Identifying Biases in a Multicenter MRI Database for Parkinson's Disease Classification: Is the Disease Classifier a Secret Site Classifier?

Raissa Souza, Anthony Winder, Emma A. M. Stanley, Vibujithan Vigneshwaran, Milton Camacho, Richard Camicioli, Oury Monchi, Matthias Wilms, and Nils D. Forkert

***Abstract*—Sharing multicenter imaging datasets can be advantageous to increase data diversity and size but may lead to spurious correlations between site-related biological and non-biological image features and target labels, which machine learning (ML) models may exploit as shortcuts. To date, studies analyzing how and if deep learning models may use such effects as a shortcut are scarce. Thus, the aim of this work was to investigate if site-related effects are encoded in the feature space of an established deep learning model designed for Parkinson's disease (PD) classification based on T1-weighted MRI datasets. Therefore, all layers of the PD classifier were frozen, except for the last layer of the network, which was replaced by a linear layer that was exclusively re-trained to predict three potential bias types (biological sex, scanner type, and originating site). Our findings based on a large database consisting of 1880 MRI scans collected across 41 centers show that the feature space of the established PD model (74% accuracy) can be used to classify sex (75% accuracy), scanner type (79% accuracy), and site location (71% accuracy) with high accuracies despite this information never being explicitly provided to the PD model during original training. Overall, the results of this study suggest that trained image-based classifiers may use unwanted shortcuts that are not meaningful for the actual clinical task at hand. This finding may explain why many image-based deep learning models do not perform well when applied to data from centers not contributing to the training set.**

***Index Terms*—Biases, deep learning, shortcut learning.**

I. Introduction

SHARING multicenter medical imaging data is assumed to lead to several benefits, such as increasing interdisciplinary collaboration, avoiding duplication of clinical trials and data collection (e.g., healthy control cohort), supporting novel medical insights, and training more robust machine learning (ML) models for computer-aided diagnosis [1]. Thus, there have been significant investments in creating large multicenter databases, such as the Parkinson's Progression Markers Initiative (PPMI)[1] and the UK Biobank [2], among others. Compiling data from multiple acquisition sites can improve the diversity and size of the database as a whole, which is assumed to be beneficial for training and testing of ML models, especially for medical image analysis where data is often rare at single sites [3], [4], [5], [6]. The utilization of large multicenter datasets is expected to enhance the performance of trained ML models, enable and improve the identification of subtle disease expressions in the data, and increase their generalizability to unseen data. However, a lack of intra-site data variation may result in spurious correlations between 'site-related effects' (image features) and the target label, which ML models could exploit. For MRI-based neuroimaging data, these site-related effects or biases could originate from complex factors such as biological differences (e.g., age, sex, pathological characteristics) and non-biological differences (e.g., sample size and bias, scanner type and model, acquisition parameters, head coil, and magnetic field properties) across sites. In the context of disease classification, imbalanced distributions of pathological characteristics between sites can also cause ML models to learn spurious associations between site-related effects and a patient's disease state, introducing biases and potentially resulting in a concealed site classifier rather than an accurate disease classifier [7].

Manuscript received 9 August 2023; revised 5 December 2023; accepted 6 January 2024. Date of publication 10 January 2024; date of current version 5 April 2024. This work was supported in part by the Parkinson Association of Alberta, in part by the Hotchkiss Brain Institute, in part by the Canadian Consortium on Neurodegeneration in Aging (CCNA), in part by the Canadian Open Neuroscience Platform, in part by the Natural Sciences and Engineering Research Council of Canada, in part by the Canada Research Chairs program, in part by the River Fund at Calgary Foundation, in part by the Canadian Institutes for Health Research, and in part by the Tourmaline Chair in Parkinson disease. Image data used were provided in part by the OASIS-3 Project (Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352), in part by the PPMI-a, public-private partnership funded by Michael J. Fox Foundation, in part by the OpenfMRI database (accession number ds000245), and in part by the UK Biobank (application number 77508). *(Mathias Wilms and Nils D. Forkert shared last authorship.) (Correspondent author: Raissa Souza.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Conjoint Health Research Ethics Board at the University of Calgary under Application No. REB21-0454, and performed in line with the declaration of Helsinki.

Please see the Acknowledgment section of this article for the author affiliations.

This article has supplementary downloadable material available at https://doi.org/10.1109/JBHI.2024.3352513, provided by the authors.

Digital Object Identifier 10.1109/JBHI.2024.3352513

[1] [Online]. Available: https://www.ppmi-info.org/.

Potential shortcut learning based on site-related effects has been observed in various neuroimaging applications. For instance, sociodemographic factors such as socioeconomic status, race, and pubertal development stage were found to influence the accuracy of a deep learning model trained to classify sex using multicenter T1-weighted brain MRI datasets from pediatric subjects [8]. Disparities in performance associated with site-related effects, such as scanner types and magnetic field, have also been identified in an ML model for Alzheimer's disease diagnosis using multicenter T1-weighted brain images [9]. Additionally, it has been demonstrated that scanner vendors can be distinguished in a multicenter functional MRI database [10].

