 2)$ *# insert a hidden layer with m neurons after the input layer*
      Else: $A[2] \leftarrow m$ *# assign m neurons to the first hidden layer*
    $p_A \leftarrow$ train and validate A on $\{T_s, V_s\}$
    If $p_A \geq p^l_{best} + toll$ **do**: $p^l_{best} \leftarrow p_A$ **and** $A^l_{best} \leftarrow A$ **and** $c_n \leftarrow 0$ *# update architecture A^l_{best}*
      Else: $c_n \leftarrow c_n + 1$
    If $c_n > C_N$ **do**: $F_n \leftarrow false$ *# early-stop for neurons reached*
    $m \leftarrow m + 1$
  If $p^l_{best} \geq p_{best} + toll$ **do**: $p_{best} \leftarrow p^l_{best}$ **and** $A_{best} \leftarrow A^l_{best}$ **and** $c_l \leftarrow 0$ *# update best architecture A_{best}*
    Else: $c_l \leftarrow c_l + 1$
  If $c_l > C_L$ **do**: $F_l \leftarrow false$ *# early-stop for layers reached*
  $A \leftarrow A^l_{best}$ *# initialize A, to be updated, with A^l_{best}*
  $l \leftarrow l + 1$

MLP: MultiLayer Perceptron. $T_s \in \mathbb{R}^{N,f}$: training set with N observations on the rows, each described with f features, $V_s \in \mathbb{R}^{M,f}$: validation set with $M < N$ observations, C_N: early-stop threshold when adding neurons, C_L: early-stop threshold when adding layers, L_{max}: maximum number of hidden layers, M_{max}: maximum number of neurons per hidden layer, $toll$: minimum performance improvement tolerance, A_{best}: best architecture, A^l_{best}: best architecture with l hidden layers, l: count of layers, m: count of neurons per hidden layer, c_n: count of failed neuron additions, c_l: count of failed layer additions, p_{best}: accuracy of the best architecture A_{best}, p^l_{best}: best performance for the architecture A with l hidden layers, F_l: Boolean flag to continue adding layers, F_n: Boolean flag to continue adding neurons in the l^{th} hidden layer.

ReliefF ranking. The optimization method was conducted twice: the first time using MERs after artifact removal (ML-STIM) and the second time using MERs without artifact removal (ML-STIM$_{noAR}$). **Algorithm 1** sequentially adds hidden layers backwards from the output- to the input-layer, making sure that the network growth converges toward the output layer. The network growth progresses firstly adding neurons to the current hidden layer and then adding hidden layers, until a stopping criterion is met. The stopping criterion adopted is the so-called *early stopping*.

2.6 Comparison with the state-of-the-art

ML-STIM was compared to the state-of-the-art deep learning method proposed by Ciecierski *et al.* [32], which uses a residual neural network with temporal-domain attention (ResNet-AT). This architecture is built on the traditional ResNet and is enhanced with a self-attention layer added after each of the ResNet layers. ResNet-AT was originally designed to process raw recording segments of 500 ms (12000 samples), which are converted into 129 (frequency) by 53 (time) spectrograms. The model was trained by its authors on the training partition of Dataset A and the trained model was made publicly available [32, 35].

2.7 Performance evaluation

Performance evaluation was performed using custom Python scripts (version 3.12), exploiting the PyTorch library (release 2.5.1) for training and validating ML-STIM and ML-STIM$_{noAR}$ models, and for validating the freely available ResNet-AT model [32]. Scripts were run on a machine with CPU Intel Core i7 (2.10 GHz), 32 GB of RAM, and NVIDIA GeForce RTX 3060 GPU. The classifiers take as input 1-second (ML-STIM and ML-STIM$_{noAR}$) or 500-milliseconds (ResNet-AT) segments and output the probability of belonging to the STN. For each MER, these probabilities are averaged across all segments within the same recording to obtain a probability for each recording. Subsequently, performance metrics are computed at the patient level by aggregating the results across all MERs from the same patient. The performances of each model were assessed with standard evaluation metrics: accuracy, sensitivity, specificity, F_1-score, area under the Receiver Operating Curve (ROC), and computational time for analysing each raw recording. Computational times are computed on the same machine for all methods, to highlight the computational efficiency under similar conditions.

An ablation study was conducted to compare the classification performances of ML-STIM and ML-STIM$_{noAR}$, with the aim of investigating the contribution of the artifact removal step to the overall pipeline. Following this internal comparison, the best-performing configuration was then compared against the representative state-of-the-art method ResNet-AT [32]. This comparison was carried out by directly applying the model trained on the training partition of Dataset A to both the test partition of Dataset A and the external Dataset B, the latter being used to evaluate generalizability across independent patient data.

2.8 Statistical analysis

Statistical comparisons are first performed between ML-STIM and ML-STIM$_{noAR}$ performance to assess the impact of the artifact removal step on classification outcomes. Subsequently, the best among the two configurations (ML-STIM or ML-STIM$_{noAR}$) is compared with the state-of-the-art method ResNet-AT [32].

Table 2. Key parameters chosen for STN classifier.

MLP hyperparameters		**ML-STIM**	**ML-STIM$_{noAR}$**
network parameters	input neurons	9	8
	hidden layers	[7,4,4,2]	[10,8,4,2]
	output neurons	1	1
	activation function for hidden neurons	hyperbolic tangent	hyperbolic tangent
	activation function for output neurons	sigmoid (threshold = 0.51)	sigmoid (threshold = 0.57)
training parameters	optimizer = *ADAM*; loss function = Mean Square Error (*MSELoss*); batch size = 32; learning rate = 0.001; weight decay = 1e-5; maximum number of iterations = 200;		

MLP architecture optimization parameters (Algorithm 1)	
maximum number of consecutive worsening iterations by adding neurons	$C_N = 3$
maximum number of consecutive worsening iterations by adding layers	$C_L = 3$
maximum number of hidden layers	$L_{max} = 15$
maximum number of neurons per hidden layer	$M_{max} = 128$
minimum accuracy increment to be considered an improvement	$toll = 0.1\ \%$

MLP: Multi-Layer Perceptron; ML-STIM: Machine Learning for SubThalamic nucleus Intraoperative Mapping; ML-STIM$_{noAR}$: ML-STIM without the intermediate artifact removal step.

