



**HAL**  
open science

# Machine learning in deep brain stimulation: A systematic review

M. Peralta, Pierre Jannin, J.S.H. Baxter

► **To cite this version:**

M. Peralta, Pierre Jannin, J.S.H. Baxter. Machine learning in deep brain stimulation: A systematic review. *Artificial Intelligence in Medicine*, 2021, 122, pp.102198. 10.1016/j.artmed.2021.102198 . hal-03468517

**HAL Id: hal-03468517**

**<https://hal.science/hal-03468517>**

Submitted on 8 Dec 2021

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

# Machine learning in deep brain stimulation: a systematic review<sup>★</sup>

Maxime Peralta<sup>a</sup>, Pierre Jannin<sup>a</sup> and John S. H. Baxter<sup>a,\*</sup>

<sup>a</sup>*Univ Rennes, Inserm, LTSI - UMR 1099, F-35000 Rennes, France*

## ARTICLE INFO

### *Keywords:*

Systematic review

Deep brain stimulation

Machine learning

## ABSTRACT

Deep Brain Stimulation (DBS) is an increasingly common therapy for a large range of neurological disorders, such as abnormal movement disorders. The effectiveness of DBS in terms of controlling patient symptomatology have made this procedure increasingly used over the past few decades. Concurrently, the popularity of Machine Learning (ML), a subfield of artificial intelligence, has skyrocketed and its influence has more recently extended to medical domains such as neurosurgery. Despite its growing research interest, there has yet to be a literature review specifically on the use of ML in DBS. We have followed a fully systematic methodology and to obtain a corpus of 73 papers. In each paper, we identified the clinical application, the type/amount

---

<sup>★</sup> This document is the result of the research project funded by the Fondation pour la Recherche Médicale.

\* Corresponding author

✉ [jbaxter@univ-rennes1.fr](mailto:jbaxter@univ-rennes1.fr) (J.S.H.B.)

ORCID (s): 0000-0001-9799-423X (M. Peralta); 0000-0002-7415-071X (P. Jannin); 0000-0003-3548-4343 (J.S.H.B.)

of data used, the method employed, and the validation strategy of each paper, further decomposed into 12 different sub-categories. The papers overall illustrated some existing trends in how ML is used in the context of DBS, including the breath of the problem domain and evolving techniques, as well as common frameworks and limitations. This systematic review analyzes at a broad level how ML have been recently used to address clinical problems on DBS, giving insight into how these new computational methods are helping to push the state-of-the-art of functional neurosurgery. DBS clinical workflow is complex, involves many specialists, and raises several clinical issues which have partly been addressed with artificial intelligence. However, several areas remain and those that have been recently addressed with ML are by no means considered “solved” by the community nor are they closed to new and evolving methods.

## **1. Introduction**

Deep Brain Stimulation (DBS) is a neurosurgical procedure, introduced in 1987 by Pr. Benabid [1], in which electrodes are implanted into deep regions of the brain to correct for abnormal neural behavior. Continuous stimulation of these regions typically greatly enhances the quality of life of the patient by reducing the severity their symptoms. Parkinson’s Disease (PD), Essential Tremor (ET), dystonia, Tourette’s syndrome or Obsessive Compulsive Disorders (OCD) are all among the pathologies now commonly treated using DBS, and the most commonly targeted structures are the bilateral Ventral Intermediate nucleus of the thalamus (VIM), Globus Pallidus internus (GPi) and Subthalamic Nucleus (STN).

DBS interventions have a complex clinical workflow, involving several steps (presented in Figure 1) that involve many challenges, both clinical challenges and challenges for the use of computer assistance. A large amount of expertise and domain knowledge is crucial for the success

of this procedure, and computer-assisted tools, referred to as Computational Decision Support System (CDSS), have been designed to support clinicians in neurosurgery since the 1980s [2]. CDSS's are commonly divided in two categories: knowledge-based and non-knowledge-based. In knowledge-based CDSS's, the tool's intelligence is explicitly integrated into the system by a human, often the system programmer. Decision rules are explicitly programmed according to medical domain knowledge such as guidelines and definitions [3], and the purpose of the CDSS is solely to retrieve the data, to evaluate the rule and to display the result with an User Interface (UI). In non-knowledge-based CDSS's, machine learning replaces expert medical knowledge in order to address new challenges and to reach new levels of performance: the intelligence is implicitly generated by learning from a database using Machine Learning (ML) tools designed by data scientists, without requiring external domain knowledge.

**Figure 1:** Surgical workflow of a typical DBS, composed of four major steps.

The term 'machine learning' appeared for the first time in 1959 in the works of Samuel *et al.* [4]. It is a branch of Artificial Intelligence (AI) which consists in constructing an algorithm that learns how to perform a task using a database of experiences without requiring any explicit programming of the user or knowledge of the task. On its most common form, called supervised learning, it consists in predicting an output from a set of inputs, called features. To model the link between the inputs and the output, the algorithm processes a dataset of multiple known pairs of inputs and output. If the training has been successful, the trained model should predict the unknown output of a new sample from its known inputs. The other common form of machine learning is called unsupervised learning, where there is no specific output to predict. In this case,

the model performs a task, such as clustering, based solely on the set of inputs without an explicit function determining the applicability of the learned representations to the larger problem. Thus, these unsupervised methods require some additional programming to apply the results of the ML model to a particular clinical problem

One common use of ML models in CDSS's is to automate a process that would otherwise be performed by a human operator, such as determining the location of a DBS electrode for example. Another use is to predict the outcome of the stimulation itself, allowing the clinical team to explore a broader array of alternatives in a non-invasive manner.

The interest of ML to assist clinicians in healthcare has been underlined for a long time. Celtikci *et al.* [5] and Bulchlak *et al.* [6] conducted systematic reviews of the utilization of ML to assist the decision-making in neurosurgery. They highlighted how ML can outperform some traditional statistical methods for the analysis of retrospective data, for example by leveraging non-linearities in high-dimensional and large databases. Senders *et al.* [7] also conducted a review of the utilization of ML in neurosurgery. They also underlined the applicability of ML in neurosurgical care, notably by saying that “In the last few decades, the volume and complexity of bio-medical data have grown beyond the physician’s ability to extract all meaningful data patterns using conventional statistical methods alone. [...] The complex diagnostic and therapeutic modalities used in neurosurgery provide a vast amount of data that is ideally suited for ML models.” [7] In another review, they compared the performance of ML to human experts for diagnosis, surgical planning, and outcome prediction in neurosurgery [8]. They concluded that machine intelligence had overall superior results, but pointed out a publication bias which tends to overestimate the performance of ML, as negative results are less likely to be published.

These reviews have elucidated the current and potential use cases of ML in neurosurgery,

highlighting interesting ML applications, but none of them focused specifically on DBS. The use of ML in DBS is a broad area of research as the methods, data modalities and clinical problems addressed are numerous. In order to draw a landscape of this research field, extracting major trends, better identifying recurring methodological limits, we conducted a systematic review focused on ML in DBS, answering the broad question of how ML is currently used to address clinical problems in DBS.

The next sections will present the methodology used to compile a corpus of paper to analyse, the data acquired from each paper, the results of said analysis, and a discussion about these results.

**Figure 2:** Workflow to select the corpus of 73 papers to classify.

## 2. Material and Methods

We performed this systematic review by following the Preferred Reporting Items for Systematic review and Meta-Analysis Protocols (PRISMA-P) [9] relevant and applicable recommendations. Figure 2 presents the workflow used to select papers for further analysis.

### Initial literature selection strategy

We researched relevant papers in the literature through three queries on two search engines, on September 18th 2021:

- PubMed, with MeSH method, with the following query: (“Machine Learning”[Mesh] OR “Artificial Intelligence”[MeSH:noexp] OR “Neural Networks, Computer”[MeSH]) AND “Deep Brain Stimulation”[Mesh]

- Google Scholar, with the following query: (“machine learning” OR “deep learning” OR “neural networks” OR “data-driven” OR “learning-based” OR “artificial intelligence”)(“deep brain stimulation”)
- Google Scholar, with the following query: (“prediction”)(“deep brain stimulation”)

We chose to use PubMed with MeSH method as it is a proven way of browsing papers in medical research, providing that they are tagged with the appropriate MeSH terms. We did not add the MeSH term ‘Deep Learning’ to the query as it is already included in the ‘Machine Learning’ hierarchy. We turned off automatic explosion of the MeSH term ‘Artificial Intelligence’ in order not to include other unrelated topics such as ‘Robotics’. We made two different queries on Google Scholar as it is more comprehensive than PubMed. The first one was composed of targeted keywords, and the second one consisted of a broader term (‘prediction’) in order to include additional papers that could have been missed by the first two queries. Due to the high number of papers returned by queries on Google Scholar, we analyzed the results page by page and stopped when we retained no new papers on two consecutive pages after title and abstract screening (the results being sorted by relevance). We decided not to merge both Google Scholar queries into a single as the term ‘prediction’ is broad and returned a lot of irrelevant papers. Therefore, including this term in the first Google Scholar query could have made relevant items sparser.

In order to keep a fully systematic methodology, to avoid flaws in the results and to make our screening method reproducible, we chose not to manually include additional papers in the corpus.

## **Selection process**

The first author screened each paper by reading the title and abstract with the following

criteria:

- The paper must be methodological, i.e. validate at least one method.
- ML must be at the heart of the methodology employed.
- The paper must be validated on patients. Papers validated with synthetic data, or using a non-human cohort were discarded.
- It must address a clearly identified clinical problem.
- The paper must be peer-reviewed. If we couldn't obtain the published version, the pre-print version was used. Thesis manuscripts and reviews were discarded.
- It must be written in English.

The number of papers returned by each query and the number of papers kept after title and abstract screening is indicated in Figure 2. We merged the results provided by the three queries, removed duplicates, and obtained a corpus of 117 papers. The first author then re-screened each paper by reading the full text according to the same selection criteria. 44 papers were discarded (notably six papers for being a preliminary version of another retained paper).

**Figure 3:** Data was acquired concerning four classes: data used, clinical application, method, and validation. These classes can be composed of several items.