Techniques to minimize site-related effects, such as bias mitigation [11] and data harmonization methods [12], are nowadays frequently utilized in medical image analysis. However, explicitly identifying all site-related effects (confounders) is very challenging and often impossible. Moreover, even with the most advanced mitigation strategies, a successful and complete removal of unwanted site-related effects may not be possible, as it would necessitate the disentanglement of site-related effects from disease-related features without harming the performance of the model. However, a ML model may rely heavily on information that co-occurs with the site-related effects when learning disease characteristics, which can lead to bad performance when applied to data from centers that did not contribute to the training set.

Currently, most studies [12], [13], [14] that apply techniques to reduce site-related effects in neuroimaging data simply assume that some effects exist without explicitly specifying them. Moreover, most previous works in this domain restrict their evaluations to a quantification of improved accuracy and/or generalizability. In contrast to that, studies analyzing if and how deep learning models really exploit such effects as a shortcut are scarce. In T1-weighted brain MRI, differences in imaging protocols, scanner vendors/models, and magnetic field strengths have been associated with two factors (1) image quality (e.g., signal-to-noise ratio and contrast-to-noise ratio) and (2) brain anatomy (e.g., cortical thickness and brain volumes)[15], [16]. Moreover, site characteristics, such as differences in the number of datasets available for ML model training, diagnostic criteria, and training and experience of the medical experts establishing the ground truth can lead to systematic site-specific distinguishable features. These non-biological differences (site biases) may serve as potential shortcuts for a deep learning model in its primary task of disease classification.

Biological variations related to the manifestation of the disease under study in medical images could also introduce bias(es) in the model. For instance, in the case of Parkinson's disease (PD), researchers have observed significant sex differences, with males exhibiting earlier disease onset, higher occurrence, and higher prevalence [17], [18]. Males also tend to experience more severe motor and cognitive symptoms compared to females. A deep learning model may exploit these inherent biological biases (disease biases) to distinguish individuals with PD from healthy individuals using MRI data, particularly if there are variations in patient characteristics across sites. Even worse, powerful deep learning models may use a complex combination of biological and non-biological biases as shortcuts.

Therefore, this work aims to investigate if site-related effects are encoded in the feature space of an established deep learning model designed for PD classification based on T1-weighted brain MRI [19]. For our analysis, we utilize a diverse database of 1880 subjects with imaging data acquired across 41 different sites. Our major contributions include: (1) an in-depth analysis of how deep learning models encode site-related information in the feature space of the original disease classifier and (2) an evaluation of which effect(s) are potentially used as shortcuts. Such an analysis provides, for the first time, a better understanding of how biased variables are encoded in deep learning models, which could eventually result in a more appropriate selection and/or development of mitigation strategies.

II. Material and Methods

In this work, we used an established deep learning-based PD classifier trained using a large and diverse multicenter database to identify potential shortcuts caused by differences in multiple biological and non-biological features across centers. Thus, we investigated if the feature space of a deep learning model trained to classify patients with PD and healthy subjects (HS) holds information related to the originating site that acquired the data, the scanner type used to acquire the data, and the sex of the participants without ever being provided this information directly during training. The presence of such information would suggest that the model may use them as shortcuts for the disease classification.

A. Dataset

All analyses described in this work were performed using a multicenter PD database consisting of 1880 T1-weighted MRI scans (867 PD and 1013 HS) collected across 41 centers[2], [3], [4][2], [20], [21], [22], [23], [24], [25]. Datasets available at each center were split into 80% for training and 20% for testing, totaling in 1478 images (680 PD [418 males and 262 females] and 798 HS [497 males and 301 females]) used for training and 402 images (187 PD [124 males and 63 females] and 215 HS [138 males and 77 females]) used for testing. Known non-biological variabilities in this database include differences in scanner vendors (i.e., Siemens, GE, and Phillips), 19 scanner models, and MRI magnet strength (i.e., magnetic field strengths of 1.5T or 3.0T). Center-specific details can be found in the supplementary material. Briefly described, the in-house developed pre-processing pipeline (same as used in [19]) that was applied to all datasets consisted of skull-striping [26], resampling to an isotropic resolution of 1 mm using linear interpolation, bias field correction [27], affine image registration to the PD25-T1-MPRAGE-1mm brain atlas [28], and cropping to 160 × 192 × 160 voxels via center-cropping to reduce background information.

Each center received ethics approval from their local ethics board and received written informed consent from all the participants in accordance with the declaration of Helsinki.

[2] [Online]. Available: https://www.ppmi-info.org/

[3] [Online]. Available: https://www.mcgill.ca/neuro/open-science

[4] [Online]. Available: https://openneuro.org/datasets/ds000245/versions/00001

B. Parkinson's Disease Classifier

The state-of-the-art simple fully convolutional network (SFCN) [29], which achieved an accuracy of 78.8% in detecting PD using the same multicenter T1-weighted MRI datasets, was utilized [19] in this work. Due to the necessary exclusion of participants with unavailable site and scanner information for this analysis, only 867 of the original 1051 patients with PD and 1013 of the original 1026 healthy subjects from [19] were included in this study. As a result, the SFCN was retrained on this reduced number of included subjects, but still achieved a comparable classification accuracy of 74% on the test set. The Adam optimizer with an initial learning rate of 0.001, a decay rate of 0.003, and batch size 5 was used during training. The best model (lowest binary cross entropy testing loss) was saved for evaluation based on early stopping with patience of 10 epochs.