The distributions of performance metrics across patients were firstly tested for normality with the Shapiro-Wilk test. Depending on the result of the normality test, a Students' paired *t*-test (for normally distributed data) or a Wilcoxon's signed-rank test (for non-normally distributed data) was used to assess statistical differences between methods. The effect size of statistically significant differences was computed through the Hedges' g statistic [49], including correction for small sample size. The significance level α was set equal to 0.05 for all the analyses, and all the variables were expressed as mean ± standard error across the population.

All statistical analyses were performed using the standard software package SPSS Statistical Software, version 27.0, SPSS Inc., Chicago, IL.

3. Results

First, this section presents the results of the feature extraction and selection process, which defines the set of features used for STN targeting, along with the final architecture of the MLP obtained through the optimization algorithm. Next, the results from the ablation study comparing the performance of the proposed method with and without artifact removal (ML-STIM vs. ML-STIM$_{noAR}$), are reported, in order to highlight the contribution of this component to the overall pipeline. The best-performing configuration is then compared against a previously published method (ResNet-AT). Finally, we assess the generalizability of the classification models using the unseen Dataset B.

3.1 Feature extraction and selection

Using a feature-feature correlation threshold $\xi = 0.9$, 6 out of 31 features ($Kurtosis$, $Skewness$, pr_13_20Hz, pr_30_70Hz, pr_1_2kHz, and $PSDindex$) were identified as weakly correlated. The remaining 25 features were grouped into 4 subgroups of strongly correlated features, as represented in **Figure 3.C** by means of blue arcs. From each cluster, the most discriminative feature was selected, resulting in the following 4 features: $avgAbsDiff$ ($BC = 0.62$), ZC ($BC = 0.88$), $PSDratio$ ($BC = 0.73$), and pr_8_13Hz ($BC = 0.62$). This resulted in a final set of 10 features, which were ranked using the ReliefF algorithm as follows: $avgAbsDiff$, pr_8_13Hz, pr_30_70Hz, $Skewness$, $PSDratio$, ZC, pr_1_2kHz, $PSDindex$, $Kurtosis$, and pr_13_20Hz.

3.2 STN classifier design

The optimal MLP architecture was determined using the sequential optimization method described in **Algorithm 1**. A total of 224 different MLP architectures were tested for ML-STIM and 235 for ML-STIM$_{noAR}$. **Table 2** shows the properties of the MLP model that achieved the highest validation accuracy (89.2 % for ML-STIM and 85.7 % for ML-STIM$_{noAR}$).

3.3 Performance evaluation

The performance of the three tested classifiers (i.e., ML-STIM, ML-STIM$_{noAR}$, and ResNet-AT) were estimated in

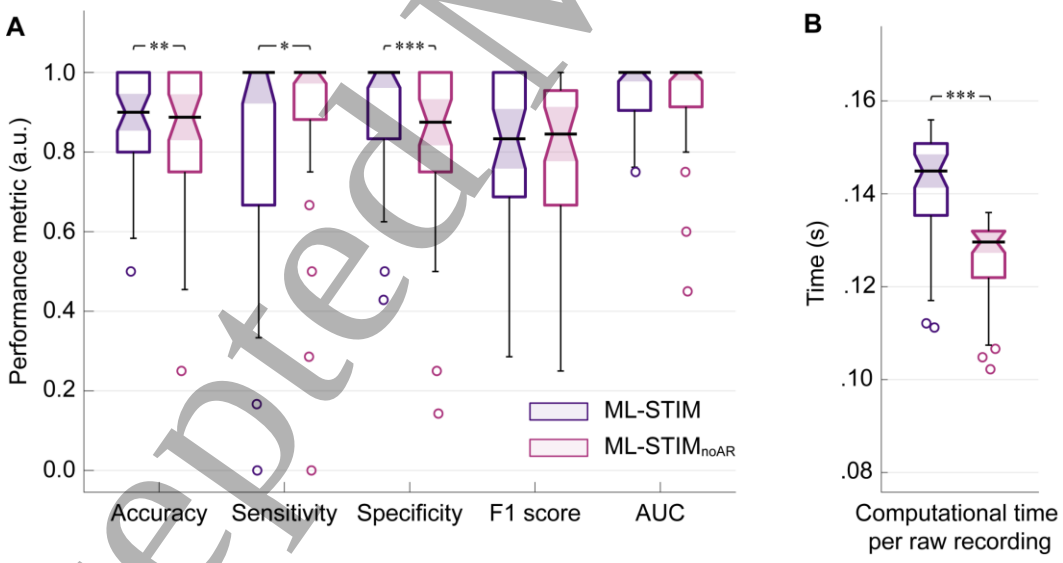

Figure 4. Ablation study results. (A) Comparison of the classification performances of the methods tested: our method with artifact removal (ML-STIM) and without artifact removal (ML-STIM$_{noAR}$). (B) Comparison of the computational time required to process a single raw recording. Metrics' distributions across patients are illustrated through boxplots representing the distribution's central tendency (minimum, 25th percentile, median, 75th percentile and maximum) with a notch delimiting the 95% confidence interval of the median; circles outside of the boxes represent outliers. Asterisks highlight statistically significant differences ($p < 0.05$) between classifiers, as determined by pairwise Wilcoxon's signed-rank tests: ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$. ML-STIM: Machine Learning for SubThalamic nucleus Intraoperative Mapping; ML-STIM$_{noAR}$: ML-STIM without intermediate step of artifact removal.