### **Data obtained from each paper**

Each of the 73 papers in the final corpus was described according to four classes, as presented on Figure 3. The 'data' class describes the cohort used in the experiment, as well as the nature of the inputs of the ML model. We evaluated the following items:

- The input data modality type, such as imaging or Micro-electrode Recordings (MER).



- The number of patients in the cohort.
- The pathology of the patients in the cohort, such as PD or ET.

The ‘application’ class corresponds to the clinical problem addressed. We evaluated the following items:

- The stage of the DBS workflow this problem is encountered, using the instances ‘screening’, ‘planning’, ‘surgery’ and ‘post-op’, as showed in Figure 1.
- The nature of the task that has to be addressed by the method, such as ‘classification’, ‘segmentation’, ‘regression’ or ‘clustering’.

The ‘method’ class describes the specific ML algorithm or framework employed in the paper and is composed of the following items:

- The method used to handle input data, decomposed into three sub-items: the data compression method used (if any), whether or not an automatic feature selection method was employed, and if the method is feature-based in which the input data was transformed into a synthetic set of features requiring a significant amount of feature engineering. For example, we did not consider common domain transformations (such as the Fourier or wavelet domains) as feature-based unless additional operations were performed on said domain which would require additional domain knowledge (such as the selection of particular frequency bands).
- The ML model used to perform the task (ie. the classification, the regression, etc.). If several models were benchmarked, we only reported the one(s) giving the better results or the one(s) highlighted in the paper’s abstract, discussion and conclusion sections.

The ‘validation’ class describes how the methods were evaluated, according to the following items:

- The method employed to split the data between training and testing sets, such as ‘hold-out’ or ‘LOOCV’.
- Whether or not the validation method is performed in a patient-wise manner, implying that data collected from a single patient cannot be simultaneously in the training set and the testing set.
- The primary metric used to evaluate the performance of the method(s). If several metrics were used, we reported the one highlighted in the paper’s abstract, discussion and conclusion sections, or the one the most extensively used in the experiments.

### 3. Results

Data obtained from each study was recorded on an Excel spreadsheet and is reproduced in Tables 1 and 2.

Paper	Data		Application		
	Input Modality	C. size	Pathology	Phase	Task
Orozco <i>et al.</i> (2006) [10]	MER		PD	surgery	class.
Muniz <i>et al.</i> (2009) [11]	external sensors	45	PD	post-op.	class.
Wong <i>et al.</i> (2009) [12]	MER	27	PD	surgery	clust.
Wu <i>et al.</i> (2010)	LFP	1	PD	post-op.	class.

[13]					
Guillén <i>et al.</i> (2011) [14]	MER	4	PD	surgery	class.
Shukla <i>et al.</i> (2012) [15]	external sensors	2	PD	post-op.	class.
Loukas <i>et al.</i> (2012) [16]	LFP	1	PD	post-op.	class.
Jiang <i>et al.</i> (2013) [17]	LFP	9	PD	post-op.	class.
Niketeghad <i>et al.</i> (2014) [18]	LFP	9	PD	post-op.	class.
Connolly <i>et al.</i> (2015) [19]	LFP	15	PD	post-op.	class.
Shamir <i>et al.</i> (2015) [20]	clinical, imaging, stimulation, MER	10	PD	post-op.	class.
Rajpurohit <i>et al.</i> (2015) [21]	MER	26	PD	surgery	class.
Kim <i>et al.</i> (2015) [22]	imaging	46	PD	planning	reg.
Khobragade <i>et al.</i> (2015) [23]	external sensors	1	PD	post-op.	class.

Yohanandan <i>et al.</i> (2016) [24]	external sensors	9	ET	post-op.	class.
Baumgarten <i>et al.</i> (2016) [25]	imaging, stimulation	10	PD	planning	class.
Liu <i>et al.</i> (2016) [26]	imaging	100	PD	planning	reg.
Kostoglou <i>et al.</i> (2016) [27]	clinical, imaging, MER	20	PD	surgery	class.
Guillén <i>et al.</i> (2016) [28]	MER	5	PD	surgery	class.
Baumgarten <i>et al.</i> (2016) [29]	imaging, stimulation	20	PD	planning	class.
Angeles <i>et al.</i> (2017) [30]	external sensors	7	PD	post-op.	class.
Houston <i>et al.</i> (2017) [31]	ECoG	1	ET	post-op.	class.
Milletari <i>et al.</i> (2017) [32]	imaging	89	N/C	planning	seg.
Valsky <i>et al.</i> (2017) [33]	MER	81	PD	surgery	reg.
Mohammed <i>et al.</i> (2017) [34]	LFP	9	PD	post-op.	class.
Golshan <i>et al.</i>	LFP	9	PD	post-op.	class.

(2018) [35]					
Baumgarten <i>et al.</i> (2018) [36]	imaging, stimulation	30	PD	planning	class.
Khosravi <i>et al.</i> (2018) [37]	MER	20	PD	surgery	class.
LeMoyne <i>et al.</i> (2018) [38]	external sensors	1	PD	post-op.	class.
Cardona <i>et al.</i> (2018) [39]	MER	5,4	PD	surgery	class.
Khobragade <i>et al.</i> (2018) [40]	external sensors	2	mixed	post-op.	class.
Oliveira <i>et al.</i> (2018) [41]	external sensors	38	PD	inc., post.	class.
Shah <i>et al.</i> (2018) [42]	LFP	7	PD	post-op.	class.
Yao <i>et al.</i> (2018) [43]	LFP	12	PD	post-op.	class.
Golshan <i>et al.</i> (2018) [44]	LFP	3	PD	post-op.	class.
Wang <i>et al.</i> (2018) [45]	LFP	12	PD	post-op.	class.
Houston <i>et al.</i> (2018) [46]	ECoG	3	ET	post-op.	class.

Koch <i>et al.</i> (2019) [47]	EEG	40	PD	screening	class.
Kim <i>et al.</i> (2019) [48]	imaging	80	PD	planning	seg.
Chen <i>et al.</i> (2019) [49]	LFP	12	PD	post-op.	class.
Tan <i>et al.</i> (2019) [50]	LFP	7	ET	post-op.	class.
Park <i>et al.</i> (2019) [51]	imaging	102	mixed	planning	seg.
LeMoyne <i>et al.</i> (2019) [52]	external sensors	1	ET	post-op.	class.
Klempíř <i>et al.</i> (2019) [53]	MER	58	PD	surgery	class.
Stuart <i>et al.</i> (2019) [54]	EEG	16	mixed	post-op.	class.
Habets <i>et al.</i> (2019) [55]	clinical	90	PD	screening	class.
Camara <i>et al.</i> (2019) [56]	LFP	4	PD	post-op.	class.
Singer <i>et al.</i> (2019) [57]	clinical, imaging	114	PD	planning	reg.
Bermudez <i>et al.</i>	imaging	187	mixed	planning	class.

(2019) [58]					
Ciecierski <i>et al.</i>	MER	115	PD	surgery	clust.
(2019) [59]					
Mohammed <i>et al.</i>	LFP	9	PD	post-op.	class.
(2020) [60]					
Hosny <i>et al.</i>	MER	17	PD	surgery	class.
(2020) [61]					
Farrokhi <i>et al.</i>	clinical	501	mixed	screening	class.
(2020) [62]					
Valsky <i>et al.</i>	MER	42	mixed	surgery	class.
(2020) [63]					
Baxter <i>et al.</i>	imaging	9	N/C	planning	reg.
(2020) [64]					
Shah <i>et al.</i>	clinical, imaging, TMS	133	dystonia	screening	class.
(2020) [65]					
Shang <i>et al.</i>	imaging	50	PD	screening	reg.
(2020) [66]					
Golshan <i>et al.</i>	LFP	10	PD	post-op.	class.
(2020) [67]					
Khosravi <i>et al.</i>	MER	100	PD	surgery	class.
(2020) [68]					
Lu <i>et al.</i> (2020)	MER	16	PD	surgery	class.
[69]					

Peralta <i>et al.</i> (2020) [70]	MER	57	PD	surgery	class.
Yao <i>et al.</i> (2020) [71]	LFP	12	PD	post-op.	class.
Karthick <i>et al.</i> (2020) [72]	MER	26	PD	surgery	class.
Park <i>et al.</i> (2021) [73]	MER	34	PD	surgery	class.
Peralta <i>et al.</i> (2021) [74]	clinical, imaging	196	PD	screening	reg.
Boutet <i>et al.</i> (2021) [75]	imaging	67	PD	post-op.	class.
Hosny <i>et al.</i> (2021) [76]	MER	21	PD	surgery	class.
Martin <i>et al.</i> (2021) [77]	MER	57	PD	surgery	class.
Martin <i>et al.</i> (2021) [78]	MER	57	PD	surgery	class.
Baxter <i>et al.</i> (2021) [79]	imaging	10	N/C	planning	seg.
Geraedts <i>et al.</i> (2021) [80]	EEG	40	PD	screening	class.
Liebrand <i>et al.</i>	imaging	57	OCD	screening	class & reg.



(2021)[81]					
Solomon <i>et al.</i>	imaging	101	mixed	planning	seg.
(2021) [82]					

**Table 1**

Data obtained from each of the 55 papers in the corpus for the ‘data’ and ‘application’ classes.

