PREDICTIVE MODELING OF

EARLY STAGE PARKINSON'S DISEASE

CHARLES STEVENS LEGER

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES IN
PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN PSYCHOLOGY
YORK UNIVERSITY
TORONTO, ONTARIO

OCTOBER 2020

Abstract

Background: Early stage (preclinical) detection of Parkinson's disease (PD) remains challenged yet is crucial to both differentiate it from other disorders and facilitate timely administration of neuroprotective treatment as it becomes available.

Objective: In a cross-validation paradigm, dual binary classifications analyses were conducted: early PD versus controls and early PD versus SWEDD (scan without evidence of dopaminergic deficit). It was hypothesized that five distinct model types using combined non-motor and biomarker features would distinguish early PD from controls with > 80% cross-validated AUC, but that the diverse nature of SWEDD would reduce early PD versus SWEDD CV classification AUC and alter model-based rank of predictor importance among model types.

Methods: Baseline data was acquired from the Parkinson's Progressive Markers Initiative (PPMI). Logistic regression, general additive (GAM), decision tree, random forest and XGBoost models were fitted using non-motor clinical and biomarker features. Randomized train and test data partitions were used. Model classification CV performance was compared using the area under the curve (AUC), accuracy, sensitivity, specificity and the Kappa statistic.

Results: All five models achieved >.80 AUC CV accuracy to distinguish early PD from controls using non-motor clinical and biomarker features. The GAM (CV AUC .928, sensitivity .898, specificity .897) and XGBoost (CV AUC .923, sensitivity .875, specificity .897) models were the top classifiers. Performance across all models was consistently lower in the early PD/SWEDD analyses. The two highest performing models were XGBoost (CV AUC .863, sensitivity .905, specificity .748) and random forest (CV AUC .822, sensitivity .809, specificity .721); XGBoost detection of non-PD SWEDD matched 1-2yr curated diagnoses in 81.25% (13/16) cases. In both early PD/control and early PD/SWEDD analyses, and across all models, olfactory function was the single most important feature to classification; rapid eye movement behaviour disorder and cognition were the next most commonly high ranked features. Alpha-synuclein was a feature of import to early PD/control but not to early PD/SWEDD classification and daytime sleepiness was antithetically important to the latter but not former.

Interpretation: Non-motor clinical and biomarker variables enable high CV discrimination of early PD versus controls but are less effective discriminating early PD from SWEDD.

For Madeleine

ACKNOWLEDGEMENTS

Joseph DeSouza, Tuan Cao-Huu, and Monique Herbert

Like all of us, Joe, Tuan and Monique are defined by a spectrum of traits unique to each. Thankfully, these educators, with their differing and enviable skills, provided a rich tapestry of advice in support of my efforts.

Special thanks and genuine gratefulness to Psychology Graduate Program Staff members Lori-Ann Santos, Freda Soltau, and Barbara Thurston. Also, constructive and thoughtful feedback from David Flora and Suzanne MacDonald was sincerely appreciated.

TABLE OF CONTENTS

Abstract ii
Dedication iii
Acknowledgements iv
Table of contents v
List of Tables vi
List of Figures vii
Citation and contributions viii