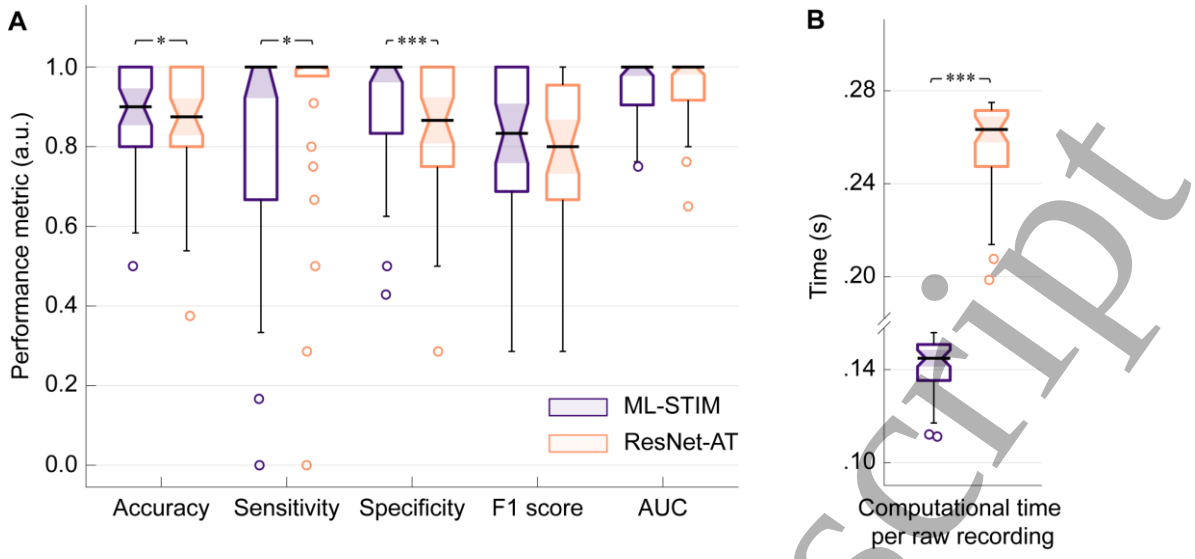

Figure 5. Comparison with the state-of-the-art. (A) Comparison of the classification performances of our method with the artifact removal step (ML-STIM) and ResNet-AT. (B) Comparison of the computational time required to process a single raw recording. Metrics' distributions across patients are illustrated through boxplots representing the distribution's central tendency (minimum, 25th percentile, median, 75th percentile and maximum) with a notch delimiting the 95% confidence interval of the median; circles outside of the boxes represent outliers. Asterisks highlight statistically significant differences ($p < 0.05$) between classifiers, as determined by pairwise Wilcoxon's signed-rank tests: ***$p < 0.001$, **$p < 0.01$, and *$p < 0.05$. ResNet-AT: Residual Neural Network with Attention in the Temporal domain ; ML-STIM: Machine Learning for SubThalamic nucleus Intraoperative Mapping.

terms of accuracy, sensitivity, specificity, F_1-score, area under the ROC, and computational time for analysing each raw recording.

3.3.1 Ablation study. ML-STIM achieved higher accuracy and specificity, but lower sensitivity compared to its variant ML-STIM$_{noAR}$. The two methods showed comparable performance in terms of F_1-score and AUC. Additionally, ML-STIM required more time to process raw recordings than ML-STIM$_{noAR}$. Specifically: statistically significant differences were found in terms of accuracy ($p = 0.006$; $g = 0.30$), sensitivity ($p = 0.008$; $g = 0.28$), specificity ($p < 0.001$; $g = 0.44$), and computational time per raw recording ($p < 0.001$; $g = 1.49$). **Figure 4** and **Table 3** compare the performances of the two variants tested, with the indication of statistically significant differences highlighted through asterisks (*) (**Figure 4**) and daggers (†) (**Table 3**).

3.3.2 Comparison with the state-of-the-art. Given the higher performance of ML-STIM over ML-STIM$_{noAR}$, the comparison with the state-of-the-art was performed for ML-STIM against ResNet-AT. ML-STIM reached higher accuracy and specificity, but lower sensitivity with respect to ResNet-AT, while no significant differences were observed for F_1-score and AUC. Additionally, ML-STIM required less time to process raw recordings than ResNet-AT. Statistically significant differences were found in terms of accuracy ($p = 0.036$; $g = 0.20$), sensitivity ($p = 0.014$; $g = 0.33$), specificity ($p < 0.001$; $g = 0.37$), and computational time per raw recording ($p < 0.001$; $g = 6.70$). **Figure 5** and **Table 3** compare the performances of the two classifiers, with the indication of statistically significant differences highlighted through asterisks (*) (**Figure 5**) and double daggers (‡) (**Table 3**).

3.4 Generalizability testing

The generalizability of STN classification was tested solely for the ML-STIM and ResNet-AT models. The assessment was conducted in terms of accuracy, sensitivity, specificity,

Table 3. STN classifier performances on the test partition of Dataset A.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F_1-score (%)	AUC (%)	Time[1] (ms)
ML-STIM$_{noAR}$	83.6 ± 2.4†	88.7 ± 3.3†	82.2 ± 2.9†	77.9 ± 2.9	93.7 ± 1.7	123.5 ± 1.7†
ML-STIM	87.8 ± 1.7†,‡	81.7 ± 4.0†,‡	90.0 ± 2.1†,‡	80.7 ± 2.8	94.4 ± 1.1	139.4 ± 2.1†,‡
ResNet-AT	85.2 ± 2.0‡	89.8 ± 3.2‡	84.0 ± 2.5‡	79.2 ± 2.7	95.7 ± 1.1	252.0 ± 3.7‡

Metrics' values are reported as *mean* ± *standard error* over the sample population. Daggers (†), and double daggers (‡), represent statistically significant differences between methods, as determined by pairwise Wilcoxon's signed-rank test (†: ML-STIM *vs.* ML-STIM$_{noAR}$; ‡: ML-STIM *vs.* ResNet-AT). [1]Time: computational time per raw recording. ResNet-AT: Residual Neural Network with Attention in the Temporal domain ; ML-STIM: MultiLayer Perceptron for SubThalamic nucleus Intraoperative Mapping; ML-STIM$_{noAR}$: ML-STIM without intermediate step of artifact removal.