Paper	Method		Data splitting	Patient-wise?	Validation
	Dim. Red.	Model			Metrics
Orozco <i>et al.</i> (2006) [10]	PCA, FB	HMM	LOOCV	no/not specified	accuracy
Muniz <i>et al.</i> (2009) [11]	PCA, FS	ANN	hold-out	yes	AUC-ROC
Wong <i>et al.</i> (2009) [12]	FB	unsupervised	hold-out	yes	MAE
Wu <i>et al.</i> (2010) [13]	PCA, FB	ANN	hold-out	no/not specified	accuracy
Guillén <i>et al.</i> (2011) [14]	FB	SVM	k-CV	no/not specified	accuracy, kappa
Shukla <i>et al.</i> (2012) [15]	FB	ANN	hold-out	no/not specified	accuracy, sensitivity
Loukas <i>et al.</i> (2012) [16]	FB	ANN	none	no/not specified	accuracy
Jiang <i>et al.</i> (2013) [17]	FB	HMM	hold-out	no/not specified	accuracy

Niketeghad <i>et al.</i> (2014) [18]	PCA	SVM, kNN	k-CV	no/not specified	accuracy
Connolly <i>et al.</i> (2015) [19]	FB, FS	SVM	LOOCV	no/not specified	nb. errors
Shamir <i>et al.</i> (2015) [20]	FB	SVM, RF, EL, NB	LOOCV	no/not specified	accuracy
Rajpurohit <i>et al.</i> (2015) [21]	FB, FS	kNN	LOOCV	yes	accuracy
Kim <i>et al.</i> (2015) [22]	FB, FS	EL	hold-out	yes	MSE
Khobragade <i>et al.</i> (2015) [23]	FB	ANN	hold-out	no/not specified	R-ratio, sensitivity, accuracy
Yohanandan <i>et al.</i> (2016) [24]	FB	RF	k-CV	no/not specified	kappa
Baumgarten <i>et al.</i> (2016) [25]	none	ANN	k-CV	no/not specified	kappa
Liu <i>et al.</i> (2016) [26]	FB	RF	k-CV	yes	MAE
Kostoglou <i>et al.</i> (2016) [27]	FB, FS	RF	bootstrapping	yes	MCC
Guillén <i>et al.</i> (2016) [28]	FB	ANN	hold-out	no/not specified	accuracy

Baumgarten <i>et al.</i> (2016) [29]	none	ANN	LOOCV	yes	sensitivity, specificity, precision, NPV, kappa
Angeles <i>et al.</i> (2017) [30]	FB	kNN	k-CV	no/not specified	accuracy
Houston <i>et al.</i> (2017) [31]	FB	LogReg	k-CV	no/not specified	accuracy, sensitivity, specificity
Milletari <i>et al.</i> (2017) [32]	none	CNN	hold-out	yes	dice, failure rate, CMD
Valsky <i>et al.</i> (2017) [33]	FB, FS	SVM, HMM	hold-out	yes	MAE
Mohammed <i>et al.</i> (2017) [34]	MRM	kNN	hold-out	no/not specified	F1
Golshan <i>et al.</i> (2018) [35]	PCA	SVM, EL	LOOCV	no/not specified	accuracy
Baumgarten <i>et al.</i> (2018) [36]	none	ANN	LOOCV	yes	sensitivity, specificity
Khosravi <i>et al.</i> (2018) [37]	none	SVM	k-CV	no/not specified	accuracy
LeMoyne <i>et al.</i> (2018) [38]	FB	ANN	k-CV	no/not specified	accuracy

Cardona <i>et al.</i> (2018) [39]	FB	GPR	hold-out	no/not specified	accuracy, AUC-ROC
Khobragade <i>et al.</i> (2018) [40]	FB	ANN	hold-out	no/not specified	accuracy, sensitivity, beta ratio
Oliveira <i>et al.</i> (2018) [41]	t-SNE,FB	SVM	hold-out	no/not specified	accuracy, AUC-ROC
Shah <i>et al.</i> (2018) [42]	FB	LogReg	k-CV	no/not specified	AUC-ROC
Yao <i>et al.</i> (2018) [43]	FB, FS	XGB	k-CV	no/not specified	AUC-ROC
Golshan <i>et al.</i> (2018) [44]	PCA	SVM	not specified	no/not specified	accuracy, precision, sensitivity, AUC-ROC
Wang <i>et al.</i> (2018) [45]	FB	LDA	hold-out	no/not specified	accuracy, sensitivity
Houston <i>et al.</i> (2018) [46]	none	LogReg	k-CV	no/not specified	accuracy, sensitivity
Koch <i>et al.</i> (2019) [47]	FB, FS	RF	k-CV	yes	accuracy, AUC-ROC, precision, sensitivity

Kim <i>et al.</i> (2019) [48]	FB	RF	LOOCV	yes	dice, MSD, MAE
Chen <i>et al.</i> (2019) [49]	FB, FS	SVM	hold-out	no/not specified	accuracy, sensitivity, specificity
Tan <i>et al.</i> (2019) [50]	FB	LogReg	hold-out	no/not specified	AUC-ROC
Park <i>et al.</i> (2019) [51]	none	CNN	hold-out	yes	accuracy, Dice, IoU
LeMoyne <i>et al.</i> (2019) [52]	FB	SVM, HMM, kNN, LogReg	k-CV	no/not specified	accuracy
Klempř <i>et al.</i> (2019) [53]	none	CNN	hold-out	no/not specified	accuracy, MCC, AUC-ROC
Stuart <i>et al.</i> (2019) [54]	PCA, FB, FS	SVM, RF	k-CV	yes	accuracy, F1, precision, sensitivity
Habets <i>et al.</i> (2019) [55]	none	LogReg	k-CV	yes	AUC-ROC, accuracy, sensitivity, FPR
Camara <i>et al.</i> (2019) [56]	FB	SVM	k-CV	no/not specified	accuracy, sensitivity,

					specificity
Singer <i>et al.</i> (2019) [57]	FB, FS	SVM	hold-out	yes	MAE
Bermudez <i>et al.</i> (2019) [58]	none	CNN	k-CV	yes	AUC-ROC, sensitivity, specificity
Ciecierski <i>et al.</i> (2019) [59]	FB	unsupervised	training set	no/not specified	sensitivity, specificity
Mohammed <i>et al.</i> (2020) [60]	FS	SVM	hold-out	no/not specified	MCC, WCE
Hosny <i>et al.</i> (2020) [61]	FB	LSTM	hold-out	yes	accuracy, sensitivity, specificity
Farrokhi <i>et al.</i> (2020) [62]	FS	XGB	hold-out	yes	AUC-ROC
Valsky <i>et al.</i> (2020) [63]	FB	HMM	hold-out	yes	accuracy
Baxter <i>et al.</i> (2020) [64]	none	CNN	LOOCV	yes	MAE
Shah <i>et al.</i> (2020) [65]	FS	DT	k-CV	yes	accuracy, sensitivity, specificity
Shang <i>et al.</i>	FB, FS	GBRT	LOOCV	yes	MAE, MSE, r

(2020) [66]					
Golshan <i>et al.</i> (2020) [67]	none	CNN	hold-out	no/not specified	accuracy, sensitivity, specificity
Khosravi <i>et al.</i> (2020) [68]	none	ANN	hold-out	no/not specified	accuracy
Lu <i>et al.</i> (2020) [69]	FS	SVM	hold-out	yes	accuracy, AUC-ROC, sensitivity, specificity, PPV, NPV
Peralta <i>et al.</i> (2020) [70]	none	CNN	k-CV	yes	BACC, sensitivity, specificity, F1
Yao <i>et al.</i> (2020) [71]	FB, FS	GBDT	hold-out	no/not specified	sensitivity, specificity, F1
Karthick <i>et al.</i> (2020) [72]	FB	RF	LOOCV	yes	accuracy, sensitivity, specificity, precision, F1
Park <i>et al.</i> (2021) [73]	none	CNN	hold-out	yes	accuracy, sensitivity, specificity

Peralta <i>et al.</i> (2021) [74]	FB	SVM, ANN	k-CV	yes	R
Boutet <i>et al.</i> (2021) [75]	FB	LDA	k-CV	yes	accuracy
Hosny <i>et al.</i> (2021) [76]	none	CNN	LOOCV	yes	accuracy, sensitivity, specificity, precision, F1
Martin <i>et al.</i> (2021) [77]	none	CNN, GMM	k-CV	yes	accuracy, sensitivity, specificity, BACC
Martin <i>et al.</i> (2021) [78]	none	CNN, GMM	k-CV	yes	sensitivity, specificity, FPR, FNR, BTQ
Baxter <i>et al.</i> (2021) [79]	none	CNN	LOOCV	yes	dice
Geraedts <i>et al.</i> (2021) [80]	FB, FS	RF	k-CV	yes	accuracy, sensitivity, specificity
Liebrand <i>et al.</i> (2021) [81]	FB	SVM	k-CV	yes	accuracy, AUC-ROC,



					MAE, MSE, sensitivity, specificity, r, $R^2$
Solomon <i>et al.</i> (2021) [82]	none	CNN	hold-out	yes	MAE, dice, MSD

**Table 2**

Data obtained from each of the 55 papers in the corpus for the ‘method’ and ‘validation’ classes.

‘FB’ stands for ‘feature-based’. ‘FS’ stands for ‘feature selection’.

(a) Input data modality type.

(b) Cumulative plot showing the number of patients in each paper’s cohort. The blue dashed line shows the median value (17 patients).

**Figure 4:** Charts presenting the results for the ‘data’ class.

## Data

Figure 4a presents the distribution of the input modality type used by the model. Site-specific electrophysiological signals represent nearly half of the modality used (22 papers used MER recorded by micro-electrodes and 17 works used Local Field Potential (LFP)). Information arising from imaging data were used 20 times. Third, external sensors were used nine times (eight times in the form of wearable sensors such as smartwatches and once with a force platform). Fourth, clinical data, such as demographics or patients’ questionnaires and clinical testing were used seven times. Fifth, stimulation information were only used in four papers, as well as global electrophysiological signals (two works used electroencephalography (EEG) and two

works used electrocorticography). Finally, Transcranial Magnetic Stimulation (TMS) was used once.

Figure 4b shows the distribution of cohort sizes. The average cohort size was 43 patients, with a median at 17. We counted 10 papers with cohorts larger than 100 patients, with a maximum at 501 and six papers that used data from a single patient.

Most cohorts (56 occurrences) are solely composed of PD patients. Cohorts of ET patients come second, with a total of five occurrences. Seven papers used a more heterogeneous cohort by mixing patients suffering from different pathologies, by mixing PD and ET patients, and/or by also studying patients suffering from dystonia or Tourette's. One paper used a cohort of patients suffering from dystonia, and one from OCD. Finally, three papers did not explicitly communicate the patients' condition.

## **Application**

Figure 5a presents the distribution of the clinical phase studied in the corpus. We can observe that the post-operative phase was the most extensively studied with 31 occurrences. The surgery phase and the planning phase came next with 21 and 13 occurrences respectively. Finally, the screening phase was the less studied with only nine occurrences.