C. Shortcut Learning Analysis

To identify possible shortcuts in the trained model, all layers of the pre-trained PD classifier model were frozen, except for the final layer, which was replaced with a customized linear layer designed to classify the specific biases of interest, namely sex ($n = 2$), site ($n = 41$), and scanner ($n = 19$). In doing so, we aimed to determine if bias-related information is present in the penultimate layer of the PD classifier, despite this information not being directly available during model training. In this context, it is important to highlight that the linear layer serves as the model's output. Each node in this layer represents a fixed feature related to the likelihood of a specific sex, site, or scanner. Consequently, the linear layer cannot create new intermediate features for the bias variables beyond the frozen layers. It essentially places a linear decision plane through the existing and fixed feature space. For sex classification, we employed a binary output layer with a sigmoid activation function, while for site and scanner type classification, multi-class output layers with softmax activation functions were used, comprising of 41 and 19 neurons, respectively. For all cases, the Adam optimizer and early stopping were used as for the original PD model as described in the previous section. Fig. 1 summarizes the workflow.

Once the bias training was completed, our next step was to investigate whether individual classifications based on sex, site, or scanner were associated with a significantly higher or lower occurrence of PD. This analysis aimed to confirm potential model shortcuts. To accomplish this, we conducted a Fisher exact test to examine the relationship between PD status and sex, and Fisher-Freeman-Halton exact tests to assess the association between PD status and site or scanner. For the potential shortcut features that were significantly associated with PD status, we performed linear regression-based mediation analyses to determine the extent to which the association between a patient's true and predicted PD status is mediated by the perceived values of these shortcut features, as determined by the sex, site, and scanner classifiers. More specifically, we performed linear regressions as described in [30] to estimate the total and direct effects. Then, we ran the PROCESS macro [31] to perform significance testing on the indirect effects. Finally, we performed bootstrapping as recommended in [32] to compute the 95% confidence intervals for the mediation effect of a derived variable representing the occurrence of PD disease. Our code is available at https://github.com/RaissaSouza/bias-identification and all statistical analyses were performed using the IBM SPSS v29 software package.

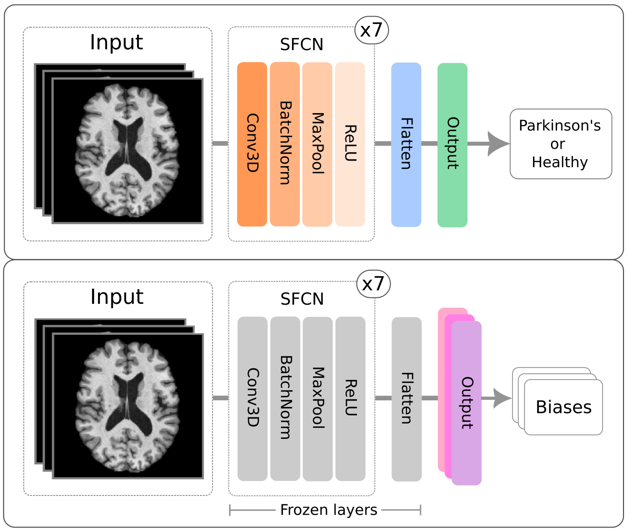

Fig. 1. Workflow diagram of the Parkinson's disease classifier and each bias of interest.

D. Evaluation Metrics

To quantitatively evaluate our results, we computed the accuracy, F1-score, and performed regression-based mediation analyses for the PD classifier as well as for each of the potential shortcut classifiers (sex, site, and scanner). Model-level F1-scores were computed as the weighted average of the F1-score across all the model's class labels, where each class label was weighted according to its frequency within the test data. Additionally, we utilized confusion matrices to visually represent sensitivity and specificity, providing further insights into the performance of the classifiers. For qualitative analysis, we employed UMAP projections [33], which enable an exploration of how the information in the feature space was organized and clustered based on the potential shortcut variables of interest.

III. Results

The results of this work suggest that the feature space of the PD classifier indeed encodes information from the biases investigated: sex, site, and scanner, even with this simple re-training setup of only replacing and retraining the final layer of the PD model with a single dense layer 'head' while all other layers remained frozen. As can be seen in Table I, the feature space of the model that was initially trained to classify patients with PD and healthy subjects (HS) without ever being given any direct information about sex, site, or scanner type can be used directly to classify sex, site, and scanner with high accuracies (75%, 71%, and 79%) similar to the PD classifier (74%).