F_1-score, and AUC. Overall, ML-STIM reached higher accuracy, sensitivity, and F_1-score, but lower specificity compared to ResNet-AT. The two methods showed comparable performances in terms of AUC. Specifically, statistically significant differences were found in terms of accuracy (p = 0.002; g = 0.58), sensitivity (p < 0.001; g = 0.68), specificity (p = 0.016; g = 0.31), and F_1-score (p < 0.001; g = 0.72). Detailed results of these evaluations, with the indication of the statistically significant differences between methods, are reported in **Table 4**.

4. Discussion

Deep Brain Stimulation (DBS) of the Subthalamic Nucleus (STN) is effective in alleviating motor symptoms in medication-refractory patients with Parkinson's Disease (PD), but its therapeutic efficacy strongly depends on the accurate placement of the stimulating electrode. Intraoperative identification of the STN relies on MicroElectrode Recordings (MERs), which are traditionally interpreted by experienced clinicians. However, this approach can be time-consuming and subject to variability. This work presents ML-STIM, a novel machine learning-based pipeline designed to automate STN classification from MERs intraoperatively. The newly proposed framework integrates signal pre-processing, feature extraction, and classification steps, each optimized for high accuracy and real-time performance. Across patients, ML-STIM achieved a high classification accuracy (mean ± standard error; 87.8 ± 1.7%) and demonstrated real-time capability, with an average processing and classification time of 139.4 ± 2.1 ms for raw 10-second recordings.

The proposed pipeline first introduces a real-time artifact removal step that compares signal variances of consecutive overlapping windows to quickly identify and remove artifacts. This step ensures that the features extracted from clean signals are of high quality, which is essential for achieving accurate classification. Second, ML-STIM performs feature extraction considering an optimized pool of features specifically selected to maximize discriminative power for STN classification and minimize redundancy, thereby enhancing the input space for the machine learning classifier. Finally, a MultiLayer Perceptron (MLP) is used to separate the feature space into two different classes: inside the STN and outside the STN. The MLP architecture is optimized with a custom algorithm that balances simplicity and computational efficiency while guaranteeing good performance.

Consistently with a previously published open-source dataset and algorithm for STN targeting [32], ML-STIM performance was tested against manual labels provided by expert raters (i.e., ground truth). Specifically, ML-STIM reached higher performances in terms of accuracy and specificity when compared to the ResNet-AT model proposed by Ciecierski and colleague [32]. In terms of sensitivity, the performance of the proposed pipeline is slightly lower than that of ResNet-AT (81.7 ± 4.0% for ML-STIM, 89.8 ± 3.2% for ResNet-AT); however, it achieves the best balance between true positive identification and false negative minimization, as demonstrated by a higher F_1-score (80.7 ± 2.8% for ML-STIM, 79.2 ± 2.7% for ResNet-AT), although this difference is not statistically significant (p = 0.61). In terms of computational time required to process each raw recording, ML-STIM is significantly faster given the simplicity of the MLP compared to a CNN. When processing whole bilateral surgeries, consisting of approximately 50 recordings per hemisphere, ML-STIM achieves a total classification time of 14.3 ± 0.3 s, compared to 25.9 ± 0.6 s for Ciecierski's ResNet-AT [32]. This corresponds to a relative percentage reduction in processing time of -42.1 ± 0.6% (relative percentage difference ± standard error).

The artifact removal step prior to feature extraction has a significant positive impact on the classifier's performance. By mitigating the impact of false positives, it significantly increases (p < 0.001) the predictions' specificity with a relative percentage increase of +16.8 ± 5.6%, and the overall accuracy, with an average relative increase of +8.3 ± 2.9% . However, this improvement comes at the cost of a slight increase in computational time per raw recording (+12.8 ± 0.5%), which remains lower than the values computed considering the model proposed by Ciecierski and colleague [32].

ML-STIM and ResNet-AT [32] are further tested on an external dataset (Dataset B) to assess their generalizability and performance in different scenarios. Results showed the superior performance of ML-STIM in maintaining a good balance between sensitivity and specificity achieving an F_1-

Table 4. STN classifier performance on Dataset B.

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	F_1-score (%)	AUC (%)
ML-STIM	83.8 ± 1.6	73.9 ± 4.2	91.3 ± 2.2	77.1 ± 2.5	96.0 ± 1.0
ResNet-AT	77.1 ± 2.2	53.9 ± 5.6	95.1 ± 1.9	64.1± 4.2	94.8 ± 1.1
Wilcoxon's test (p-value)	**0.002**	**< 0.001**	**0.016**	**< 0.001**	0.294

Parameters' values are reported as *mean ± standard error* over the sample population. *P*-values that are below the significance level α = 0.05 are displayed in bold. ResNet-AT: Residual Neural Network with Attention in the Temporal domain; ML-STIM: Machine learning for SubThalamic nucleus Intraoperative Mapping.

score of 77.1 ± 2.5 %, while ResNet-AT achieves a F_1-score of 64.1 ± 4.2 %. The observed drop in performance across both methods could also be attributed to differences in data distributions introduced by differences in acquisition systems and annotation strategies between datasets [50]. To reduce these differences, we applied a resampling procedure to harmonize the sampling frequency of Dataset B with that of Dataset A. As shown in **Figure S.1** of the supplementary material, this helped to align most feature distributions across datasets. Nonetheless, a few features, such as pr_1_2kHz and $kurtosis$, exhibited noteworthy differences, although the relative shift between classes was preserved.