Figure 5b presents the distribution of the tasks addressed. Classification is, by far, the most represented with 59 papers. Regression, segmentation, and clustering are far less represented with eight, five, and two papers respectively.

(a) Clinical phase corresponding to the problem.

(b) Task performed by the method.

**Figure 5:** Chart presenting the results for the ‘application’ class.

(a) Strategy to handle input data.

(b) Main ML model used, function of the input data modality type.

**Figure 6:** Charts presenting the results for the ‘method’ class.

## Method

Figure 6a shows the methods used to handle the input data. If several methods were employed, the best performing one was reported. The two main categories are feature-based methods and non-feature-based methods. The first category represents the contributions that require some initial domain knowledge and non-insignificant effort to transform (and/or filter) the initial input space into a more readily usable feature space. Additionally, we noted if compression methods and/or automatic feature selection method were used. Papers relying on feature-based methods were used more than 60% of the time, with 44 occurrences. In order to reduce the dimensionality of the input space, on average, feature selection or data compression techniques were used nearly 33% of the time. The raw input space was used in 20 papers, representing 27% of the occurrences.

Figure 6b shows the model used, or the best performing model(s) if several were benchmarked. The two most commonly used models were Support Vector Machines (SVM) and Artificial Neural Networks (ANN) (mostly through the form of shallow feed-forward neural networks such as a Multi-Layer Perceptron (MLP)), with 18 and 13 occurrences respectively. More advanced and/or specialized neural networks were also used such as Convolutional Neural Networks (CNN) (13 occurrences) and Recurrent Neural Networks (RNN), specifically with Long Short-Term Memory (LSTM) (one occurrence). In order to study temporal sequences, Hidden

Markov Models (HMM) were more regularly used than RNN with five occurrences. Random Forests (RF) are the fourth most commonly used models with nine occurrences, closely followed by logistic regression and k-Nearest Neighbors (kNN), with respectively six and five occurrences. Three papers used Ensemble Learning (EL) (excluding RF and gradient boosting methods) in order to combine the predictions of several base models. Gaussian Mixture Modeling (GMM) were also used three times, including two times on top of a CNN. Linear Discriminant Analysis (LDA) and Extreme Gradient Boosting (XGB) were used twice, and Naive Bayes (NB), Decision Trees (DT), Gaussian Process Regression (GPR), Gradient Boosting Decision Trees (GBDT) and Gradient Boosting Regression Trees (GBRT) and Gradient Boosting Regression Trees (GBRT) were used once. Finally, supervised learning represents the large majority of the papers, unsupervised models having been used only twice.

(a) Validation strategy regarding patient splitting strategy used for validation, both for patient-wise and non-patient wise validation methods.

(b) Metrics used to report the performance. Corresponding task is also indicated.

**Figure 7:** Charts presenting the results for the ‘validation’ class.

## Validation

Figure 7a presents the validation strategy with respect to how data was split between the training set and the testing set. A patient-wise validation strategy, employed by 36 papers, implies that data collected from a single patient is not simultaneously included in the training set and the testing set, reducing bias. 37 papers (ie. more than half) did not employ such a strategy, possibly containing information from a single patient in both the training and testing sets, or did not specify

how such splitting was achieved which leaves them open to potential data leakage.

Three different strategies were mainly used to split datapoints between training and testing sets. *Hold-out*, arguably the most common technique in the broader machine learning community, corresponds to separating a portion of the datapoints in a testing set at the beginning of the experiment in order to isolate it entirely from the training process, and was used 30 times in total, which makes it the most regularly used strategy. *k*-fold Cross-Validation (CV) (*k*-CV) comes second, with 26 occurrences in total. This strategy consists in partitioning the datapoints into *k* different disjoint sets. An independent training procedure is then performed *k* times, each using a different set as testing data and the remaining *k* - 1 as training data. When performed correctly (i.e. ensuring that the training procedures are completely independent and isolating from each other) this has the benefit of evaluating the algorithm on all the data possible although at the expense of re-training the algorithm several times. Leave-one-out CV (LOOCV) is a specific case *k*-CV for very small datasets, in which *k* is equal to the number of datapoints (or the number of patients, if the validation was done patient-wise) in the database. LOOCV is beneficial in that it also ensures that a maximum amount of data is present in the training set at any given time. LOOCV has been used 13 times. Two papers do not validate their method on a separated testing set, and one did not specify the validation method employed. Finally, *bootstrapping* refers to evaluation procedures in which the algorithm is trained once, but the effect of any individual datapoint can be isolated and removed from the model, thus allowing for said datapoint to be used in evaluating the reduced model. For methods that are inherently based on averaging an ensemble of simpler models (such as Random Forests), bootstrapping can be easily performed by removing the simpler models that had access to a particular datapoint in training time, and then evaluating the performance of the remaining simple models on said datapoint. This process is repeat and

averaged over all datapoints to estimate the perform of the model as a whole. Due to being highly model-specific, bootstrapping was only used once.

Figure 7b presents the evaluation metrics, as well as the corresponding task, for the proposed methods. For classification tasks, accuracy is the most consistently used metric (42 occurrences), followed by sensitivity and specificity (28 and 19 occurrences, respectively) and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) (14 occurrences). Several other metrics are less frequent including: precision and  $F1$  (six occurrences each), Cohen's Kappa (four occurrences), Matthews Correlation Coefficient (MCC) (three occurrences) , False Positive Rate (FPR) and balanced accuracy (BACC) (two occurrences each). Finally, number of classification errors, Weighted Classification Error (WCE), Negative Predictive Value (NPV), False Negative Rate (FNR), Positive Predictive Value (PPV), Balanced Ternary Quality (BTQ),  $R$ -ratio and  $\beta$ -ratio were used only once each. The metrics mostly used for regression were Mean Absolute Error (MAE) and Mean Squared Error (MSE), with six occurrences and three occurrence respectively. Pearson's  $r$  was used twice, and correlation coefficient ( $R$ ) and coefficient of determination ( $R^2$ ) were used once each. For segmentation, the Dice coefficient was the most consistently used metric with five occurrences, followed by Mean Squared Distance (MSD) and MAE (two occurrences each). Finally, accuracy, Intersection-over-Union (IoU), failure rate and Contour Mean Distance (CMD) were used once each. For clustering, sensitivity, specificity and MAE were each used once.

## **4. Discussion**

### **4.1. Clinical problems**

## Post-operative problems

The most consistent post-operative problem addressed with ML methods is the real-time analysis of LFP signals recorded by macro-electrode, in the perspective of delivering adaptive (rather than continuous) stimulation [13, 17, 18, 34, 35, 42, 43, 44, 45, 49, 50, 56, 60, 67, 71]. Indeed, the ability to process and analyze LFP signals in real time could allow for the design of closed-loop stimulation systems that deliver the therapy only when needed, thus limiting undesirable side effects and extending battery life. A complementary approach was investigated by Loukas *et al.* [16] who proposed a complete system to record, process, and display LFP signals. Houston *et al.* [31, 46] designed a close-loop system using cortical activity analysis, and Shukla *et al.* [15] and Khobragade *et al.* [23, 40] using surface electromyography (sEMG) and accelerometer signals. These works towards designing closed-loop stimulation systems represent 19 of the 28 papers focused on the post-operative phase of DBS interventions.

The second most common post-operative clinical problem is the analysis and quantification of the motor symptomatology of patients with external sensors (wearable sensors [24, 30, 38, 41, 52] or a force platform [11]). Such automatic systems can be clinically valuable by providing objective, automatic and quick feedback for therapies in order to, for example, compare several therapy parameters and combinations. Indeed, the degree of treatment parameter tuning, including stimulation parameters and drug dosage, is large and can't be assessed exhaustively, which may result in sub-optimal configurations. Other works have been done toward automatizing this post-operative phase: Connolly *et al.* [19] proposed a system predicting which contact is optimal from LFP recordings. Boutet *et al.* [75] proposed a method to predict whether the contact and stimulation voltage are optimal or not by analysing post-operative functional Magnetic Resonance Imaging (MRI). Shamir *et al.* [20] proposed a predictive system using inputs from

several modalities (clinical, therapy (medication and stimulation) and demographic data) in order to narrow the research space both for medication dosage and stimulation parameters. Lastly, Stuart *et al.* [54] used EEG to predict effective stimulation in real time.

### **Surgery problems**

The second most commonly investigated phase is the surgery itself, with the clinical problem frequently being the inter-operative identification of the DBS target. Every paper addressing this problem used MER analysis [10, 14, 21, 28, 33, 37, 39, 59, 63, 68, 70, 72, 76, 77, 78]. Instead of helping clinicians to aim for an anatomical structure, Lu *et al.* [69] proposed a method to predict, by analyzing MER, whether or not the electrode lead is inside a clinically predefined therapeutic site of activation. Complementarily, Wong *et al.* [12] proposed a method to project MER in a 2D plan in order for clinicians an alternative way to visualize and interpret it. Park *et al.* [73] proposed a method to predict motor outcomes six months after surgery by analysing MERs, which was used to find the optimal lead location during surgery. From the perspective of curating these signals for downstream analysis, Klempivir *et al.* [53] and Hosny *et al.* [61] proposed systems for artifact detection and correction. Kostoglou *et al.* [27] used MER features, a few clinical scores, demographics and contact location to predict the clinical improvement that could result from the stimulation of various locations, in a perspective of placing the electrode based on functional criteria rather than anatomical ones.

As MERs are electrophysiological signals, these papers are similar in methodology to those using LFPs.

### **Planning problems**



The planning phase comes third, the clinical problem being how to select the stimulation targets and determine the electrode trajectories prior to the operation. A first strategy is to assist the surgeon by automatically segmenting, or localising the subcortical structures of interest from pre-operative images ([22, 26, 32, 48, 51, 64, 79, 82]). Such works are important as the surgical targets are often small with low contrast in clinical images. An alternative strategy is to propose to the surgeon functional criteria instead of anatomical ones for the choice of a stimulation site. Baumgarten *et al.* [25, 29, 36] and Bermudez *et al.* [58] proposed clinical efficacy probability maps to visualize the expected clinical effects of stimulation of several locations around the structure of interest. Singer *et al.* [57] went even further in this idea by directly predicting the optimal electrode location without the use of intermediate representations such as anatomical segmentations or clinical-effect probability maps.

