Generalisation capabilities of machine-learning algorithms for the detection of the subthalamic nucleus in micro-electrode recordings

Thibault Martin, Pierre Jannin, John S H Baxter

Generalisation Capabilities of Machine-Learning Algorithms for the Detection of the Subthalamic Nucleus in Micro-Electrode Recordings

Thibault Martin[1], Pierre Jannin[1], John S. H. Baxter[1*]

[1]Laboratoire Traitement du Signal et de l'Image (LTSI, INSERM UMR 1099), Université de Rennes, Rennes, France.

*Corresponding author(s). E-mail(s): john.baxter@univ-rennes.fr;
Contributing authors: thibault.martin@univ-rennes.fr;
pierre.jannin@univ-rennes.fr;

Abstract

Purpose: Micro-electrode recordings (MERs) are a key intra-operative modality used during deep brain stimulation (DBS) electrode implantation, which allow for a trained neurophysiologist to infer the anatomy in which the electrode is placed. As DBS targets are small, such inference is necessary to confirm that the electrode is correctly positioned. Recently, machine learning techniques have been used to augment the neurophysiologist's capability. The goal of this paper is to investigate the generalisability of these methods with respect to different clinical centres and training paradigms.

Methods: Five deep learning algorithms for binary classification of MER signals have been implemented. Three databases from two different clinical centres have also been collected with differing size, acquisition hardware, and annotation protocol. Each algorithm has initially been trained on the largest database, then either directly tested or fine-tuned on the smaller databases in order to estimate their generalisability. As a reference, they have also been trained from scratch on the smaller databases as well in order to estimate the effect of the differing database sizes and annotation systems.

Results: Each network shows significantly reduced performance (on the order of a 6.5% to 16.0% reduction in balanced accuracy) when applied out-of-distribution. This reduction can be ameliorated through fine-tuning the network on the new database through transfer learning, although even for these small databases, it appears that retraining from scratch may still offer equivalent performance as fine-tuning with transfer learning. However, this is at the expense of significantly longer training times.

Conclusion: Generalisability is an important criterion for the success of machine learning algorithms in clinic. We have demonstrated that a variety of recent machine learning algorithms for MER classification are negatively affected by domain shift, but that this can be quickly ameliorated through simple transfer learning procedures that can be readily performed for new centres.

Keywords: Micro-electrode recordings, machine learning, generalisability

1 Introduction

Deep brain stimulation (DBS) is an interventional treatment used to control the symptoms of several neurological disorders. The most common of these is undoubtedly Parkinson's disease. Although the small subcortical structures targeted in DBS interventions, such as the subthalamic nucleus (STN), are visible on a pre-operative MRI, sources of error arising from brain shift and small uncontrollable deviations from the pre-planned electrode trajectory necessitate an intra-operative data modality to inform the interventionalist as to if the electrode is correctly positioned [1].

One of these potential modalities is Micro-Electrode Recording (MER) in which a listening electrode is first implanted, allowing for the clinical team to hear the neural behaviour happening at a particular depth along the trajectory. From this information, a trained neurophysiologist could then determine what anatomy lies underneath the electrode's current position and infer whether or not said depth is appropriate for final electrode positioning. Furthermore, recent studies have show than using MER as an intra-operative data modality produces similarly patient outcomes as the use of intra-operative MRI, despite much lower cost and wider accessibility [1].

However, the use of MER also implies a longer intervention duration [2]. This is because the interpretation of an MER signal is difficult and time-consuming as well as necessitates a large amount of experience and judgement on the part of the human expert. Recently, several machine learning algorithms [3–13] have been proposed as decision-making support tools for this task in order to better control for this subjectivity and to make the overall DBS electrode implantation intervention more efficient. These algorithms have relied on a variety of different machine learning paradigms. Most recent methods use traditional approaches which classify signals based on specific features that are known to be of interest to the neuroscientific community, such as the power in particular frequency bands [5, 6, 8, 11] or various spike-dependant or spike-independent features [3, 4] as well as more technical features such as wavelet decompositions [7]. Others [9, 10, 12, 13] have taken a more modern, deep learning approach in which the entire signal is provided to the algorithm rather than reducing it to a smaller vector of features.