Our approach addresses several challenges that make it stand out from previously proposed methods for STN classification. While prior studies have demonstrated high classification accuracy, most have not addressed the crucial issues of real-time applicability and generalizability. Indeed, many studies adopted strategies that could go from patient-wise data normalization [33] to the employment of complex spike sorting algorithms [27], that are not suited for real-time application. Moreover, many studies train and validate their methods on complementary partitions of the same dataset, rather then on a independent dataset, raising concerns about how they would perform in different scenarios (i.e., different acquisition systems, patient cohorts, etc.). Nonetheless, recent works have started to address these limitations, for example the study by Martin and collegues [50] provides an important benchmarking effort across datasets and annotation strategies to assess generalizability, while Martin *et al.* [51] proposes a framework aimed at enabling real-time feasibility. In this context, our method was trained and validated on a publicly available dataset (Dataset A [35]) and its generalizability was tested on an 'unseen' dataset (Dataset B). This approach ensures a fairer comparison of the classification performance. Our results showed that the proposed pipeline not only achieves competitive performance on Dataset A but also generalizes effectively to Dataset B, with a relative drop in average accuracy of -4.7%. In contrast, ResNet-AT [32] shows a larger performance drop of -9.5% on Dataset B. Moreover, Dataset B is made freely available at https://doi.org/10.5281/zenodo.14894226 (accessed on 29 July 2025) to serve for testing the generalizability of previously proposed methods or to train and validate newly proposed approaches.

We chose to use traditional ML rather than more advanced Deep Learning (DL) techniques because of explainability. Unlike deep learning models which are black boxes [52], traditional ML methods extract features that give more insight into the classifier reasoning. By focusing on known STN signal characteristics such as increased background noise, reflected in increased RMS or its correlated measures (e.g., $avgAbsDiff$), or bursting spiking activity in specific frequency bands (13-20 Hz and 30-70 Hz), our pipeline is transparent and interpretable. In this way, the model's decisions can be clearly interpreted by clinicians, a crucial requirement for the application of Artificial Intelligence (AI) in medical procedures. Although DL techniques such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have demonstrated impressive classification performance, they typically require large datasets and often lack the interpretability necessary for clinical adoption. Our classification pipeline emulates the manual process of identifying key signal characteristics, effectively balancing classification performance with interpretability.

Despite the promising results, there are some limitations to acknowledge. One of the primary concerns is the dataset used for training and validation (i.e., Dataset A) [35]. As highlighted in **Table 1**, Dataset A includes data from the same patients in both training and validation sets, which may lead to overfitting and biased performance estimates. Nevertheless, to ensure comparability with the results reported by Ciecierski *et al.* [32], we decided to maintain the same training/validation split. Additionally, limitations in the test partitioning strategy also affect the distribution of performance metrics, particularly for sensitivity. It can occur that a test subject has recordings from only one class. For example, Subject 2 in the test partition of Dataset B includes only OUT-STN recordings, which leads to a null sensitivity score for that subject. Furthermore, the dataset is highly imbalanced in favour of the OUT-STN class, and while this imbalance is handled during training through the use of balanced mini-batches [32], no such correction is applied during testing. To mitigate this limitation and better assess the generalizability of our method, we further evaluated it on an independent dataset, Dataset B, which resulted in a slight reduction in performance. This approach provides a more realistic evaluation of the model's ability to perform in different clinical settings and highlights the critical importance of training AI methods on diverse datasets to avoid biases introduced by variations in acquisition systems. We also acknowledge the limitation regarding the choice of the comparative method. While several prior works [23–29] are conceptually closer to our pipeline, neither the data nor the trained models are publicly available. In contrast, ResNet-AT [32] was chosen as a baseline because both its dataset and trained model are publicly shared. Another limitation lies in the non-uniform preprocessing applied to the training and testing datasets. Our approach includes explicit band-pass filtering and artifact rejection prior to feature extraction, while ResNet-AT processes raw input signals [31]. ResNet-AT addresses lower frequency artifacts through a self-attention mechanism, whereas our method removes these artifacts explicitly through preprocessing. Moreover, since Dataset B includes a hardware high-pass filter with a cutoff at 200 Hz, the lower-frequency components emphasized by ResNet-AT are already removed at acquisition. While we

chose not to apply additional filtering to Dataset A before testing ResNet-AT in order to preserve the model's original evaluation conditions, we acknowledge that this may lead to a slight amount of information leakage and reflects a broader challenge in designing methods that are robust and generalizable across different acquisition systems.

Future research should focus on improving the generalizability and robustness of the model by enlarging the dataset to gather a greater variety of patient cohorts and acquisition devices. Furthermore, integrating the pipeline with more advanced Explainable Deep Learning techniques [52] could potentially boost performance while keeping the classification process transparent. Prospective clinical trials would also be useful in evaluating the pipeline's usefulness in real surgical settings.

5. Conclusion

In conclusion, this work offers a reliable and effective machine-learning pipeline for classifying STN activity during DBS surgeries. The system is a promising tool for improving surgical decision-making due to its high accuracy, real-time applicability, and generalizability. Results revealed that the proposed pipeline is suitable for clinical application, providing fast and accurate feedback on electrode positioning, which can substantially aid intraoperative decision-making. By reducing the time spent by trained operators in the visual and auditory analysis of MERs, this machine learning-based method can decrease patient exposure to risks and improve overall procedure safety and efficiency. In addition, this pipeline can be used as a training tool for neurosurgery residents, providing an interactive platform to learn and refine their skills in STN targeting.