### **Screening problems**

Finally, the least commonly addressed phase using ML is the screening phase. Oliveira *et al.* [41] proposed a method to visualize on a two-dimension space the motor symptomatology of the patient from electromyography sensors in order to facilitate the clinical interpretation of patient motor scores. Habets *et al.* [55] proposed a predictive system to identify weak motor responders from pre-operative clinical data and demographics for patient selection purposes. In the same screening assisting tool objectives, Koch *et al.* [47] proposed a system to classify patients regarding their cognition from EEG. Using pre-operative EEG too, Geraedts *et al.* [80] proposed a method to predict post-operative cognitive functions. Farrokhi *et al.* [62] attempted to find factors of surgery adverse outcomes, such as infections or hemorrhages. Shah *et al.* [65] proposed a method to predict DBS outcomes from pre-operative demographics, clinical tests, expert

interpretation of anatomical abnormalities using MRI, and TMS. Shang *et al.* [66] proposed a method to predict post-operative motor outcomes with functional connectivity. Peralta *et al.* [74] proposed a method to predict 21 different clinical scores (including motor, cognitive, and quality-of-life scores) three months, six months, one year, and three years after surgery by using pre-operative clinical tests, demographic information, and the shape of particular anatomical structures using T1-MRI. Finally, Liebrand *et al.* [81] attempted to predict DBS good responders from pre-operative MRI but did not obtain satisfactory results.

## 4.2. Wide variety of models

When it comes to the choice of ML model, the most widely used remains the SVM which is not surprising as SVMs are known to perform well and are simple to train for both classification and regression problems, even for small databases. Among the papers comparing several models, SVMs were amongst the top-performing [52, 54]. Shallow feed-forward ANNs come second, likely due to the recent advances in deep learning and its growing popularity in the research community.

More specialized ANN structures, such as CNNs and RNNs were also used. CNNs were used 13 times in total, notably six times for image analysis: three times for subcortical structures segmentation or localisation with the VGG model [51], a modified ResNet structure [58], a custom structure called Hough-CNN [32] based on Hough voting, and U-net-based structures [64, 79, 82]. CNNs are also extensively used for MER spectrogram analysis: three times with a structure based on 1D separable convolutions [70, 77, 78], once with the AlexNet model [53], once with a CNN based on VGG16 and trained with multi-task learning [73], and once with a custom structure based on 1D-convolution [76]. RNNs were used once with LSTM for MER artifact detection [61].

We can also mention the usage of an interesting technique called EL, which consists in using several models in combination to make a more accurate prediction. Bagging was used eight times with RFs, and once by Kim *et al.* [22]. Gradient boosting was used four times, twice with the XGB model [43, 62] and twice with GBRT [66] or GBDT [71], and stacking was used by Golshan *et al.* [35] and Shamir *et al.* [20].

The methods used in the literature seem relatively independent of the input data used as shown in Figure 6b, which outlines a great heterogeneity of model used for each input data type, with the notable exception of CNNs, which were used 12 out of 13 times to analyze MER spectrograms and imaging data, and once to analyze LFPs. This is not a surprising result as CNNs are tailored to find spatial patterns automatically in high-dimensional data. Nevertheless, even though the use of Deep Learning (DL) models is becoming more and more common, other ML models (along with handcrafted features) still represent the majority of the methods employed. This is outlined by the analysis of site-specific electrophysiological signals such as MER and LFP where both options are possible, but where non-DL based methods are still used more than 85% of the time.

One difficulty in model training and optimization that particularly affects DL-based methods is the extensive amount of computational time and resources required. Indeed, on top of being usually heavyweight, CNNs require the tuning of several architectural and training hyper-parameters, necessitating the training of possibly hundreds of neural networks for a typical research paper. This is exacerbated when k-fold CV is employed, as  $k$  training iterations are performed to validate the model's performance. To this extent, Graphics Processing Units (GPU), computational clusters, or external computational clouds can be necessary to train some methods.

### **4.3. The prominent role of pre-processing and feature engineering for handling high-dimension input data**

An important consideration for ML studies is the curse of dimensionality: the greater the input dimensionality is, the exponentially greater the number of training samples are required to guarantee that data points are not too sparse in the input space. In DBS, the number of training samples are usually limited. Therefore, limiting the dimensionality of the input space to ease the training process can be an interesting strategy, even if it comes at the cost of reducing the amount of information available to the ML model.

Two common strategies can be to unsupervisedly compress the data and/or to automatically select the features, but such approaches are not in the majority. The most common strategy is to transform the original, raw input space into a set of features thanks to expert knowledge of the domain. For example, Kostoglou *et al.* [27] synthesized the MER signals by computing domain-specific features such as the mean inter-spike interval, or the power band ratio of the signal in different pre-determined frequency bands.

A minority of papers chose a fully data-driven method by straightforwardly feeding the model with raw data, without reducing it in the form of features, compressing it or automatically selecting some of them. Naturally, the 13 papers using CNNs fall under this umbrella. Indeed, the purpose of a CNN is to take advantage of the spatial inter-correlations in the input images by applying a series of learnable convolution kernels to it. Therefore, CNNs learn an optimal dimensionality reduction strategy in a supervised manner. Among the other occurrences are the three papers from Baumgarten *et al.* [25, 29, 36] and the paper from Habets *et al.* [55], because the number of inputs is low (respectively four stimulation parameters and 15 clinical scores and demographics).

We can interestingly mention five papers which compared a feature-based method with a fully data-driven method. First, for STN localisation using MER's, Khosravi *et al.* [37, 68] compared the utilization of state-of-the-art features versus the raw signal's Fourier coefficients, and Hosny *et al.* [76] compared the utilization of state-of-the-art features versus the raw signal directly. Second, Baumgarten *et al.* [29] compared two methods for predicting the occurrence of Pyramidal Tract Side Effect (PTSE) during the stimulation: a Volume of Tissue Activated (VTA)-based method versus a method using the raw information straightforwardly, which is the three-dimensional location of the contact and the stimulation voltage. All of the four papers showed a superiority of the fully data-driven paradigm which outperformed the feature-engineering approaches. Lastly, Yao *et al.* [71] compared the utilisation of handcrafted features versus a CNN for LFP signals classification, and got better results with handcrafted features, likely due to overfitting.

Feature engineering could be thought of as a particular, very involved type of pre-processing, that is, the modification of the input data before it is presented to the machine learning algorithm. From this perspective, pre-processing can be seen as a matter of degree. For example, for machine learning methods that process MERs as input, some use a spectral representation of the entire signal [37, 63], whereas others use a spectrogram [61] or Haar wavelets [10], i.e. a mixed temporal/frequential representation, and others process the temporal signal directly without explicitly representing the frequency components [14]. These can be interpreted as a gradation between methods that require a large amount of pre-processing in order to make the informative components more accessible to the machine learning algorithm and others that rely more heavily on the algorithm itself. That being said, some amount of pre-processing appears to always be necessary, even for complex machine learning approaches, to ensure that the input data

is meaningful to the particular algorithm or adheres to some pre-defined scope, however that scope has appeared to get broader as newer methods are developed, which is evidenced by half of the feature-selection-free methods being published in the last two years.

#### **4.4. Inter-patient variability: the elephant in the room**

Small cohorts are problematic in ML because they limit the performance of predictive systems (several papers [35, 49, 54, 60, 62] stated lack of data as a limitation). Furthermore, the pathologies treated by DBS are heterogeneous causing a high inter-patient variability, on top of intra-patient variability (as the clinical state of the patients can fluctuate or because the recording conditions may vary). Therefore, a system trained on one patient or on one recording configuration is not likely to have good performance on another one, or later in time, which limits its prospective usability.

The contribution of Khobragade *et al.* [40] illustrates this phenomenon well. They gathered surface electromyography and accelerometer data from two patients through several trials spread on different sessions, with at least a week between consecutive sessions. They did two experiments: the first one by training one model per patient and per session, therefore testing the model on trials of the same sessions. For the second one, they trained one model per patient, but trained the model on a set of sessions and tested it on other sessions. They obtained a perfect accuracy for the first experiment, but the median performance dropped to 46.15% for the second one. On the same extent, Rajpurohit *et al.* [21] reported results with a patient-specific feature normalization scheme (therefore not applicable prospectively), and with a patient-independent normalization scheme. Not surprisingly, the classification error rates of the patient-specific scheme are much lower (in the range of 0.0711 to 0.1353) than the patient-independent scheme

error rates (in the range of 0.102 to 0.1979).

The majority of the papers did not employ a patient-wise validation method [10, 13, 14, 15, 16, 17, 18, 19, 20, 23, 24, 25, 30, 31, 28, 34, 35, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 49, 50, 52, 53, 56, 59, 60, 67, 68, 71]. A large number of those papers reported performance on a single patient, training a different model for each patient and validating it on that same patient. This is often because a singular patient has multiple signals and the model uses some for training and others for evaluation, treating the learning of a model as a form of calibration. However, there were some papers that simply did not separate training and testing patients, could lead to biased measurements of their performance. The results reported by these contributions are interesting as preliminary work, but cannot safely be considered as representative of a real, prospective usage.

Others measured their performance by employing a patient-wise validation method [11, 12, 21, 22, 26, 27, 32, 29, 33, 36, 47, 48, 51, 54, 55, 57, 58, 61, 62, 63, 64, 65, 66, 69, 70, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82]. While it does not ensure the complete absence of data leakage (that could occur, for example, by mixing the validation and testing sets, or by selecting or normalize features with the whole database), these results can be considered as more reliable, and more representative of prospective-usage performance. Inter-patient variability remains an open problem for most of the contributions and is likely only solvable thanks to extensive data collection. Several papers stated that the lack of variability in the cohort limits the generalizability of the results [20, 47], that further validation has to be done on other recording conditions or other centers [33, 65, 70, 77, 82] or that inter-patient variability was a limiting problem [46, 68].