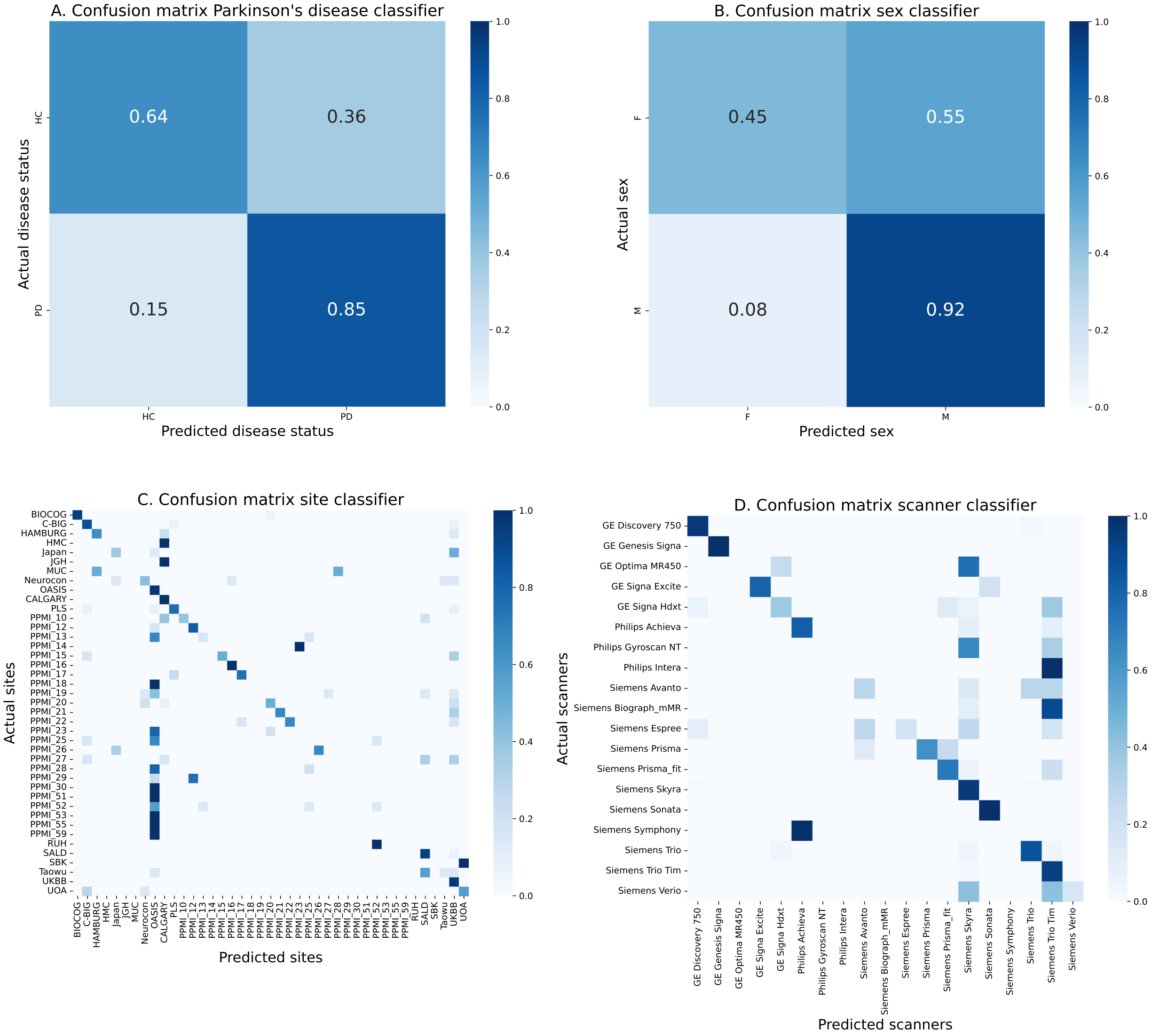

Fig. 2. Confusion matrices showing sensitivity and specificity for each classifier evaluated in this study.

TABLE I
ACCURACY AND F1-SCORE FOR PARKINSON'S DISEASE (PD) CLASSIFIER AND SHORTCUT LEARNING MODELS

Bias of interest	Models		Accuracy	F1-score
	Features	Classifier		
n/a	PD	PD	0.74	0.73
Sex	PD (frozen)	Sex	0.75	0.74
Site	PD (frozen)	Site	0.71	0.65
Scanner	PD (frozen)	Scanner	0.79	0.75

Even though the database as a whole was well-balanced (867 PD and 1013 HS) and the F1-score (73%), which considers the proportion of PD and HS in the testing set, was similar compared to the accuracy (74%) for PD vs. HS classification, the model achieved a considerably higher sensitivity (85%) than specificity (64% - Fig. 2(a)). This demonstrates that the PD classifier was better at identifying patients with Parkinson's disease than classifying healthy subjects.

Similar results were observed when analyzing the presence of sex information in the PD classifier feature space. Fig. 2(b) shows that a better performance was achieved for males (92%) than females (45%), although their representation is well-balanced (males [542 patients with PD and 635 HS] and females [325 patients with PD and 378 HS]) for the disease status, resulting in a non-significant Fisher exact test ($p = 0.676$, Table II).