One of the key limitations to all of these approaches does not lie specifically in the paradigm or machine learning architecture chosen, but in methodological design. Each of these studies [8–13] was performed with databases arising from a uniform protocol in a singular clinical centre with the annotations provided by a singular expert neurophysiologist. In addition, these studies (as well as many others in the

field of machine learning in DBS research) use a variety of different procedures for quantitatively validating their algorithms, leading to a large degree of uncertainty in comparing their numerical results [14]. Given the technical and relatively early stage of this research, such study designs are par for the course but do lead to questions regarding their generalisability as well as comparative performance [14].

As deep learning algorithms are known to be sensitive to subtle or imperceptible differences in dataset distributions, generalisability and comparative performance studies are crucial for determining the performance of algorithms outside of their training conditions in those that are more similar to diverse clinical contexts [15]. Such an investigation of simpler binary problems, such as MER signal classification, is important in order to ensure that more complex and ambiguous tasks, such as optimal placement planning [16], can also be constructed in a more robust and generalisable manner.

1.1 Contributions

This article examines a series of recently published deep learning algorithms [8–10, 12, 13] that are all designed to classify whether or not a given MER arises from inside the STN or from outside of it. In order to do so, we have re-implemented each of these algorithms in order to control what data they access during training, what data they are evaluated against, as well as the training paradigm used to potentially correct these domain shift errors.

2 Methods

2.1 Patient Data

The primary database used for initially training the STN classification algorithms consists of 57 Parkinsonian patients undergoing either single or bilateral DBS electrode implantation. The MER signals arising through five channels (anterior, posterior, medial, lateral, and central) were recorded using the *Leadpoint 5* (Medtronic) station. The MER signals were captured from 10.0mm prior to the pre-operatively estimated STN boundary to 4.0/5.0mm after said boundary, leading to the acquisition of 11,162 signals each 9s (or 192000 samples) long. Each signal was recorded at 24kHz and then bandpass filtered (500-5000Hz, notch: 60Hz). In order to mitigate for artefacts, amplitude clipping was performed and the signal intensity rescaled to a -249 to 250 range. The collection and use of this data was approved by the Research Ethics Board at Western University, Canada (REB # 109045). From now on, this dataset will be referred to as the *London database*.

The first of the secondary *Rennes databases* used for evaluating the generalisability of the algorithms was acquired at the Rennes University Hospital Centre between 2015 and 2022 from 63 patients undergoing DBS electrode implantation. A Ben-Gun configuration with three channels (anterior, central, and posterior) was used and the signals were recorded using the Dantex Keypoint G station at 24kHz. 20-40 MER signals were collected per channel per patient leading to a total of 8,630 signals, each 3 seconds in length. The data was then digitally bandpass filtered (500-5000Hz, notch:

50Hz). Similar to the London database, amplitude clipping was applied with the same parameters. This data collection was approved by the Rennes University Hospital Centre ethics committee (Ethical authorisation declaration n2205295).

The second, smaller secondary database was also collected at the Rennes University Hospital Centre between 2005 and 2015 from 50 patients. This configuration included 5 channels (anterior, posterior, medial, lateral, and central) recorded using the Dantex Keypoint G station at 24kHz and then bandpass filtered (500-5000Hz, notch: 50Hz). This database is smaller, consisting of 1880 signals, each 10 seconds long. This data collection was also approved by the Rennes University Hospital Centre ethics committee (Ethical authorisation declaration n2205295).

Each of the databases were provided with manual annotations regarding whether or not the signal arose from the STN. For the London database, this annotation was performed by an expert neurosurgeon at University Hospital (Western University, London, Canada) and for the two Rennes databases, it was performed by an expert neurophysiologist at the Rennes University Hospital Centre (Rennes, France).