Authors' contributions

Fabrizio Sciscenti: Methodology, Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review and editing, Visualization; **Valentina Agostini**: Conceptualization, Validation, Writing – original draft, Writing – review and editing, Funding acquisition; **Laura Rizzi**: Validation, Resources, Writing – review and editing; **Michele Lanotte**: Validation, Resources, Funding acquisition, Writing - review and editing; **Marco Ghislieri**: Conceptualization, Methodology, Data curation, Writing – original draft, Writing – review and editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Funding

This study was carried out within the «PD-DBS» project ("Effect of bilateral subthalamic nucleus deep brain stimulation on gait analysis and muscle synergy patterns of patients affected by Parkinson's disease during dual-task walking", protocol N° 2022KWSJJT) – funded by European Union – Next Generation EU within the PRIN 2022 program (D.D. 104 - 02/02/2022 Ministero dell'Università e della Ricerca). This manuscript reflects only the authors' views and opinions and the Ministry cannot be considered responsible for them.

Institutional Review Board Statement

This study received approval from the Ethics Committee of A.O.U. Città della Salute e della Scienza di Torino – A.O. Ordine Mauriziano – A.S.L. "Città di Torino" (Approval No. 0092029, granted on September 11, 2018). In accordance with the Declaration of Helsinki, all participants provided written informed consent for the experimental procedure.

Availability of data and materials

The test dataset and the newly proposed Python algorithm (ML-STIM) are made freely available on Zenodo (https://doi.org/10.5281/zenodo.14894226; accessed on 29 July 2025) and GitHub (https://github.com/Biolab-PoliTO/ML-STIM; accessed on 29 July 2025), respectively.

Conflict of interest

The authors declare that they have no conflict of interest.

References

[1] Feigin, V.L., Nichols, E., Alam, T., Bannick, M.S., Beghi, E., Blake, N., Culpepper, W.J., Dorsey, E.R., Elbaz, A., Ellenbogen, R.G., et al. (2019). Global, regional, and national burden of neurological disorders, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet Neurology*, 18, 459–480. https://doi.org/10.1016/S1474-4422(18)30499-X

[2] Dauer, W. and Przedborski, S. (2003). Parkinson's disease: Mechanisms and models. *Neuron*, 39, 889–909. https://doi.org/10.1016/S0896-6273(03)00568-3

[3] Fahn, S. (2003). Description of parkinson's disease as a clinical syndrome. *Annals of the New York Academy of Sciences*, 991, 1–14. https://doi.org/10.1111/j.1749-6632.2003.tb07458.x

[4] Shah, H., Usman, O., Ur Rehman, H., Jhaveri, S., Avanthika, C., Hussain, K., Islam, H., and Sailesh, I.S.K. (2022). Deep brain stimulation in the treatment of parkinson's disease. *Curēus*, 14(9), e28760. https://doi.org/10.7759/cureus.28760

[5] Benabid, A.L. (2003). Deep brain stimulation for Parkinson's disease. *Current Opinion in Neurobiology*, 13, 696–706. https://doi.org/10.1016/j.conb.2003.11.001

[6] Ghislieri, M., Lanotte, M., Knaflitz, M., Rizzi, L., and Agostini, V. (2023). Muscle synergies in Parkinson's disease before and after the deep brain stimulation of the bilateral subthalamic nucleus. *Scientific Reports*, 13, 6997. https://doi.org/10.1038/s41598-023-34151-6

[7] Ghislieri, M., Agostini, V., Rizzi, L., Fronda, C., Knaflitz, M., and Lanotte, M. (2024). Foot–Floor Contact Sequences: A Metric for Gait Assessment in Parkinson's Disease after

Deep Brain Stimulation. *Sensors*, 24, 6593. https://doi.org/10.3390/s24206593

[8] Schulder, M., Mishra, A., Mammis, A., Horn, A., Boutet, A., Blomstedt, P., Chabardes, S., Flouty, O., Lozano, A.M., Neimat, J.S., et al. (2023). Advances in technical aspects of deep brain stimulation surgery. *Stereotactic and Functional Neurosurgery*, 101, 112–134. https://doi.org/10.1159/000529040

[9] van den Munckhof, P., Bot, M., and Schuurman, P.R. (2021). Targeting of the subthalamic nucleus in patients with parkinson's disease undergoing deep brain stimulation surgery. *Neurology and Therapy*, 10, 61–73. https://doi.org/10.1007/s40120-021-00233-8

[10] Verhagen, R., Schuurman, P.R., Munckhof, P.V.D., Contarino, M.F., Bie, R.M.A.D., and Bour, L.J. (2016). Comparative study of microelectrode recording-based STN location and MRI-based STN location in low to ultra-high field (7.0 T) T2-weighted MRI images. *Journal of Neural Engineering*, 13, 066009. https://doi.org/10.1088/1741-2560/13/6/066009

[11] Neumann, W., Köhler, R.M., and Kühn, A.A. (2022). A practical guide to invasive neurophysiology in patients with deep brain stimulation. *Clinical Neurophysiology*, 140, 171–180. https://doi.org/10.1016/j.clinph.2022.05.004

[12] Wan, K.R., Maszczyk, T., See, A.A.Q., Dauwels, J., and King, N.K.K. (2019). A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease. *Clinical Neurophysiology*, 130, 145–154. https://doi.org/10.1016/j.clinph.2018.09.018

[13] Peralta, M., Jannin, P., and Baxter, J.S. (2021). Machine learning in deep brain stimulation: A systematic review. *Artificial Intelligence in Medicine*, 122, 102198. https://doi.org/10.1016/j.artmed.2021.102198

[14] Bakštein, E., Sieger, T., Wild, J., Novák, D., Schneider, J., Vostatek, P., Urgošík, D., and Jech, R. (2017). Methods for automatic detection of artifacts in microelectrode recordings. *Journal of Neuroscience Methods*, 290, 39–51. https://doi.org/10.1016/j.jneumeth.2017.07.012

[15] Klempíř, O., Krupička, R., Bakštein, E., and Jech, R. (2019). Identification of microrecording artifacts with wavelet analysis and convolutional neural network: An image recognition approach. *Measurement Science Review*, 19, 222–231. https://doi.org/10.2478/msr-2019-0029