Chapter One: Introduction 1
The basal ganglia, imaging and deep brain stimulation 2
Scan without evidence of dopaminergic deficit (SWEDD) 7
The staging of Parkinson's disease (PD), and PD non-motor prodromal variables 8
Predictive models 14
Prediction of Parkinson's pathology using non-motor clinical and biomarker variables 17
Chapter Two: Methods 19
2.0 Procedures 19
2.1 Participant data 20
2.2 Feature elimination and hyper-parameter tuning 21
2.3 Clinical assessments and cerebral spinal fluid assays 22
2.4 Screening 23
2.5 Imaging 23
2.6 Statistical analyses 23
2.7 Classification performance metrics 27
Chapter Three 29
3.1 Descriptive statistics and t-tests 29
3.2 Bivariate analyses 33
Chapter Four: Early PD versus control model classification 36
4.1 Decision tree classification 37
4.2 Logistic regression and general additive model classification 40
4.3 Random forest classification 46
4.4 XGBoost classification 49
4.5 Long-term conversion of SWEDD to PD 52
4.6 Longitudinal diagnosis vs. model predicted SWEDD to PD conversion 54
Chapter Five: Early PD versus SWEDD classification, SMOTE-based models 56
5.1 Decision tree early PD/SWEDD classification 57
5.2 Logistic regression, GAM early PD/ SWEDD classification 60
5.3 Random forest early PD/SWEDD classification 66
5.4 XGBoost classification early PD/SWEDD 69
5.5 Early PD/SWEDD: model Prediction and Long-Term Diagnosis 72
Chapter Six: Performance summary 74
Chapter Seven: Discussion 81
Conclusion 86
Limitations 87
Funding 87
Bibliography 88
Appendices (Supporting information sections I-III) 102

LIST OF TABLES

Early PD/control

Table 1: Descriptive statistics and t-tests, early PD/controls....30
Table 2: Descriptive statistics and t-tests, early PD/SWEDD....31
Table 3: Predictor importance: decision tree model, early PD/control....37
Table 4a: Decision tree confusion matrices, early PD/control....39
Table 4b: Decision tree performance metrics, early PD/control....39
Table 5: Predictor importance: logistic model, early PD/control....40
Table 6: Logistic regression early PD/controls (parameters)....42
Table 7a: Logistic regression GLM and GAM confusion matrices, early PD/control....45
Table 7b: Logistic regression GLM and GAM performance metrics, early PD/control....45
Table 8: Predictor importance: random forest, early PD/control....46
Table 9a: Random forest confusion matrices, early PD/control....48
Table 9b: Random forest performance metrics, early PD/control....48
Table 10: Predictor importance, XGBoost, early PD/control....49
Table 11a: XGBoost confusion matrices, early PD/control....51
Table 11b: XGBoost performance matrices, early PD/control....51
Table 12: XGBoost and GAM models applied to SWEDD/control test data....52

Early PD/SWEDD, SMOTE-based models

Table 13: Predictor importance: decision tree model early PD/SWEDD....57
Table 14a: Decision tree confusion matrices, early PD/SWEDD....59
Table 14b: Decision tree performance metrics, early PD/SWEDD....59
Table 15: Predictor importance, logistic regression, early PD/SWEDD....61
Table 16: Logistic regression early PD/SWEDD (parameters)....62
Table 17a: Logistic regression GLM, GAM confusion matrices, early PD/SWEDD....65
Table 17b: Logistic regression GLM, GAM performance metrics, early PD/SWEDD....65
Table 18: Predictor importance, random forest, early PD/SWEDD....67
Table 19a: Random forest confusion matrices, early PD/SWEDD....68
Table 19b: Random forest performance metrics, early PD/SWEDD....68
Table 20: Predictor importance, XGBoost....69
Table 21a: XGBoost confusion matrices, early PD/SWEDD....71
Table 21b: XGBoost performance metrics, early PD/SWEDD....71
Table 22: Performance summary....76

LIST OF FIGURES

Figure A 3
Figure B 4
Figure C 9

Descriptive statistics and bivariate analysis

Figure 1: Analyses pipeline flow chart 20
Figure 2: Density plots 29
Figure 3: Clinical predictor boxplots 32
Figure 4: Biological predictor boxplots 33
Figure 5: Predictor correlations 34
Figure 6: Mean dopamine active transporter (DAT) values against hyposmia 35

Early PD/Control

Figure 7: Tree plot, early PD versus controls 38
Figure 8: Decision tree AUC plot 40
Figure 9: Effects plots (logistic regression) 43
Figure 10: Logistic regression GLM and GAM AUC plots 46
Figure 11: Out of bag error plot 47
Figure 12: Random forest AUC plot 49
Figure 13: XGBoost AUC plot 51
Figure 14: GAM, XGBoost models AUC for SWEDD/control validation data 53