Although the classification of the originating site achieved the lowest accuracy (71%) and F1-score (65%), the results achieved are still considerably better than the chance level. As can be seen in the confusion matrix (Fig. 2(c)), approximately 20 sites were correctly identified based on the PD classifier feature space. Furthermore, the occurrence of PD differed significantly between sites ($p<0.001$) and was a partial mediator of the association between true and predicted PD status (Table II). Notably, the indirect effect size of the mediation analysis, which represents the proportion of the total association between the true and predicted PD status that is attributable to the mediation

TABLE II
STATISTIC TEST AND MEDIATION ANALYSIS FOR SEX, SITE, AND SCANNER AS POTENTIAL SHORTCUT FEATURES USED BY THE PD CLASSIFIER

Potential shortcut feature	Fisher/ FFH exact test	Mediation Analysis				
		Direct Effect		Indirect Effect		
-	p-value	Effect Size (Std. Error)	p-value	Effect Size (Std. Error)	95% Lower CI	95% Upper CI
Sex	0.676	-	-	-	-	-
Site	0.001	1.490 (0.283)	0.001	1.499 (0.225)	1.118	2.002
Scanner	0.001	2.003 (0.261)	0.001	0.728 (0.173)	0.460	1.138

FFH = Fisher-Freeman-Halton; CI = Confidence interval.

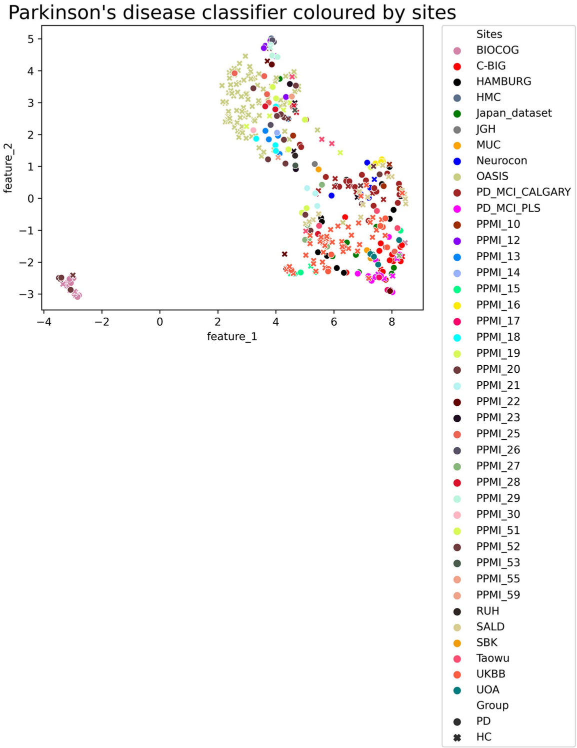

Fig. 3. Although sites present more complex variability due to the combined biological and non-biological effects, clusters relating to different sites can be seen. The small cluster on the left side may be associated with bad image quality, as ring artifacts were identified in the data provided by BIOCOG.

effect, was estimated to be greater than the direct effect size, which represents the proportion of the total association that is attributable to true PD status alone. These results suggest that even though multicenter data present more complex variability due to the combined effects (total number of samples, scanners, sex, and number of patients and healthy subjects), sites are still distinguishable (Fig. 3).

As shown in Table I, the feature space of the classifier originally trained for the PD classification task also achieved a higher accuracy (79%) and F1-score (75%) for identifying scanners than for the task it was primarily trained for, which, again, is considerably higher than the chance level. This result can be further confirmed visually in Fig. 2(d) that shows that only 5 out of 19 scanner types were completely misclassified. Moreover, Fig. 4 demonstrates how well different scanners were clustered in the feature space of the PD classifier in the UMAP, suggesting a potential strong use of the scanner type to learn the PD classification task. Quantitatively, the occurrence of PD also differed significantly between scanners, with this variable acting as a significant (partial) mediator between true and predicted PD status. Combined with the mediation effects shown for sites, these results support the hypothesis that information related to conditions under which a patient's data was acquired may contribute significantly to predicting their PD status. For instance, the model may associate Parkinson's patients and healthy subjects to the scanner used to acquire the data rather than biologically plausible brain morphology differences.

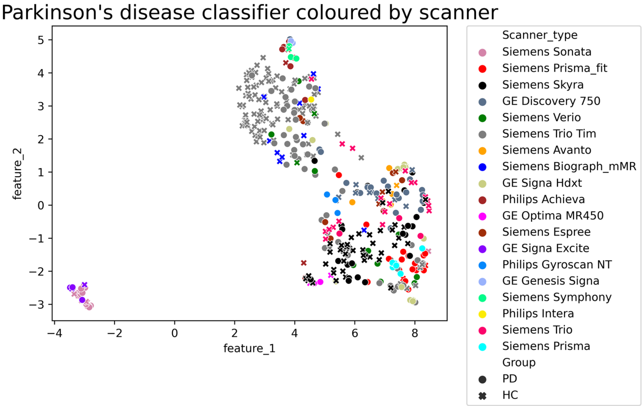

Fig. 4. Clusters relating to different scanner types can be seen. The small cluster on the left side may be associated with bad image quality, as ring artifacts were identified in the data provided by BIOCOG, which scanned their patients using a Siemens Sonata scanner.

IV. DISCUSSION

In this work, we presented an in-depth analysis investigating if and how a deep learning model designed to classify patients with PD and healthy subjects encode biological (sex) and non-biological (site and scanner) information in the feature space that may not be causally relevant from a pathological perspective. Most importantly, the results of this work show that by using the feature space of the PD classifier, it is possible to classify sex, sites, and scanners with high accuracy, although this information was never explicitly provided to the original classification model. Thus, our results raise the question if the baseline PD model actually learned to reliably distinguish patients with PD subjects using biologically relevant information or if it relies, at least to some extent, on shortcuts to achieve this task, given the high performance of the bias classifiers.