These databases differ in terms of their size, length, acquisition system, bandpass filter parameters, number of channels, and class balance. This allows for different aspects of generalisability and retrainability to be evaluated.

2.2 Algorithms

Five recent machine learning algorithms were selected from the literature. The primary requirement of these algorithms is that they be based on artificial neural networks of some variety with sufficient explanation of their methods to replicate them or code provided either open source or via communication with the paper authors. These algorithms reflect the many conceptual evolutions in deep learning, starting with multi-layer perceptrons applied to a vector of engineered features [8] before expanding to convolutional networks [9, 12], recurrent networks [10], and finally transformers [13]. Each of these networks were given the same data as input, considering their pre-processing steps (e.g. feature extraction, short-time Fourier transforms, etc...) as fundamentally part of the individual method given how much these steps vary across the selected algorithms.

ANN with engineered features by Khosravi *et al.* (2019)

Khosravi *et al.* [8] were one of the first to apply artificial neural networks to the problem of MER signal classification during DBS, although other common machine learning frameworks such as random forests and support vector machines have been applied to the problem since 2006 [17]. As with these earlier ML approaches, Khosravi *et al.* 's framework was designed to be feature-based, using the power of a number of specific frequency bands, several statistical features regarding spike (i.e. spike rate, standard deviation of inter-spike pauses, etc...), and raw signal features (i.e. curve length, number of zero-crossings, etc...) which were commonly used in previous frameworks [17]. The network created by Khosravi *et al.* [8] contains 10 hidden layers each with 50 neurons leading to a total of 6,025,601 weights. In order to control for potential overfitting, both L1 regularisation and DropOut are used during the training process.

As stated earlier, this algorithm is reminiscent of the traditional feedforward artificial neural networks architecture, the multi-layer perceptron, in that the input to the network is treated as a fixed-length vector with each layer having direct access to all the neurons in the previous layer. This has distinct advantages in that the MER signal is actively reduced to a (relatively) small number of components that either efficiently summarise the signal as a whole (i.e. power, number of zero crossings), have some known causal relationship with the underlying anatomy (i.e. spike frequency), or both (i.e. frequency band powers). In theory, this should render the framework more generalisable, provided that said input representations can be computed in the exact same way for different databases. However, this is complicated by differences in the acquisition parameters of the databases (notably the bandpass filter parameters) and the fact that differing class probabilities may effect the ideal parameters in the neural network even if the input representations were exactly equivalent.

Separable CNNs by Peralta *et al.* (2020) and their Bayesian extensions by Martin *et al.* (2021)

Shortly after the first application of artificial neural networks to MER classification, Peralta *et al.* [9] were the first to apply convolutional neural networks to the same problem. The first step in their approach was to represent the signal as a spectrogram, capturing the frequency related information from previous ML approaches while maintaining the entire signal. Afterwards, a series of five "computation blocks," each consisting of a drop-out, convolution, non-linear activation, and max-pooling layer. The final output is then calculated by a single linear layer applied to the flattened results. Separable convolutions were used to limit the number of parameters in the model, leading to a total of 16,752 weights.

By treating the MER signal as a full signal rather than summarising it through a series of engineered features, it is possible for the neural network itself to discover more immediately informative or less redundant features connecting the MER signal to the underlying anatomy. It should be noted that augmenting the input (e.g. by providing the entire signal spectrogram, rather than only the signal itself) can sometimes simplify this process.

This network architecture was then extended by Martin *et al.* [10], using a recurrent neural networks and a Bayesian frameworks to update the prediction of the network as more signal became available, leading to a framework capable of classifying arbitrary length MERs while increasing the number of weights to 25,208. Martin *et al.* found that the Bayesian approach, despite several simplifying assumptions, worked significantly better than the use of recurrent units, as the necessary features could be learnt in the initial convolutional layers and aggregated across time in a simple way without necessitating the greater flexibility (and parameterisation) of recurrent units.