Early PD/SWEDD

Figure 15: Tree plot, Early PD/ SWEDD 58
Figure 16: Decision tree AUC plot 60
Figure 17: Effects plots (logistic regression) 63
Figure 18: Logistic regression GLM, GAM AUC plots 66
Figure 19: Out of bag error rates compared 67
Figure 20: Random forest AUC plot 69
Figure 21: XGBoost AUC plot 71
Figure 22: Predictor importance ranking Early PD/controls 75
Figure 23: Predictor importance ranking Early PD/SWEDD 76
Figure 24: GAM, XGBoost AUC plots 78
Figure 25: Random forest, XGBoost AUC plots 78

Citation and contributions

Much of the current work (chapters 3 to 6) was recently published in Frontiers in Neurology:

> Citation: Leger, C., Herbert, M., DeSouza, J.F.X. (2020). Non-motor Clinical and Biomarker Predictors Enable High Cross-Validated Accuracy Detection of Early PD**.** Front. Neurol., **https://doi.org/10.3389/fneur.2020.00364**

As stipulated in the online, published version, Joseph DeSouza (JD) and Monique Herbert (MH) helped proof the work. Charles Leger (CL) originated study aims and scope; CL acquired data from the PPMI, built, tested, re-tested all models, and arrived at conclusions. MH vetted models built by CL; JD monitored overall research progress and managed the formation of the dissertation committee.

Chapter One

1. INTRODUCTION

Parkinson's disease

As a neuropathological disorder, the occurrence of Parkinson's disease (PD) is second only to Alzheimer's disease (Schapira, 2009). Despite being one of the most studied pathologies, the exact etiology of PD is unknown, its root cause remains elusive, and PD has a 10-20% misdiagnosis rate (Hess & Okun, 2016). Accuracy of diagnosis is largely a function of disease duration: more accurate diagnosis in advanced disease with more pronounced symptoms, but lower accuracy in the early preclinical (e.g. premotor dysfunction) stage (Adler, 2011). Further, there is not a definitive test of incipient idiopathic PD pathology. A long-standing sine qua non distinguishing PD is dopamine neuron loss and a paucity of dopamine. But in early, premotor stage PD, the initiation and progression of such neuron loss can proceed undetected and predate pathology diagnosis by several years (Cummings et al., 2011; Fahn et al., 2004). Currently, clinical assessment and diagnosis depends mainly on satisfactory response to levodopa and identification of cardinal PD motor symptoms (i.e. resting tremor, rigidity, postural instability, bradykinesia, asymmetrical onset) (Hughes, Daniel, & Lees,

Abbreviations

Semi-continuous scales
ESS: Epworth daytime sleepiness scale
GDS: Geriatric depression scale
MoCA: Montreal cognitive assessment
RBD: rapid eye movement behaviour disorder
RBDQ: Rapid eye movement behaviour disorder questionnaire
UPSIT: University of Pennsylvania Smell Identification Test
UPDRS: Universal Parkinson's disease rating scale

Biologics
α-synuclein: a neuronal protein (the main component of Lewy bodies)
$A\beta_{1\text{-}42}$: beta-amyloid
pTau: phosphorylated tau protein
tTau: total tau protein

Classes
Early PD: early stage Parkinson's' disease
HC: healthy controls
SWEDD: scan without evidence of dopamine deficit

Models
GAM: general additive model
Logistic model: logistic regression model
Decision tree: classification decision tree
Random forest: ensemble of forest of decision trees
XGBoost: extreme gradient boosting (ensemble of regression trees)

Metrics:
AUC: receiver operating characteristic
ACC: general accuracy
FP: false positive
FN: false negative
Kappa: Cohen's Kappa statistic
SN: sensitivity
SP: specificity
TP: true positive
TN: true negative

Resampling
K-fold: a resampling method where k is the number of folds
OOB: out of bag samples (used in Random forest)
SMOTE: synthetic minority oversampling technique

Note, models, resampling methods, and performance metrics are defined in Supporting information III

2001). However, cardinal PD motor symptoms while typically present in relatively advanced disease can be undetectable and absent in subclinical early stage PD pathology.