Our results suggest that the PD classifier can identify males with PD with higher accuracy, potentially because of the more frequent manifestation of the disease in males [17], [18]. This observation might suggest that considering sex as a feature for the classifier could be justified, given the sex-dimorphic nature of PD. However, our mediation analysis revealed that sex is not acting as a mediator between true and predicted PD status.

This result is not surprising, considering that the entire database contains a balanced representation of patients and healthy subjects when stratified by sex. As a result, the lack of a mediation effect from sex on the PD classifier's predictions is consistent with the balanced representation of male and female subjects in the dataset. Nevertheless, it is essential to be cautious when a trained model exhibits disproportional subgroup performance (e.g., males vs. females), as this could suggest model unfairness and limit its reliability for clinical use.

Our findings showed that the site acts as a significant (partial) mediator between true and predicted PD status. However, it is important to note that the site bias is not extensively explored in the literature and is often considered the same as the scanner bias. Within this context, it should be emphasized that a single site can also use multiple scanners so that the two biases are not the same. Moreover, it should be pointed out that considerable differences exist in the patient distribution among the sites included in this work. For example, OASIS provided only healthy subjects, introducing the bias of class imbalance. On the other hand, the Biocog study focused on examining representative patients across the entire disease spectrum, without dementia at baseline, encompassing biases typically associated with patients who agree to participate in studies. In contrast, PPMI primarily targeted de-novo patients with PD, leading to a potential bias related to differences in disease stage. This disparity in disease stage and class imbalance across sites could potentially be utilized as a shortcut associated with the site variable that is independent but may interact with the scanner bias.

The ability to identify scanner types from neuroimaging data has also been observed previously [10]. However, while those findings generally support our results, we demonstrated that scanner identification was possible even when the representations learned by the deep learning model were not even optimized for this task, which is in stark contrast to previous work that specifically trained ML models for this task. More importantly, we established that this variable acts as a significant (partial) mediator between true and predicted PD status. Thus, future research direction should aim to reduce the model's reliance on scanner type (i.e., unlearning scanner type effects), investigating how bias encoding may change as a function of different bias mitigation techniques. A proposed approach for unlearning scanner type effects may involve a combination of multi-task learning and adversarial debiasing methods, as, for example, presented in [12]. Their training procedure involved predicting brain age, classifying scanner type, and unlearning the scanner type using four loss functions, including a confusion loss from the adversarial debiasing field. Although their results showed promise, they did not initially demonstrate that scanner types were used as shortcuts by the model. Additionally, their work included only three sites with similar data distribution, which limits its direct application to our dataset comprising 41 unique sites. Each site in our study provides distinct distributions, including variations in scanner types, data acquisition subjects (patients/healthy subjects), and other factors. Moreover, the authors [12] suggest that scanner unlearning in a disease classification model requires either an overlapping distribution (a dataset segment with subjects having PD and healthy subjects from all intended scanners for harmonization) or a healthy cohort with data from all targeted scanners. This requirement, aimed at preserving disease-related information, limits our use of an overlapping distribution or a healthy cohort for scanner unlearning in our unique database.

Overall, our study reveals that even when information unrelated to PD pathology is not explicitly provided to the model during training, it can still discover spurious correlations and act as a hidden site classifier rather than a disease classifier. These findings emphasize the importance of conducting more comprehensive model analyses to determine whether site-related effects are contributing to shortcut learning in a classifier. The simplicity of our approach allows it to be applicable to any deep learning model task and database, making it a valuable tool for identifying potential shortcuts in trained models. Recognizing the origins of shortcuts or biases is vital for the formulation and examination of applicable mitigation strategies. In this context, a recent study by Stanley et al. [34] showed that applying bias mitigation strategies in the absence of bias can detrimentally impact a model's performance. Therefore, using this methodology to help identify biases can facilitate the selection or development of more targeted bias mitigation and data harmonization strategies, especially when dealing with multicenter datasets. Ultimately, ensuring that a classifier relies on genuine disease effects rather than spurious correlations is crucial for its clinical utility across diverse contexts. In a larger context, the results of this study may explain to some degree why so many deep learning models that perform well in cross-validation evaluations or based on an internal test set, fail to achieve a clinical meaningful accuracy when applied to data from centers that did not contribute any of the data used for the initial training of deep convolutional neural networks used for disease classification.

It is essential to highlight some of the limitations of this work. First, our shortcut investigation approach was based on adding a single dense layer ('head') to the penultimate layer of the frozen PD classifier, leading to a linear analysis. However, this linear approach also presents an advantage, as it prevented that the models learn additional complex representations from the data and changing the actual feature space of the original model. Thus, this limitation may also be beneficial for our ability to detect potential shortcuts in the PD classifier, making the study more indicative of their presence. Second, it is important to note that our analysis was focused on a single established PD classifier model. Hence, the results might differ when considering other deep learning models or other disease models. However, it should be noted in this context that the multicenter database used in this work is comparably large and was acquired in many centers compared to the data used for many other deep learning models. It may be argued that deep learning models trained based on fewer datasets acquired in a smaller number of centers may be even more prone to using biases as shortcuts. Third, our study solely employed one MRI sequence, T1-weighted images, which are not quantitative. Other image modalities may show more or less severe non-biological biases that could be exploited as shortcuts. However, it may be argued that even quantitative imaging modalities have some technical bias, for example, due to the reconstruction kernel used, so that the findings of this work are still widely relevant. Finally, while the shortcut analysis method employed in this study can be applicable to other datasets and biases of interest, it is essential to acknowledge that our research was conducted on a single disease with only three potential biases being investigated in detail. Therefore, the findings may not fully represent all possible scenarios in other contexts, but the same framework can be used to evaluate different scenarios and biases.