In comparison to previous approaches, both of these networks were designed to be lightweight and use an order of magnitude fewer parameters than the network proposed by Khosravi *et al.* [8].

Traditional CNN by Hosny *et al.* (2021)

Hosny *et al.* [12] also independently created a CNN architecture for MER classification, although following more closely to a traditional CNN architecture (i.e. alternating between convolution, non-linear activation, and max-pooling before applying a small fully-connected network to the flattened results) although with the addition of Batch Normalisation to make the training process more stable. The architecture contained three convolutional layers and two linear layers, leading to a total of 11,277,149 weights. Unlike in the previous architecture proposed by Peralta *et al.* [9], the signal was not pre-processed into a spectrogram but rather provided directly to the network. The lack of spectrogram input likely required the addition of more convolutional layer blocks in order to develop sufficient expressiveness to capture different features in the frequency-domain, leading to the highest degree of parameterisation seen in the investigated networks.

CNN with self-attention by Xiao *et al.* (2022)

A more recent conceptual addition to artificial neural networks is the concept of self-attention, which was first applied to the problem of MER classification by Xiao *et al.* [13]. This network has a similar structure as the CNNs described in the previous two subsections, although it uses addition "CBAM" layers (which combine a channel self-attention module followed by a spatial self-attention module) multiple times throughout the network. These modules involve a much larger amount of non-linearity thus allowing for a lower number of channels to be used for the intermediate images in the network.

The benefit of this narrower architecture is that it requires fewer weights, a total of 2,158,607, making it lighter than the previous CNN approach [13] despite still having a larger total number of layers. This is because, unlike Hosny *et al.* , spectrogam input was used to simplify the initial network, allowing for a relatively narrow network to be used without diminishing its capability to learn important signal features.

2.3 Training Approaches

Generalisability is a particularly important area of research for deep learning as networks often show a high sensitivity to distributional shifts between the training data and the data on which they are eventually applied [18]. Some aspects of distributional shifts can be estimated and thus addressed by augmenting the training procedure. However, characterising distributional shifts in real data, especially for highly specialised problems such as MER signal classification, to a degree where they can be modelled in the training process is not always possible. Given that, it is often hard to distinguish between the degree of generalisability of a framework from the unknown degree of distributional shift used to measure said generalisability.

The most immediate test of generalisability for any particular trained model is to apply it directly to previously unseen from a new distribution. By applying a trained model directly, one can heuristically evaluate the robustness of the patterns learned in said model to the distribution shift between the training and testing databases. However, such a test lacks baseline information such as whether or not the two databases

are fundamentally different in terms of inherent difficulty. This limitation, along with the desire to improve performance on the new database, motivation fine-tuning the network in order to overcome this distributional shift.

Transfer learning

In the context of CNNs, transfer learning is a technique used for training in which the weights for a network trained on a particular problem are used to initialise the weights for a new network for a different problem [19]. The motivation behind this is that many problems share a common set of useful features, especially low-level ones, and thus relatively little additional data is required to adapt these features to new problem domains. In the parlance of distribution shift, transfer learning allows for the fine-tuning of representations to adhere to proximal ones in the new shifted distribution using limited additional data.

The last set of experiments performed thus uses the weights learnt using the larger London database with fine-tuning being performed using the smaller Rennes databases.

Train-from-scratch

In order to determine a reference accuracy for each database, we have also retrained the network from a purely random initialisation. By performing this type of re-training, we can also measure the change in accuracy with respect to the number of training iterations in order to determine how much time (i.e. number of training epochs) is saved through training techniques such as transfer learning.

2.4 Evaluation metric and statistical analysis

The performance of each method is calculated via the Balanced Accuracy:

$$BA = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$$

which avoids the issue of class imbalance as significantly more signals were collected from outside the STN rather than inside it, noting that the balanced accuracy is equivalent to regular accuracy if the number of signals in each class is equal.