#### **4.5. A reproducibility and comparison problem**

Another recurring problem is the disparity in the validation methods which keeps from

comparing the results of different contributions together. As an example, all of the following contributions report the results of the location of the STN with MER but differ in the validation process:

- Guillén-Rondon *et al.* [14, 28] mixed patients and different portions of the same signals in training and test sets, which is an identified source of data leakage, leading to likely overestimated performance.
- Khosravi *et al.* [37] didn't specify if the validation was done on a separated set of patients.
- Rajpurohit *et al.* [21] used a leave-one-patient-one validation strategy.

While the contributions on the same topic are proliferating, it is often difficult or impossible to directly compare the quantitative results presented as the methods are validated in such heterogeneous ways with varying levels of bias and data leakage.

On the same extent, methods are not easily reproducible for benchmarking purposes, only three [41, 66, 75] mentioned having publicly shared, or being willing to share on request their code, their data or the features they computed on their data.

One factor that exacerbates this issue is a lack of openly available annotated datasets that multiple groups could use to ensure that a higher proportion of the patient variability is captured in training. To the best of our knowledge, there are no openly accessible databases of MER or LFP signals for DBS. For images and clinical questionnaires, there is an open dataset specific to Parkinson's disease, the Parkinson's Progression Markers Initiative (PPMI) [83]. However, because it is not specific to DBS, it lacks annotations for DBS-specific problems, and many of the individuals contained in the dataset may not have received DBS at all. The creation of these open datasets would be an obvious area for the community to develop, especially given how much of the



literature involves the collection of their own datasets which, if combined, would represent a sizeable if heterogeneous database for most ML problems associated with DBS.

One potential impediment to such a dataset is the relatively closed nature of medical data in general. Centres may be wary of releasing denser types of patient data such as MR images (and even less dense data such as clinical questionnaires) due to concerns about the future potential for de-anonymisation which would be ethically problematic [84].

## 5. Conclusion

We conducted, to the best of our knowledge, the first systematic review on ML for DBS, and identified several common methodological threads and limitations from the analysis of a corpus of 73 papers.

First, we have seen that only a few studies concern the screening phase. To us, there is a real opportunity for ML here because, as there exist several clinical challenges and a lot of complex and high-dimensional data arising from several pre-operative modalities (such as clinical testing, patient questionnaires, imaging data or demographics). These factors make this phase challenging to address but also ripe for data-driven methods providing that adequate methods are employed and large enough databases become available.

Second, the majority of studies use simple models with either a few features as input or aggressive dimensionality reduction methods (with automatic feature selection or feature-engineering). While it ensures the input data is readily usable for simple models, it also limits drastically the information that can be leveraged by the model, and therefore limits the performance and the practicable complexity of the prediction task. We think more ambitious problems, or higher levels of performance could be achieved by employing more bottom-up,

data-driven paradigms. Four papers comparing a feature-oriented method to a data-driven one support this hypothesis, showing better results for the latter approach [29, 37, 68, 76], where only one reported the opposite [71].

Third, small cohorts are often used: a lot of studies collect data that is not usually collected in the clinical routine, thus limiting the number of patients. Small cohorts, first, are problematic because they limit the generalization performance of the model and impose limits regarding the dimensionality of the input, because of the curse of dimensionality. Some papers overcame this problem by employing data augmentation when possible (for example, by splitting a ten seconds LFP signal into several two seconds ones). Unfortunately, this strategy does not address the second problem caused by small cohorts, which is that they cannot cover the large heterogeneity of the studied populations. Consequently, the performances reported are likely not representative of a prospective use of the system. Additionally, a number of contributions validated their method in a non-patient-wise manner which could lead to potential data leakage and thus inflated accuracy measures. Although the degree of this bias is not fully known, one study has found it to be extremely highly significant.

Lastly, to us, more effort has to be dedicated to the validation method employed and, if possible, to share the code and the data to facilitate reproducibility. We found that numerous papers are sharing similar, if not identical objectives but have not found any that benchmark different approaches by comparing the performance of their method to those proposed in the literature. This may be a result of the aforementioned problem of reproducibility, as it would imply reproducing and evaluating them with the same database and validation method. We think this area of research would benefit from more uniform validation methods and explicit validation guidelines, and further work would be required in this direction.

Overall, our survey indicates that ML is growing in this field, and we expect many of these issues to be resolved in large part with this particular research field maturing.

## Acknowledgments

Maxime Peralta's PhD is funded by the Fondation pour la Recherche Médicale (FRM). John Baxter is supported by a Post-Doctoral Fellowship from the Natural Sciences and Research Council of Canada (NSERC) and by the Institut des Neurosciences Cliniques de Rennes (INCR). These institutions were not involved in protocol development or in the data analysis.

## Declarations

**Conflicts of interest:** The authors have no relevant financial or non-financial interests to disclose.

## References

- [1] Alim-Louis Benabid, Pierre Pollak, Alain Louveau, S Henry, and J De Rougemont. Combined (thalamotomy and stimulation) stereotactic surgery of the vim thalamic nucleus for bilateral parkinson disease. *Stereotactic and functional neurosurgery*, 50(1-6):344–346, 1987.
- [2] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digital Medicine*, 3(1):1–10, 2020.

- [3] Ida Sim, Paul Gorman, Robert A Greenes, R Brian Haynes, Bonnie Kaplan, Harold Lehmann, and Paul C Tang. Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, 8(6):527–534, 2001.
- [4] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.
- [5] Emrah Celtikci. A systematic review on machine learning in neurosurgery: the future of decision-making in patient care. *Turk Neurosurg*, 28(2):167–173, 2018.
- [6] Quinlan D Buchlak, Nazanin Esmaili, Jean-Christophe Leveque, Farrokh Farrokhi, Christine Bennett, Massimo Piccardi, and Rajiv K Sethi. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. *Neurosurgical review*, pages 1–19, 2019.
- [7] Joeky T Senders, Mark M Zaki, Aditya V Karhade, Bliss Chang, William B Gormley, Marike L Broekman, Timothy R Smith, and Omar Arnaout. An introduction and overview of machine learning in neurosurgical care. *Acta neurochirurgica*, 160(1):29–38, 2018.
- [8] Joeky T Senders, Omar Arnaout, Aditya V Karhade, Hormuzdiyar H Dasenbrock, William B Gormley, Marike L Broekman, and Timothy R Smith. Natural and artificial intelligence in neurosurgery: a systematic review. *Neurosurgery*, 83(2):181–192, 2018.
- [9] David Moher, Larissa Shamseer, Mike Clarke, Davina Ghera, Alessandro Liberati, Mark Petticrew, Paul Shekelle, Lesley A Stewart, et al. Preferred reporting items for systematic review and meta-analysis protocols (prisma-p) 2015 statement. *Systematic reviews*, 4(1):1, 2015.
- [10] Alvaro Orozco, Mauricio Alvarez, Enrique Guijarro, and German Castellanos.

- Identification of spike sources using proximity analysis through hidden markov models. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5555–5558. IEEE, 2006.
- [11] AMS Muniz, W Liu, H Liu, KE Lyons, R Pahwa, FF Nobre, and J Nadal. Assessment of the effects of subthalamic stimulation in parkinson disease patients by artificial neural network. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5673–5676. IEEE, 2009.
- [12] Stephen Wong, GH Baltuch, JL Jaggi, and SF Danish. Functional localization and visualization of the subthalamic nucleus from microelectrode recordings acquired during dbs surgery with unsupervised machine learning. *Journal of neural engineering*, 6(2):026006, 2009.
- [13] Defeng Wu, Kevin Warwick, Zi Ma, Mark N Gasson, Jonathan G Burgess, Song Pan, and Tipu Z Aziz. Prediction of parkinson’s disease tremor onset using a radial basis function neural network based on particle swarm optimization. *International journal of neural systems*, 20(02):109–116, 2010.
- [14] Pablo Guillén, F Martinez-de Pison, R Sanchez, Miguel Argáez, and Leticia Velázquez. Characterization of subcortical structures during deep brain stimulation utilizing support vector machines. In *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 7949–7952. IEEE, 2011.
- [15] Pitamber Shukla, Ishita Basu, Daniel Graupe, Daniela Tuninetti, and Konstantin V Slavin. A neural network-based design of an on-off adaptive control for deep brain stimulation in movement disorders. In *2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4140–4143. IEEE, 2012.

- [16] Constantinos Loukas and Peter Brown. A pc-based system for predicting movement from deep brain signals in parkinson's disease. *Computer methods and programs in biomedicine*, 107(1):36–44, 2012.
- [17] Huaiguang Jiang, Jun Jason Zhang, Adam Hebb, and Mohammad H Mahoor. Time-frequency analysis of brain electrical signals for behavior recognition in patients with parkinson's disease. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 1843–1847. IEEE, 2013.
- [18] Soroush Niketeghad, Adam O Hebb, Joshua Nedrud, Sara J Hanrahan, and Mohammad H Mahoor. Single trial behavioral task classification using subthalamic nucleus local field potential signals. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3793–3796. IEEE, 2014.
- [19] Allison T Connolly, William F Kaemmerer, Siddharth Dani, Scott R Stanslaski, Eric Panken, Matthew D Johnson, and Timothy Denison. Guiding deep brain stimulation contact selection using local field potentials sensed by a chronically implanted device in parkinson's disease patients. In *2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 840–843. IEEE, 2015.
- [20] Reuben R Shamir, Trygve Dolber, Angela M Noecker, Benjamin L Walter, and Cameron C McIntyre. Machine learning approach to optimizing combined stimulation and medication therapies for parkinson's disease. *Brain stimulation*, 8(6):1025–1032, 2015.
- [21] Vikram Rajpurohit, Shabbar F Danish, Eric L Hargreaves, and Stephen Wong. Optimizing computational feature sets for subthalamic nucleus localization in dbs surgery with feature selection. *Clinical Neurophysiology*, 126(5):975–982, 2015.
- [22] Jinyoung Kim, Yuval Duchin, Hyunsoo Kim, Jerrold Vitek, Noam Harel, and Guillermo