Cardinal motor disruption symptoms in conjunction with identification of so-called hallmark nigrostriatal Lewy body (LB) lesions have become a PD red flag. A LB is a proteinaceous abnormal intracytoplasmic inclusion mainly consisting of excessive accumulation of abnormally folded α-synuclein but also deposits of ubiquitin(Braak & Braak, 1991; Braak et al., 2003; Warren et al., 2013). The mainstay medical management of PD is a pharmaceutical carbidopa/levodopa preparation: it treats motor symptoms only, and, over time it gradually becomes ineffective. In addition, in some cases, carbidopa/levodopa may have no effect.

Motor dysfunction in PD is largely attributed to LB-synucleinopathy accumulation and associated dopaminergic nigrostriatal neuron loss particularly in substantia nigra pars compacta (SNc) and putamen neurons of the nigrostriatal pathway (Cummings et al., 2011; Fahn et al., 2004). The approximate location of the SNc, putamen and other basal ganglia structures is provided in figure (A). There is evidence of selective loss of SNc dopamine neurons (Brichta & Greengard, 2014) with motor dysfunction but strangely preserved dopamine neurons in the ventral tegmental area (Dauer & Przedborski, 2003) (the ventral tegmental area is just lateral to the SNc; not shown in figure A). Moreover, PD pathology is not confined to altered dopaminergic neuron status but likely also involves alteration of other neurotransmitter levels related to reduced dopamine levels. It was recently reported that reduced dopamine does not, as previously assumed, increase acetylcholine availability but antithetically reduces acetylcholine availability (McKinley et al., 2019). Both dopamine and acetylcholine are involved not only in motor but also learning behaviors (Rizzi & Tan, 2017). Not surprisingly, reduced levels of these neurotransmitters is accompanied by up to 80% dementia prevalence in PD. Mild cognitive impairment in PD ranges between 20-50% (Goldman et al., 2018).

The basal ganglia, imaging and deep brain stimulation

Although LB-synucleinopathy has become virtually synonymous with PD, LBs are not exclusive to PD, are not a "hallmark" that certifies the existence of PD pathology, and can occur without lifetime manifestation of PD symptoms (Dickson et al., 2008). The SNc is a midbrain structure part of the basal ganglia. Also included in the basal ganglia are the pars reticula (a gamma-aminobutyric acid [GABAergic] nucleus that forms the other

segment of the substantia nigra) caudate, globus pallidus (external and internal segments), putamen and subthalamic nucleus. The caudate and putamen are collectively referred to as the striatum, and between them lies the nucleus accumbens (the central component of the ventral striatum). Figure (A) is based on the Montreal Neurological Institute (MNI152) average of 152 subject images. This is a standard space image and includes all basal ganglia structures except the nucleus accumbens.