To analyse the quantitative results given the number of factors in the experiment, we use a multifactorial analysis of variance model using scipy's statsmodel library's [1] ordinary least squares (OLS) model. The factors in this model include the algorithm, the training type, the database being evaluated on, the interactions between these three factors, and finally the specific fold id (in order to account for inter-fold variability). Given these interaction terms, we used Type 1 ANOVA. This analysis was performed only over the Rennes datasets with the results from the London database being given as a reference for the optimum model quality.

All experiments were performed using 5-fold cross-validation due to the relatively small sizes of the databases involved. Each network received a set 550 training

[1] https://www.statsmodels.org/stable/index.html

iterations which was considered sufficient to ensure the convergence of network training.

3 Results

The quantitative results from the 5-fold cross-validation are shown in Table 1. For each database, the folds were split according to the patient (i.e. all data arising from a single patient appears in a single fold) in order to prevent data leakage. For the generalisation experiments, the final network trained on the London database (i.e. the one trained on the last fold) was used to initialise the network for the Rennes databases. The results of the multifactorial ANOVA test are given in Table 2 specifically showing that the algorithm, dataset, training type, and interaction between the dataset and training type are the only significant factors affecting the performance.

		Rennes 1 (2015-2022)		
Algorithm	London	D.A.L.	T.L.	T.F.S.
Khosravi *et al.* [8]	68.1±9.1	51.3±1.3	63.9±11.4	68.5±9.3
Peralta *et al.* [9]	79.9±1.8	67.4±6.3	79.2±2.6	79.1±2.6
Martin *et al.* [10]	81.0±2.4	72.8±3.4	79.1±2.3	79.8±1.6
Hosny *et al.* [12]	77.5±1.9	73.8±1.9	72.7±11.1	77.5±1.5
Xiao *et al.* [13]	74.6±2.1	68.1±2.5	75.8±1.1	75.3±1.4
		Rennes 2 (2005-2015)		
		D.A.L.	T.L.	T.F.S.
Khosravi *et al.* [8]		54.9±5.2	50.6±1.1	57.1±6.8
Peralta *et al.* [9]		67.1±8.9	68.8±10.0	69.6±6.1
Martin *et al.* [10]		65.0±6.2	69.2±10.8	73.1±3.0
Hosny *et al.* [12]		64.0±6.6	62.2±8.8	64.1±6.6
Xiao *et al.* [13]		65.3±4.8	66.8±6.8	62.3±1.1

Table 1: Quantitative results for Balanced Accuracy (%) across algorithms and database or generalisation types (D.A.L. - direct application from London database, T.L. - transfer learning, T.F.S. - train from scratch). Values are shown as mean ± standard deviation across the testing folds.

Factor	DoF	F	p-value
Algorithm	**4**	**24.23**	$\mathbf{1.85 \times 10^{-14}}$
Dataset	**1**	**55.24**	$\mathbf{2.27 \times 10^{-11}}$
Training	**2**	**9.04**	$\mathbf{2.28 \times 10^{-4}}$
Algorithm × Dataset	4	0.56	6.86×10^{-1}
Algorithm × Training	8	0.99	4.47×10^{-1}
Dataset × Training	**2**	**4.84**	$\mathbf{1.10 \times 10^{-2}}$
Algorithm × Dataset × Training	8	0.88	5.34×10^{-1}
Testing Fold ID	9	1.13	3.47×10^{-1}
Residual	112		

Table 2: Multifactoral ANOVA results table. Rows marked in bold are considered statistically significant.