- Sapiro. Robust prediction of clinical deep brain stimulation target structures via the estimation of influential high-field mr atlases. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 587–594. Springer, 2015.
- [23] Nivedita Khobragade, Daniel Graupe, and Daniela Tuninetti. Towards fully automated closed-loop deep brain stimulation in parkinson’s disease patients: a lamstar-based tremor predictor. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2616–2619. IEEE, 2015.
- [24] Shivanthan AC Yohanandan, Mary Jones, Richard Peppard, Joy L Tan, Hugh J McDermott, and Thushara Perera. Evaluating machine learning algorithms estimating tremor severity ratings on the bain–findley scale. *Measurement Science and Technology*, 27(12):125702, 2016.
- [25] Clement Baumgarten, Yulong Zhao, Paul Sauleau, Cecile Malrain, Pierre Jannin, and Claire Haegelen. Image-guided preoperative prediction of pyramidal tract side effect in deep brain stimulation: proof of concept and application to the pyramidal tract side effect induced by pallidal stimulation. *Journal of Medical Imaging*, 3(2):025001, 2016.
- [26] Yuan Liu and Benoit M Dawant. Multi-modal learning-based pre-operative targeting in deep brain stimulation procedures. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, pages 17–20. IEEE, 2016.
- [27] Kyriaki Kostoglou, Konstantinos P Michmizos, Pantelis Stathis, Damianos Sakas, Konstantina S Nikita, and Georgios D Mitsis. Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings. *IEEE Transactions on Biomedical Engineering*, 64(5):1123–1130, 2016.
- [28] Pablo Guillén-Rondon and Melvin D Robinson. Deep brain stimulation signal

- classification using deep belief networks. In *2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 155–158. IEEE, 2016.
- [29] Clement Baumgarten, Yulong Zhao, Paul Sauleau, Cecile Malrain, Pierre Jannin, and Claire Haegelen. Improvement of pyramidal tract side effect prediction using a data-driven method in subthalamic stimulation. *IEEE Transactions on Biomedical Engineering*, 64(9):2134–2141, 2016.
- [30] Paolo Angeles, Yen Tai, Nicola Pavese, Samuel Wilson, and Ravi Vaidyanathan. Automated assessment of symptom severity changes during deep brain stimulation (dbs) therapy for parkinson’s disease. In *2017 International Conference on Rehabilitation Robotics (ICORR)*, pages 1512–1517. IEEE, 2017.
- [31] Brady C Houston, Margaret C Thompson, Jeffrey G Ojemann, Andrew L Ko, and Howard J Chizeck. Classifier-based closed-loop deep brain stimulation for essential tremor. In *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 316–320. IEEE, 2017.
- [32] Fausto Milletari, Seyed-Ahmad Ahmadi, Christine Kroll, Annika Plate, Verena Rozanski, Juliana Maiostre, Johannes Levin, Olaf Dietrich, Birgit Ertl-Wagner, Kai Bötzel, et al. Hough-cnn: deep learning for segmentation of deep brain regions in mri and ultrasound. *Computer Vision and Image Understanding*, 164:92–102, 2017.
- [33] Dan Valsky, Odeya Marmor-Levin, Marc Deffains, Renana Eitan, Kim T Blackwell, Hagai Bergman, and Zvi Israel. Stop! border ahead: A utomatic detection of subthalamic exit during deep brain stimulation surgery. *Movement Disorders*, 32(1):70–79, 2017.
- [34] Ameer Mohammed, Majid Zamani, Richard Bayford, and Andreas Demosthenous.



- Toward on-demand deep brain stimulation using online parkinson's disease prediction driven by dynamic detection. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12):2441–2452, 2017.
- [35] Hosein M Golshan, Adam O Hebb, Sara J Hanrahan, Joshua Nedrud, and Mohammad H Mahoor. A hierarchical structure for human behavior classification using stn local field potentials. *Journal of neuroscience methods*, 293:254–263, 2018.
- [36] Clément Baumgarten, Claire Haegelen, Yulong Zhao, Paul Sauleau, and Pierre Jannin. Data-driven prediction of the therapeutic window during subthalamic deep brain stimulation surgery. *Stereotactic and functional neurosurgery*, 96(3):142–150, 2018.
- [37] Mahsa Khosravi, Seyed Farokh Atashzar, Greydon Gilmore, Mandar S Jog, and Rajni V Patel. Electrophysiological signal processing for intraoperative localization of subthalamic nucleus during deep brain stimulation surgery. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pages 424–428. IEEE, 2018.
- [38] Robert LeMoyné, Timothy Mastroianni, Cyrus McCandless, Christopher Currivan, Donald Whiting, and Nestor Tomycz. Implementation of a smartphone as a wearable and wireless accelerometer and gyroscope platform for ascertaining deep brain stimulation treatment efficacy of parkinson's disease through machine learning classification. *Advances in Parkinson's Disease*, 7(2):19–30, 2018.
- [39] Hernán Darío Vargas Cardona, Mauricio A Álvarez, and Álvaro A Orozco. Multi-task learning for subthalamic nucleus identification in deep brain stimulation. *International Journal of Machine Learning and Cybernetics*, 9(7):1181–1192, 2018.
- [40] Nivedita Khobragade, Daniela Tuninetti, and Daniel Graupe. On the need for adaptive learning in on-demand deep brain stimulation for movement disorders. In *2018 40th*

- Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2190–2193. IEEE, 2018.
- [41] Fábio Henrique M Oliveira, Alessandro RP Machado, and Adriano O Andrade. On the use of t-distributed stochastic neighbor embedding for data visualization and classification of individuals with parkinson’s disease. *Computational and mathematical methods in medicine*, 2018, 2018.
- [42] Syed A Shah, Gerd Tinkhauser, Chiung Chu Chen, Simon Little, and Peter Brown. Parkinsonian tremor detection from subthalamic nucleus local field potentials for closed-loop deep brain stimulation. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 2320–2324. IEEE, 2018.
- [43] Lin Yao, Peter Brown, and Mahsa Shoaran. Resting tremor detection in parkinson’s disease with machine learning and kalman filtering. In *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2018.
- [44] Hosein M Golshan, Adam O Hebb, Joshua Nedrud, and Mohammad H Mahoor. Studying the effects of deep brain stimulation and medication on the dynamics of stn-lfp signals for human behavior analysis. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4720–4723. IEEE, 2018.
- [45] Taige Wang, Mahsa Shoaran, and Azita Emami. Towards adaptive deep brain stimulation in parkinson’s disease: Lfp-based feature analysis and classification. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2536–2540. IEEE, 2018.
- [46] Brady Houston, Margaret Thompson, Andrew Ko, and Howard Chizeck. A

- machine-learning approach to volitional control of a closed-loop deep brain stimulation system. *Journal of neural engineering*, 16(1):016004, 2018.
- [47] Milan Koch, Victor Geraedts, Hao Wang, Martijn Tannemaat, and Thomas Bäck. Automated machine learning for eeg-based classification of parkinson's disease patients. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 4845–4852. IEEE, 2019.
- [48] Jinyoung Kim, Yuval Duchin, Reuben R Shamir, Remi Patriat, Jerrold Vitek, Noam Harel, and Guillermo Sapiro. Automatic localization of the subthalamic nucleus on patient-specific clinical mri by incorporating 7 t mri and machine learning: Application in deep brain stimulation. *Human brain mapping*, 40(2):679–698, 2019.
- [49] Yue Chen, Chen Gong, Hongwei Hao, Yi Guo, Shujun Xu, Yuhuan Zhang, Guoping Yin, Xin Cao, Anchao Yang, Fangang Meng, et al. Automatic sleep stage classification based on subthalamic local field potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 27(2):118–128, 2019.
- [50] Huiling Tan, Jean Debarros, Shenghong He, Alek Pogosyan, Tipu Z Aziz, Yongzhi Huang, Shouyan Wang, Lars Timmermann, Veerle Visser-Vandewalle, David J Pedrosa, et al. Decoding voluntary movements and postural tremor based on thalamic lfps as a basis for closed-loop stimulation for essential tremor. *Brain stimulation*, 12(4):858–867, 2019.
- [51] Seong-Cheol Park, Joon Hyuk Cha, Seonhwa Lee, Wooyoung Jang, Chong Sik Lee, and Jung Kyo Lee. Deep learning-based deep brain stimulation targeting and clinical applications. *Frontiers in neuroscience*, 13:1128–1128, 2019.
- [52] Robert LeMoyne, Timothy Mastroianni, Cyrus McCandless, Donald Whiting, and Nestor Tomycz. Evaluation of machine learning algorithms for classifying deep brain stimulation

- respective of on and off status. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 483–488. IEEE, 2019.
- [53] Ondřej Klempíř, Radim Krupička, Eduard Bakštein, and Robert Jech. Identification of microrecording artifacts with wavelet analysis and convolutional neural network: An image recognition approach. *Measurement Science Review*, 19(5):222–231, 2019.
- [54] Morgan Stuart, Chathurika S Wickramasinghe, Daniel L Marino, Deepak Kumbhare, Kathryn Holloway, and Milos Manic. Machine learning for deep brain stimulation efficacy using dense array eeg. In *2019 12th International Conference on Human System Interaction (HSI)*, pages 143–150. IEEE, 2019.
- [55] Jeroen GV Habets, Annelien A Duits, Laura CJ Sijben, Bianca De Greef, Anne Mulders, Yasin Temel, Mark L Kuijf, Pieter L Kubben, Christian Herff, and Marcus LF Janssen. Machine learning prediction of motor response after deep brain stimulation in parkinson’s disease. *medRxiv*, page 19006841, 2019.
- [56] Carmen Camara, Narayan P Subramaniam, Kevin Warwick, Lauri Parkkonen, Tipu Aziz, and Ernesto Pereda. Non-linear dynamical analysis of resting tremor for demand-driven deep brain stimulation. *Sensors*, 19(11):2507, 2019.
- [57] Alexa Singer, Chencheng Zhang, Tao Wang, Suhao Qiu, Dianyou Li, Yiping Du, Zhi-Pei Liang, Pawel Herman, Bomin Sun, and Yuan Feng. Post-operative electrode placement prediction in deep brain stimulation using support vector regression. In *Proceedings of the Third International Symposium on Image Computing and Digital Medicine*, pages 202–207, 2019.
- [58] Camilo Bermudez, William Rodriguez, Yuankai Huo, Allison E Hainline, Rui Li, Robert Shults, Pierre D Dâ€™Haese, Peter E Konrad, Benoit M Dawant, and Bennett A Landman.