[16] Koirala, N., Serrano, L., Paschen, S., Falk, D., Anwar, A.R., Kuravi, P., Deuschl, G., Groppa, S., and Muthuraman, M. (2020). Mapping of subthalamic nucleus using microelectrode recordings during deep brain stimulation. *Scientific Reports*, 10. https://doi.org/10.1038/s41598-020-74196-5

[17] Fabietti, M., Mahmud, M., Lotfi, A., Kaiser, M.S., Averna, A., Guggenmos, D.J., Nudo, R.J., Chiappalone, M., and Chen, J. (2021). SANTIA: a Matlab-based open-source toolbox for artifact detection and removal from extracellular neuronal signals. *Brain Informatics*, 8. https://doi.org/10.1186/s40708-021-00135-3

[18] Gorlini, C., Forzanini, F., Coelli, S., Rinaldo, S., Eleopra, R., Bianchi, A.M., and Levi, V. (2024). Impact of Microelectrode Recording Artefacts on Subthalamic Nucleus Functional Identification via Features-Based Machine Learning Classifiers. In: *2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE). IEEE*, St Albans, United Kingdom, pp 13–18. https://doi.org/10.1109/MetroXRAINE62247.2024.10796772

[19] Tepper, Á., Henrich, M.C., Schiaffino, L., Muñoz, A.R., Gutiérrez, A., and Martínez, J.G. (2017). Selection of the optimal algorithm for real-time estimation of beta band power during DBS surgeries in patients with Parkinson's disease. *Computational Intelligence and Neuroscience*, 2017, 1–9. https://doi.org/10.1155/2017/1512504

[20] Valsky, D., Marmor-Levin, O., Deffains, M., Eitan, R., Blackwell, K.T., Bergman, H., and Israel, Z. (2017). Stop! border ahead: Automatic detection of subthalamic exit during deep brain stimulation surgery. *Movement Disorders*, 32, 70–79. https://doi.org/10.1002/mds.26806

[21] Karthick, P.A., Wan, K.R., Qi, A.S.A., Dauwels, J., and King, N.K.K. (2020). Automated detection of subthalamic nucleus in deep brain stimulation surgery for parkinson's disease using microelectrode recordings and wavelet packet features. *Journal of Neuroscience Methods*, 343, 108826. https://doi.org/10.1016/j.jneumeth.2020.108826

[22] Hosny, M., Zhu, M., Gao, W., and Fu, Y. (2020). A novel deep LSTM network for artifacts detection in microelectrode recordings. *Biocybernetics and Biomedical Engineering*, 40, 1052–1063. https://doi.org/10.1016/j.bbe.2020.04.004

[23] Hosny, M., Zhu, M., Gao, W., and Fu, Y. (2021). Detection of subthalamic nucleus using novel higher-order spectra features in microelectrode recordings signals. *Biocybernetics and Biomedical Engineering*, 41, 704–716. https://doi.org/10.1016/j.bbe.2021.04.016

[24] Schiaffino, L., Muñoz, A.R., Martínez, J.G., Villora, J.F., Gutiérrez, A., Torres, I.M., and Kohan, Y.D.R. (2016). STN area detection using K-NN classifiers for MER recordings in Parkinson patients during neurostimulator implant surgery. *Journal of Physics: Conference Series*, 705, 012050. https://doi.org/10.1088/1742-6596/705/1/012050

[25] Bellino, G.M., Schiaffino, L., Battisti, M., Guerrero, J., and Rosado-Muñoz, A. (2019). Optimization of the KNN supervised classification algorithm as a support tool for the implantation of deep brain stimulators in patients with Parkinson'S Disease. *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, 21, 346. https://doi.org/10.3390/e21040346

[26] Chaovalitwongse, W., Jeong, Y., Jeong, M.K., Danish, S., and Wong, S. (2011). Pattern recognition approaches for identifying subcortical targets during deep brain stimulation surgery. *IEEE Intelligent Systems*, 26, 54–63. https://doi.org/10.1109/MIS.2011.56

[27] Coelli, S., Levi, V., Del Vecchio Del Vecchio, J., Mailland, E., Rinaldo, S., Eleopra, R., and Bianchi, A.M. (2021). An intra-operative feature-based classification of microelectrode recordings to support the subthalamic nucleus functional identification during deep brain stimulation surgery. *Journal of Neural Engineering*, 18, 016003. https://doi.org/10.1088/1741-2552/abcb15

[28] Valsky, D., Blackwell, K.T., Tamir, I., Eitan, R., Bergman, H., and Israel, Z. (2020). Real-time machine learning classification of pallidal borders during deep brain stimulation surgery. *Journal of Neural Engineering*, 17, 016021. https://doi.org/10.1088/1741-2552/ab53ac

[29] Khosravi, M., Atashzar, S.F., Gilmore, G., Jog, M.S., and Patel, R.V. (2020). Intraoperative localization of STN during DBS surgery using a data-driven model. *IEEE Journal of Translational Engineering in Health and*

Medicine, 8, 1–9. https://doi.org/10.1109/JTEHM.2020.2969152

[30] Xiao, L., Li, C., Wang, Y., Si, W., Zhang, D., Lin, H., Cai, X., and Heng, P.A. (2021). Automatic identification of sweet spots from MERs for electrodes implantation in STN-DBS. *International Journal of Computer Assisted Radiology and Surgery*, 16, 809–818. https://doi.org/10.1007/s11548-021-02377-2

[31] Martin, T., Peralta, M., Gilmore, G., Sauleau, P., Haegelen, C., Jannin, P., and Baxter, J.S. (2021). Extending convolutional neural networks for localizing the subthalamic nucleus from micro-electrode recordings in Parkinson's disease. *Biomedical Signal Processing and Control*, 67, 102529. https://doi.org/10.1016/j.bspc.2021.102529