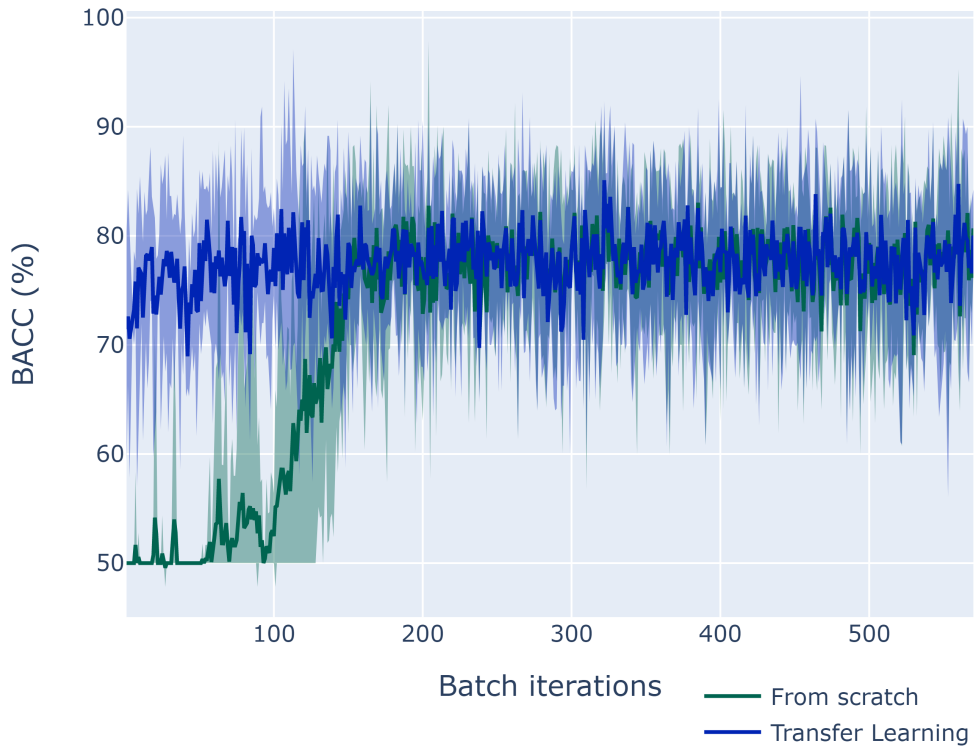

Fig. 1: Training curve showing the evolution of the BACC for the network proposed by Peralta *et al.* [9] for the Rennes database (2015-2022) initialised random (green - train from scratch) or from the weights learned from the London database (blue - transfer learning).

Figure 1 shows the evolution of the BACC over the course of the training process on a secondary dataset. It would appear that transfer learning's primary advantage is that it improves the speed at which these networks can be adapted to a new, similar domain, rather than reflective of a particular improvement in the final accuracy. This latter point is further confirmed by the numerical results in Table 1 where the difference between transfer learning and training from scratch is not only insignificant (as per a paired two-tailed t-test giving a value of $p = 0.181$) but does not even have a consistent sign. This speed increase can be highly beneficial as network training usually incurs a higher computational burden and energy cost.

4 Discussion

As hypothesised, generalisation is indeed problematic as given by the decrease of 9.54% on average for the first Rennes database and 12.96% for the second Rennes database when using parameters learnt on the London database. Although the domain shift appears to always result in a reduction in accuracy, the amount of said reduction is

highly variable with some results, such as the network proposed Hosny *et al.* [12], differing by less than 5%.

According to our analysis, there was a significant interaction between the database and the training type, meaning that the specifics of the dataset affect how much performance can be regained by modifying the training type. Regardless of the original algorithm, the reduction in accuracy caused by the domain shift can be corrected for on larger databases using re-training techniques, notably transfer learning, which improve the performance to a degree comparable to the original database, a difference of between -1.2% to 0.7% in terms of balanced accuracy. For smaller databases, such as the second Rennes database, the improvements resulting from re-training are much more modest to non-existent. This could be the result of fundamental annotation differences between the datasets, with one including more annotation uncertainty and thus fundamentally lower maximum accuracies.

Unsurprisingly given the large differences in network structure and parameterisation, some methods significantly outperformed others in general according to our ANOVA results. In terms of differences between algorithms, with the exception of [8], the difference in performance between algorithms is reduced when applied directly to a new database. Despite not finding significant evidence of this (i.e. the interaction between algorithm and training type was not significant), it does appear that certain algorithms appear to improve faster after retraining on a smaller dataset (specifically [9, 10] possibly due to their much lighter parameterisation making them less susceptible to overfitting. Further investigation would be needed to verify this observation.