V. Conclusion

While advancements in bias mitigation and data harmonization are important methodological contributions that aid in minimizing biases in trained models and may prevent potential shortcut learning to some extent, our work highlights the importance of using more detailed model analyses in order to determine whether site-related effects are being used for shortcut learning in a classifier. This can allow for bias mitigation and data harmonization strategies to be selected or developed in ways that are more targeted for the specific task, especially when using multicenter datasets. Overall, it is crucial that a classifier makes use of true disease effects and not spurious correlations so that it may be clinically useful in a broad context.

Acknowledgment

Authors' Affiliations

Raissa Souza, Emma A. M. Stanley, and Milton Camacho are with the Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada, and with the Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 1N4, Canada, and also with the Biomedical Engineering Graduate Program, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: raissa.souzadeandrad@ucalgary.ca; emma.stanley@ucalgary.ca; milton.camachocamach@ucalgary.ca).

Anthony Winder and Vibujithan Vigneshwaran are with the Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada, and also with the Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: anthony.winder@ucalgary.ca; vibujithan.vigneshwa@ucalgary.ca).

Richard Camicioli is with the Neuroscience and Mental Health Institute and Department of Medicine (Neurology), University of Alberta, Edmonton, AB T6G 2R3, Canada (e-mail: rcamicio@ualberta.ca).

Oury Monchi is with the Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 1N4, Canada, and with the Department of Clinical Neurosciences, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada, and with the Department of Radiology, Radio-oncology and Nuclear Medicine, Université de Montréal, Montréal, QC H3C 3J7, Canada, and also with the Centre de Recherche, Institut Universitaire de Gériatrie de Montréal, Montréal, QC H3C 3J7, Canada (e-mail: oury.monchi@umontreal.ca).

Matthias Wilms is with the Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 1N4, Canada, and with the Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 1N4, Canada, and also with the Department of Pediatrics and Department of Community Health Sciences, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: matthias.wilms@ucalgary.ca).

Nils D. Forkert is with the Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada, and with the Hotchkiss Brain Institute, University of Calgary, Calgary, AB T2N 1N4, Canada, and with the Department of Clinical Neurosciences, Cumming School of Medicine, University of Calgary, Calgary, AB T2N 1N4, Canada, and also with the Alberta Children's Hospital Research Institute, University of Calgary, Calgary, AB T2N 1N4, Canada (e-mail: nils.forkert@ucalgary.ca).

References

[1] F. Sardanelli, M. Alì, M. G. Hunink, N. Houssami, L. M. Sconfienza, and G. D. Leo, "To share or not to share? Expected pros and cons of data sharing in radiological research," *Eur. Radiol.*, vol. 28, pp. 2328–2335, 2018.

[2] C. Sudlow et al., "UK biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Med.*, vol. 12, 2015, Art. no. e1001779.

[3] M. Wilms, P. Mouches, J. J. Bannister, D. Rajashekar, S. Langner, and N. D. Forkert, "Towards self-explainable classifiers and regressors in neuroimaging with normalizing flows," in *Proc. Int. Workshop Mach. Learn. Clin. Neuroimaging*, 2021, pp. 23–33.

[4] M. Wilms et al., "Invertible modeling of bidirectional relationships in neuroimaging with normalizing flows: Application to brain aging," *IEEE Trans. Med. Imag.*, vol. 41, no. 9, pp. 2331–2347, Sep. 2022.

[5] R. Souza, P. Mouches, M. Wilms, A. Tuladhar, S. Langner, and N. D. Forkert, "An analysis of the effects of limited training data in distributed learning scenarios for brain age prediction," *J. Amer. Med. Inform. Assoc.*, vol. 30, pp. 112–119, 2023.

[6] K. Amador, M. Wilms, A. Winder, J. Fiehler, and N. D. Forkert, "Predicting treatment-specific lesion outcomes in acute ischemic stroke from 4D CT perfusion imaging using spatio-temporal convolutional neural networks," *Med. Image Anal.*, vol. 82, 2022, Art. no. 102610.

[7] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nature Mach. Intell.*, vol. 2, pp. 665–673, 2020, doi: 10.1038/s42256-020-00257-z.

[8] E. A. Stanley, M. Wilms, and N. D. Forkert, "Disproportionate subgroup impacts and other challenges of fairness in artificial intelligence for medical image analysis," in *Proc. Workshop Ethical Philos. Issues Med. Imag. Int. Workshop Multimodal Learn. Clin. Decis. Support MICCAI Workshop Topological Data Anal. Biomed. Imag.*, 2022, pp. 14–25.