[32] Ciecierski, K. and Mandat, T. (2024). Classification of DBS microelectrode recordings using a residual neural network with attention in the temporal domain. *Neural Networks*, 170, 18–31. https://doi.org/10.1016/j.neunet.2023.11.021

[33] Maged, A., Zhu, M., Gao, W., and Hosny, M. (2024). Lightweight deep learning model for automated STN localization using MER in Parkinson's disease. *Biomedical Signal Processing and Control*, 96, 106640. https://doi.org/10.1016/j.bspc.2024.106640

[34] Hosny, M., Zhu, M., Gao, W., and Fu, Y. (2022). A novel deep learning model for STN localization from LFPs in Parkinson's disease. *Biomedical Signal Processing and Control*, 77, 103830. https://doi.org/10.1016/j.bspc.2022.103830

[35] Ciecierski, K. and Mandat, T.(2023)RAW deep brain stimulation recordings [dataset] https://doi.org/10.17632/tr93krswn2.1

[36] Defer, G., Widner, H., Marié, R., Rémy, P., Levivier, M., cowriters for the dyskinesia/dystonia, C., and imaging sections, R.P.C. (1999). Core assessment program for surgical interventional therapies in Parkinson's disease (CAPSIT-PD). *Movement Disorders*, 14, 572–584. https://doi.org/10.1002/1531-8257(199907)14:4<572::AID-MDS1005>3.0.CO;2-C

[37] Lanotte, M., Rizzone, M., Bergamasco, B., Faccani, G., Melcarne, A., and Lopiano, L. (2002). Deep brain stimulation of the subthalamic nucleus: anatomical, neurophysiological, and outcome correlations with the effects of stimulation. *Journal of Neurology, Neurosurgery & Psychiatry*, 72, 53–58. https://doi.org/10.1136/jnnp.72.1.53

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

[39] Pesenti, A., Rohr, M., Egidi, M., Rampini, P., Tamma, F., Locatelli, M., Caputo, E., Chiesa, V., Bianchi, A., Barbieri, S., et al. (2004). The subthalamic nucleus in Parkinson?s disease: power spectral density analysis of neural intraoperative signals. *Neurological Sciences*, 24, 367–374. https://doi.org/10.1007/s10072-003-0191-2

[40] Cagnan, H., Dolan, K., He, X., Contarino, M.F., Schuurman, R., Munckhof, P.V.D., Wadman, W.J., Bour, L., and Martens, H.C.F. (2011). Automatic subthalamic nucleus detection from microelectrode recordings based on noise level and neuronal activity. *Journal of Neural Engineering*, 8, 046006. https://doi.org/10.1088/1741-2560/8/4/046006

[41] Rajpurohit, V., Danish, S.F., Hargreaves, E.L., and Wong, S. (2015). Optimizing computational feature sets for subthalamic nucleus localization in DBS surgery with feature selection. *Clinical Neurophysiology*, 126, 975–982. https://doi.org/10.1016/j.clinph.2014.05.039

[42] Thompson, J.A., Oukal, S., Bergman, H., Ojemann, S., Hebb, A.O., Hanrahan, S., Israel, Z., and Abosch, A. (2019). Semi-automated application for estimating subthalamic nucleus boundaries and optimal target selection for deep brain stimulation implantation surgery. *Journal of Neurosurgery*, 130, 1224–1233. https://doi.org/10.3171/2017.12.JNS171964

[43] Cao, L., Li, J., Zhou, Y., Liu, Y., Zhao, Y., and Liu, H. (2019). Online identification of functional regions in deep brain stimulation based on an unsupervised random forest with feature selection. *Journal of Neural Engineering*, 16, 066015. https://doi.org/10.1088/1741-2552/ab2eb4

[44] Cao, L., Jie, L., Zhou, Y., Liu, Y., and Liu, H. (2020). Automatic feature group combination selection method based on GA for the functional regions clustering in DBS. *Computer Methods and Programs in Biomedicine*, 183, 105091. https://doi.org/10.1016/j.cmpb.2019.105091

[45] Moran, A. and Bar-Gad, I. (2010). Revealing neuronal functional organization through the relation between multi-scale oscillatory extracellular signals. *Journal of Neuroscience Methods*, 186, 116–129. https://doi.org/10.1016/j.jneumeth.2009.10.024

[46] Ding, C. and Peng, H. (2003). Minimum redundancy feature selection from microarray gene expression data. In: *Proceedings of the 2003 IEEE Bioinformatics Conference. CSB2003*, Stanford, CA, USA, 2003, pp. 523-528. https://doi.org/10.1109/CSB.2003.1227396

[47] Kurita, T., Otsu, N., and Abdelmalek, N. (1992). Maximum likelihood thresholding based on population mixture models. *Pattern Recognition*, 25, 1231–1240. https://doi.org/10.1016/0031-3203(92)90024-D

[48] Robnik-Sikonja, M. and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53, 23–69. https://doi.org/10.1023/A:1025667309714

[49] Hedges, L.V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, 6, 107–128. https://doi.org/10.3102/10769986006002107

[50] Martin, T., Jannin, P., and Baxter, J.S.H. (2024). Generalisation capabilities of machine-learning algorithms for the detection of the subthalamic nucleus in micro-electrode recordings. *International Journal of Computer Assisted Radiology and Surgery*, 19, 2445–2451. https://doi.org/10.1007/s11548-024-03202-2

[51] Martin, T., Gilmore, G., Haegelen, C., Jannin, P., and Baxter, J.S.H. (2021). Adapting the listening time for micro-electrode recordings in deep brain stimulation interventions. *International Journal of Computer Assisted Radiology and Surgery*, 16, 1371–1379. https://doi.org/10.1007/s11548-021-02379-0

[52] Loh, H.W., Ooi, C.P., Seoni, S., Barua, P.D., Molinari, F., and Acharya, U.R. (2022). Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022). *Computer Methods and Programs in Biomedicine*, 226, 107161. https://doi.org/10.1016/j.cmpb.2022.107161