One of the limitations of this work is a large portion of the literature on MER classification still uses more traditional machine learning techniques rather than deeply learned artificial neural networks [3–6, 11, 17]. Although the first test of generalisation (the direct application of a method trained on a different dataset) can still easily be applied, it is more difficult to determine if simple methods for retraining using these previous parameters as a starting-point can be applied as well.

4.1 Future work

This study into the generalisability of deep learning approaches to MER classification in the context of DBS electrode implantation provides a preliminary look into some of the challenges one needs to consider for the implementation of these algorithms in practice. It also raises several important theoretical questions that could further contextualise and inform the results. Notably, we have evidence that distribution shift within similar domains (e.g. MERs acquired and annotated in different centres using different protocols) not only affects the distribution of the input but also the annotation uncertainty. One area of future work is to collect this information through measuring the inter- and intra-operator variability of the annotations in these databases. This additional information would not only verify whether or not the automated approaches perform with equivalent accuracy as human operators, but also provide a basis for improving these algorithms through better adaptation towards noisy labels [20] or even leveraging Bayesian approaches [21] for predicting the annotation uncertainty for a given input MER signal directly rather than a singular label.

5 Conclusion

Generalisability is a critical concern in the application of modern deep learning methods to biomedical signal processing. Measuring how an algorithm copes with distribution shifts that may exist between the context in which its training data was collected and the context in which it is applied is crucial to ensuring its safe application. Certain aspects of distribution shift can however be corrected for, notably using transfer learning as a computationally efficient approach to update learned weights to adapt to their new context.

In this article, several algorithms which span the gamut of deep learning approaches towards signal processing have been implemented, trained, evaluated, and re-trained across three different databases of interventional micro-electrode recordings collected during deep brain stimulation electrode implantation. This experiment quantifies the degree of generalisability and susceptibility to distribution shift across similar domains characteristic of how these algorithms may be used in clinic. Although every algorithm we investigated was susceptible to performance degradation (and many to approximately the same degree) we also found that simply fine-tuning the weights (i.e. transfer learning) can allow these algorithms to quickly regain their previous higher performance.

Declarations

- **Funding:** Thibault Martin is supported through the Fondation France Parkinson.
- **Conflict of interest:** The authors have no conflicts of interest to declare.
- **Availability of data and materials:** Data was collected through collaboration with Western University, Canada, and the Rennes University Hospital Centre, France.
- **Ethics approval:** The London database was approved by the Research Ethics Board at Western University, Canada (REB # 109045). Both Rennes databases were approved by the Rennes University Hospital Centre ethics committee (Ethical authorisation declaration n2205295).

References

[1] Lee, P.S., Weiner, G.M., Corson, D., Kappel, J., Chang, Y.-F., Suski, V.R., Berman, S.B., Homayoun, H., Van Laar, A.D., Crammond, D.J., *et al.*: Outcomes of interventional-mri versus microelectrode recording-guided subthalamic deep brain stimulation. Frontiers in neurology **9**, 241 (2018)

[2] Iess, G., Bonomo, G., Levi, V., Aquino, D., Zekaj, E., Mezza, F., Servello, D.: Mer and increased operative time are not risk factors for the formation of pneumocephalus during dbs. Scientific Reports **13**(1), 9324 (2023)

[3] Rajpurohit, V., Danish, S.F., Hargreaves, E.L., Wong, S.: Optimizing computational feature sets for subthalamic nucleus localization in dbs surgery with feature selection. Clinical Neurophysiology **126**(5), 975–982 (2015)

[4] Schiaffino, L., Muñoz, A.R., Martínez, J.G., Villora, J.F., Gutiérrez, A., Torres, I.M., *et al.*: Stn area detection using k-nn classifiers for mer recordings in parkinson patients during neurostimulator implant surgery. In: Journal of Physics: Conference Series, vol. 705, p. 012050 (2016). IOP Publishing