[9] A. A. Chen, J. C. Beer, N. J. Tustison, P. A. Cook, R. T. Shinohara, and H. Shou, "Mitigating site effects in covariance for machine learning in neuroimaging data," *Hum. Brain Mapping*, vol. 43, pp. 1179–1195, 2022.

[10] D. M. Nielson et al., "Detecting and harmonizing scanner differences in the ABCD study - annual release 1.0," *BiorXiv*, 2018, doi: 10.1101/309260.

[11] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in - *Proc. AAAI/ACM Conf. AI, Ethics, Soc.*, 2018, pp. 335–340.

[12] N. K. Dinsdale, M. Jenkinson, and A. I. Namburete, "Unlearning scanner bias for MRI harmonisation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2020, pp. 369–378.

[13] V. M. Bashyam et al., "Deep generative medical image harmonization for improving cross-site generalization in deep learning predictors," *J. Magn. Reson. Imag.*, vol. 55, pp. 908–916, 2022.

[14] M. Liu et al., "Style transfer using generative adversarial networks for multi-site MRI harmonization," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2021, pp. 313–322.

[15] C. L. Tardif, D. L. Collins, and G. B. Pike, "Sensitivity of voxel-based morphometry analysis to choice of imaging protocol at 3 T," *NeuroImage*, vol. 44, pp. 827–838, 2009.

[16] F. H. Alhazmi, O. M. Abdulaal, A. A. Qurashi, K. M. Aloufi, and V. Sluming, "The effect of the MR pulse sequence on the regional corpus callosum morphometry," *Insights Imag.*, vol. 11, 2020, Art. no. 17.

[17] S. Meoni, A. Macerollo, and E. Moro, "Sex differences in movement disorders," *Nature Rev. Neurol.*, vol. 16, no. 2, pp. 84–96, 2020.

[18] I. N. Miller and A. Cronin-Golomb, "Gender differences in Parkinson's disease: Clinical characteristics and cognition," *Movement Disord.*, vol. 25, pp. 2695–2703, 2010.

[19] M. Camacho et al., "Explainable classification of Parkinson's disease using deep learning trained on a large multi-center database of T1-weighted MRI datasets," *NeuroImage: Clin.*, vol. 38, 2023, Art. no. 103405.

[20] S. Duchesne et al., "The Canadian dementia imaging protocol: Harmonizing national cohorts," *J. Magn. Reson. Imag.*, vol. 49, pp. 456–465, 2019.

[21] H. J. Acharya, T. P. Bouchard, D. J. Emery, and R. M. Camicioli, "Axial signs and magnetic resonance imaging correlates in Parkinson's disease," *Can. J. Neurological Sci.*, vol. 34, pp. 56–61, 2007.

[22] S. Lang et al., "Network basis of the dysexecutive and posterior cortical cognitive profiles in Parkinson's disease," *Movement Disord.*, vol. 34, pp. 893–902, 2019.

[23] P. J. LaMontagne et al., “OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer disease,” *medRxiv*, 2019, doi: 10.1101/2019.12.13.19014902.

[24] D. Wei et al., “Structural and functional MRI from a cross-sectional southwest university adult lifespan dataset (SALD),” *BiorXiv*, 2018.

[25] A. S. Talai, J. Sedlacik, K. Boelmans, and N. D. Forkert, “Utility of multi-modal MRI for differentiating of Parkinson’s disease and progressive supranuclear palsy using machine learning,” *Front. Neurol.*, vol. 12, 2021, Art. no. 648548.

[26] F. Isensee et al., “Automated brain extraction of multisequence MRI using artificial neural networks,” *Hum. Brain Mapping*, vol. 40, pp. 4952–4964, 2019.

[27] N. J. Tustison et al., “N4ITK: Improved N3 bias correction,” *IEEE Trans. Med. Imag.*, vol. 29, no. 6, pp. 1310–1320, Jun. 2010.

[28] Y. Xiao et al., “A dataset of multi-contrast population-averaged brain MRI atlases of a Parkinson’s disease cohort,” *Data Brief*, vol. 12, pp. 370–379, 2017.

[29] H. Peng, W. Gong, C. F. Beckmann, A. Vedaldi, and S. M. Smith, “Accurate brain age prediction with lightweight deep neural networks,” *Med. Image Anal.*, vol. 68, 2021, Art. no. 101871.

[30] R. M. Baron and D. A. Kenny, “The moderator-mediator variable distinction in social psychological research. conceptual, strategic, and statistical considerations,” *J. Pers. Social Psychol.*, vol. 51, pp. 1173–1182, 1986.

[31] A. F. Hayes, *Introduction to Mediation, Moderation, and Conditional Process Analysis - A Regression-Based Approach*. New York, NY, USA: Guilford Press, 2022.

[32] P. E. Shrout and N. Bolger, “Mediation in experimental and nonexperimental studies: New procedures and recommendations,” *Psychol. Methods*, vol. 7, pp. 422–445, 2002.

[33] L. McInnes, J. Healy, N. Saul, and L. Großberger, “UMAP: Uniform manifold approximation and projection,” *J. Open Source Softw.*, vol. 3, 2018, Art. no. 861.

[34] E. A. M. Stanley et al., “Towards objective and systematic evaluation of bias in medical imaging AI,” 2023, *arxiv:2311.02115*.