- Towards machine learning prediction of deep brain stimulation (dbs) intra-operative efficacy maps. In *Medical Imaging 2019: Image Processing*, volume 10949, page 1094922. International Society for Optics and Photonics, 2019.
- [59] Konrad A Ciecierski and Tomasz Mandat. Unsupervised machine learning in classification of neurobiological data. In *Intelligent Methods and Big Data in Industrial Applications*, pages 203–212. Springer, 2019.
- [60] Ameer Mohammed, Richard Bayford, and Andreas Demosthenous. A framework for adapting deep brain stimulation using parkinsonian state estimates. *Frontiers in Neuroscience*, 14:499, 2020.
- [61] Mohamed Hosny, Minwei Zhu, Wenpeng Gao, and Yili Fu. A novel deep lstm network for artifacts detection in microelectrode recordings. *Biocybernetics and Biomedical Engineering*, 2020.
- [62] Farrokh Farrokhi, Quinlan D Buchlak, Matt Sikora, Nazanin Esmaili, Maria Marsans, Pamela McLeod, Jamie Mark, Emily Cox, Christine Bennett, and Jonathan Carlson. Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms. *World Neurosurgery*, 134:e325–e338, 2020.
- [63] Dan Valsky, Kim T Blackwell, Idit Tamir, Renana Eitan, Hagai Bergman, and Zvi Israel. Real-time machine learning classification of pallidal borders during deep brain stimulation surgery. *Journal of Neural Engineering*, 17(1):016021, 2020.
- [64] John SH Baxter, Ehouarn Maguet, and Pierre Jannin. Localisation of the subthalamic nucleus in mri via convolutional neural networks for deep brain stimulation planning. In *Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11315, page 113150M. International Society for Optics and Photonics, 2020.

- [65] Syed Ahmar Shah, Peter Brown, Hortensia Gimeno, Jean-Pierre Lin, and Verity M McClelland. Application of machine learning using decision trees for prognosis of deep brain stimulation of globus pallidus internus for children with dystonia. *Frontiers in neurology*, 11:825, 2020.
- [66] Ruihong Shang, Le He, Xiaodong Ma, Yu Ma, and Xuesong Li. Connectome-Based Model Predicts Deep Brain Stimulation Outcome in Parkinson’s Disease. *Frontiers in computational neuroscience*, 14:98, 2020.
- [67] Hosein M Golshan, Adam O Hebb, and Mohammad H Mahoor. LFP-Net: A deep learning framework to recognize human behavioral activities using brain STN-LFP signals. *Journal of neuroscience methods*, 335:108621, 2020.
- [68] Mahsa Khosravi, S Farokh Atashzar, Greydon Gilmore, Mandar S Jog, and Rajni V Patel. Intraoperative localization of STN during DBS surgery using a data-driven model. *IEEE journal of translational engineering in health and medicine*, 8:1–9, 2020.
- [69] Charles W Lu, Karlo A Malaga, Kelvin L Chou, Cynthia A Chestek, and Parag G Patil. High density microelectrode recording predicts span of therapeutic tissue activation volumes in subthalamic deep brain stimulation for Parkinson disease. *Brain stimulation*, 13(2):412–419, 2020.
- [70] Maxime Peralta, Quoc Anh Bui, Antoine Ackaouy, Thibault Martin, Greydon Gilmore, Claire Haegelen, Paul Sauleau, John SH Baxter, and Pierre Jannin. SepaConvNet for Localizing the Subthalamic Nucleus using One Second Micro-Electrode Recordings. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 888–893. IEEE, 2020.
- [71] Lin Yao, Peter Brown, and Mahsa Shoaran. Improved detection of Parkinsonian resting

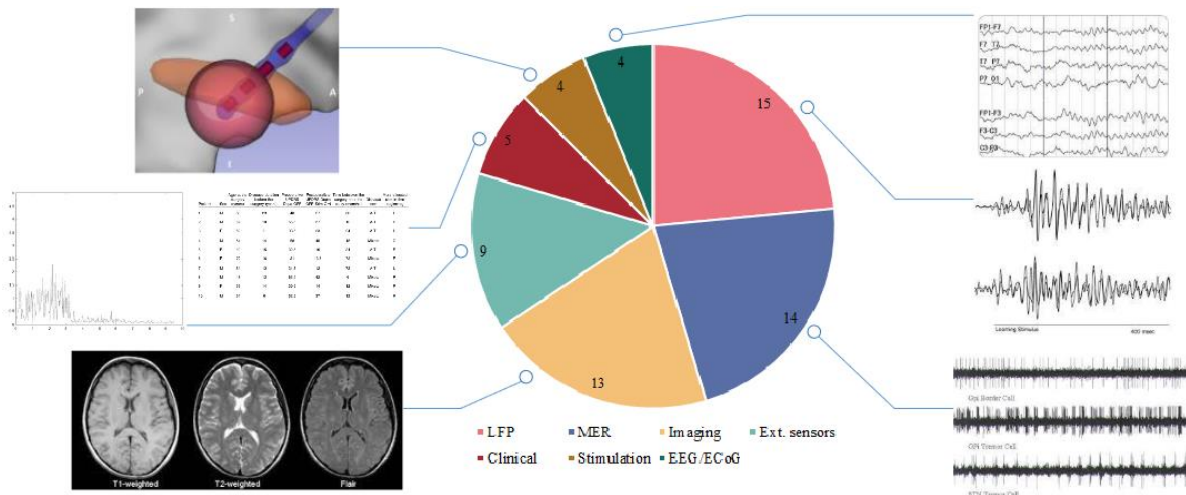
- tremor with feature engineering and Kalman filtering. *Clinical Neurophysiology*, 131(1):274–284, 2020.
- [72] PA Karthick, Kai Rui Wan, Angela See An Qi, Justin Dauwels, and Nicolas Kon Kam King. Automated detection of subthalamic nucleus in deep brain stimulation surgery for Parkinson’s disease using microelectrode recordings and wavelet packet features. *Journal of Neuroscience Methods*, 343:108826, 2020.
- [73] Kwang Hyon Park, Sukkyu Sun, Yong Hoon Lim, Hye Ran Park, Jae Meen Lee, Kawngwoo Park, Beomseok Jeon, Hee-Pyoung Park, Hee Chan Kim, and Sun Ha Paek. Clinical outcome prediction from analysis of microelectrode recordings using deep learning in subthalamic deep brain stimulation for Parkinsons disease. *PloS one*, 16(1):e0244133, 2021.
- [74] Maxime Peralta, Claire Haegelen, Pierre Jannin, and John Baxter. PassFlow: a multimodal workflow for predicting Deep Brain Stimulation Outcomes. In *CARS: Computer Assisted Radiology and Surgery 2021*, 2021.
- [75] Alexandre Boutet, Radhika Madhavan, Gavin JB Elias, Suresh E Joel, Robert Gramer, Manish Ranjan, Vijayashankar Paramanandam, David Xu, Jurgen Germann, Aaron Loh, et al. Predicting optimal deep brain stimulation parameters for Parkinson’s disease using functional MRI and machine learning. *Nature communications*, 12(1):1–13, 2021.
- [76] Mohamed Hosny, Minwei Zhu, Wenpeng Gao, and Yili Fu. Deep convolutional neural network for the automated detection of Subthalamic nucleus using MER signals. *Journal of Neuroscience Methods*, 356:109145, 2021.
- [77] Thibault Martin, Maxime Peralta, Greydon Gilmore, Paul Sauleau, Claire Haegelen, Pierre Jannin, and John SH Baxter. Extending convolutional neural networks for localizing the

- subthalamic nucleus from micro-electrode recordings in Parkinson's disease. *Biomedical Signal Processing and Control*, 67:102529, 2021.
- [78] Thibault Martin, Greydon Gilmore, Claire Haegelen, Pierre Jannin, and John SH Baxter. Adapting the listening time for micro-electrode recordings in deep brain stimulation interventions. *International Journal of Computer Assisted Radiology and Surgery*, pages 1–9, 2021.
- [79] John SH Baxter, Ehouarn Maguet, and Pierre Jannin. Segmentation of the subthalamic nucleus in MRI via Convolutional Neural Networks for deep brain stimulation planning. In *Medical Imaging 2021: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 11598, page 115981K. International Society for Optics and Photonics, 2021.
- [80] VJ Geraedts, M Koch, MF Contarino, HAM Middelkoop, H Wang, JJ van Hilten, THW Bäck, and MR Tannemaat. Machine learning for automated EEG-based biomarkers of cognitive impairment during Deep Brain Stimulation screening in patients with Parkinson's Disease. *Clinical Neurophysiology*, 132(5):1041–1048, 2021.
- [81] Luka C Liebrand, Paul Zhutovsky, Eva K Tolmeijer, Ilse Graat, Nienke Vulink, Pelle de Koning, Martijn Figee, P Richard Schuurman, Pepijn van den Munckhof, Matthan WA Caan, et al. Deep brain stimulation response in obsessive–compulsive disorder is associated with preoperative nucleus accumbens volume. *NeuroImage: Clinical*, 30:102640, 2021.
- [82] Oren Solomon, Tara Palnitkar, Re'mi Patriat, Henry Braun, Joshua Aman, Michael C Park, Jerrold Vitek, Guillermo Sapiro, and Noam Harel. Deep-learning based fully automatic segmentation of the globus pallidus interna and externa using ultra-high 7 Tesla MRI. *Human brain mapping*, 42(9):2862–2879, 2021.



- 
- [83] Kenneth Marek, Danna Jennings, Shirley Lasch, Andrew Siderowf, Caroline Tanner, Tanya Simuni, Chris Coffey, Karl Kieburtz, Emily Flagg, Sohini Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.
- [84] Khaled El Emam, Sam Rodgers, and Bradley Malin. Anonymising and sharing individual patient data. *bmj*, 350, 2015.

Graphical abstract



## Highlights

- We survey 73 recent papers on machine learning (ML) in deep brain stimulation (DBS)
- ML has been increasingly applied in DBS to process electrical signals and MR images
- ML in DBS is largely dominated by classification problems and traditional ML algorithms
- Validation is heterogeneous with numerous different metrics and techniques employed
- Many data processing problems in deep brain stimulation are still largely unanswered

# Conflicts of Interest Statement

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.