[5] Valsky, D., Marmor-Levin, O., Deffains, M., Eitan, R., Blackwell, K.T., Bergman, H., Israel, Z.: Stop! border ahead: Automatic detection of subthalamic exit during deep brain stimulation surgery. Movement Disorders **32**(1), 70–79 (2017)

[6] Vargas Cardona, H.D., Álvarez, M.A., Orozco, Á.A.: Multi-task learning for subthalamic nucleus identification in deep brain stimulation. International Journal of Machine Learning and Cybernetics **9**, 1181–1192 (2018)

[7] Karthick, P., Wan, K.R., Qi, A.S.A., Dauwels, J., King, N.K.K.: 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)

[8] Khosravi, M., Atashzar, S.F., Gilmore, G., Jog, M.S., Patel, R.V.: 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) https://doi.org/10.1109/JTEHM.2020.2969152 . Conference Name: IEEE Journal of Translational Engineering in Health and Medicine

[9] Peralta, M., Bui, Q.A., Ackaouy, A., Martin, T., Gilmore, G., Haegelen, C., Sauleau, P., Baxter, J.S., Jannin, P.: 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), pp. 888–893 (2020). IEEE

[10] Martin, T., Peralta, M., Gilmore, G., Sauleau, P., Haegelen, C., Jannin, P., Baxter, J.S.: 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)

[11] Coelli, S., Levi, V., Del Vecchio, J.D.V., Mailland, E., Rinaldo, S., Eleopra, R., Bianchi, A.M.: 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**(1), 016003 (2021)

[12] Hosny, M., Zhu, M., Gao, W., Fu, Y.: Deep convolutional neural network for the automated detection of Subthalamic nucleus using MER signals. Journal of Neuroscience Methods **356**, 109145 (2021) https://doi.org/10.1016/j.jneumeth.2021.109145 . Accessed 2021-09-20

[13] Xiao, L., Li, C., Wang, Y., Si, W., Lin, H., Zhang, D., Cai, X., Heng, P.-A.:

Amplitude-frequency-aware deep fusion network for optimal contact selection on STN-DBS electrodes. Science China Information Sciences **65**(4), 140404 (2022) https://doi.org/10.1007/s11432-021-3392-1 . Accessed 2022-04-06

[14] Peralta, M., Jannin, P., Baxter, J.S.: Machine learning in deep brain stimulation: A systematic review. Artificial Intelligence in Medicine **122**, 102198 (2021)

[15] Pooch, E.H., Ballester, P., Barros, R.C.: Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2, pp. 74–83 (2020). Springer

[16] Park, K.H., Sun, S., Lim, Y.H., Park, H.R., Lee, J.M., Park, K., Jeon, B., Park, H.-P., Kim, H.C., Paek, S.H.: Clinical outcome prediction from analysis of microelectrode recordings using deep learning in subthalamic deep brain stimulation for parkinsons disease. PloS one **16**(1), 0244133 (2021)

[17] Wan, K.R., Maszczyk, T., See, A.A.Q., Dauwels, J., King, N.K.K.: A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for parkinson's disease. Clinical Neurophysiology **130**(1), 145–154 (2019)

[18] Taori, R., Dave, A., Shankar, V., Carlini, N., Recht, B., Schmidt, L.: Measuring robustness to natural distribution shifts in image classification. Advances in Neural Information Processing Systems **33**, 18583–18599 (2020)

[19] Ribani, R., Marengoni, M.: A survey of transfer learning for convolutional neural networks. In: 2019 32nd SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T), pp. 47–57 (2019). IEEE

[20] Huang, Y., Bai, B., Zhao, S., Bai, K., Wang, F.: Uncertainty-aware learning against label noise on imbalanced datasets. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 6960–6969 (2022)

[21] Zheng, R., Zhang, S., Liu, L., Luo, Y., Sun, M.: Uncertainty in bayesian deep label distribution learning. Applied Soft Computing **101**, 107046 (2021)