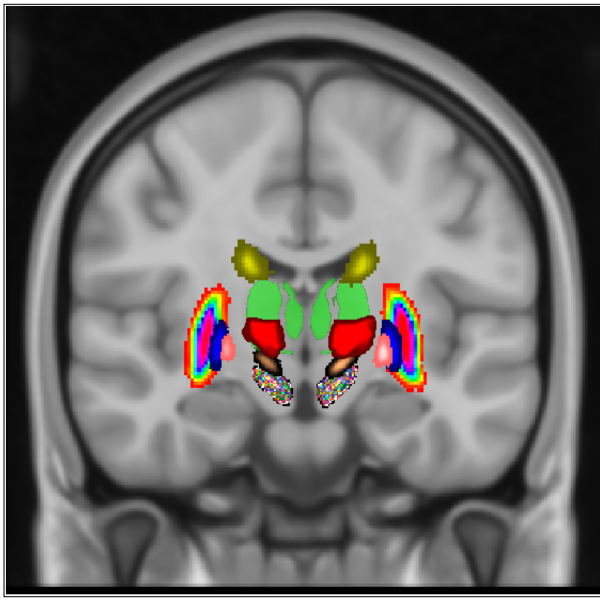


FIG A | Basal ganglia. MNI152 probabilistic image. Color codes from top to bottom: yellow = caudate, green= thalamus, rainbow = putamen, blue = globus pallidus external (GPe), pink = globus pallidus internal (GPi), solid red = ventral intermediate thalamic nucleus (VIM), copper = subthalamic nucleus (STN), mixed = substantia nigra (both pars compacta [SNc], and pars reticulata [SNr]). Note, the globus pallidus and striatum jointly account for the largest portion of the basal ganglia. This MNI152 image was created using FSLeyes (https://git.fmrib.ox.ac.uk/fsl/fsleyes/fsleyes/). Also, the thalamus and VIM are not part of the basal ganglia. The thalamus serves as a landmark; the VIM is a target of particular interventions, such as deep brain stimulation.

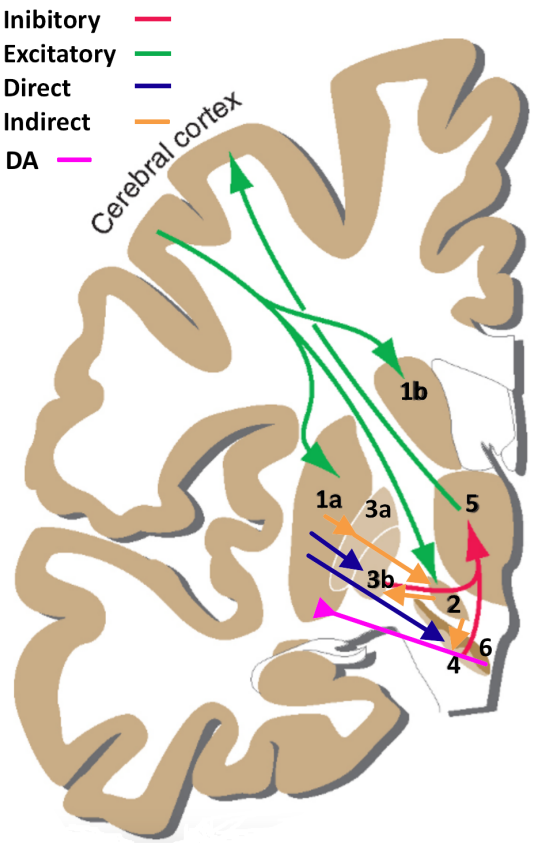


FIG B | Direct and indirect basal ganglia circuitry (Adapted from Nolte, 2009). Reprinted with permission from Mosby Inc., an affiliate of Elsevier Inc. Red = main inhibitory; green = main excitatory . 1a = putamen; 1b = caudate; 2 = subthalamic nucleus (STN); 3a = external globus pallidus segment (GPe); 3b = internal globus pallidus (GPi); 4 = reticular part of the substantia nigra (SNr); 5 = the thalamus (i.e. ventral lateral anterior [VL], ventral anterior [VA] and intralaminar [IL] nuclei); 6 = substantia nigra pars compacta (SNc); DA = dopamine. Note, all direct and indirect pathway connections are inhibitory (GABAergic), except indirect STN (gluaminergic, excitatory) pathway connections to the GPi and SNr. The main cortical inputs to the striatum and STN are excitatory (gluaminergic). The main output basal ganglia nuclei, the GPi and SNr, are inhibitory. SNc projections of dopamine have an excitatory effect on D1 receptors involved in the direct pathway, but an inhibitory effect on D2 receptors involved in the indirect pathway. It warrants mention, that with the exception of the primary auditory and visual cortices, all neocortex areas project to the striatum.

Figure (B) conveys a schematic of the direct and indirect pathway (primate originated) model of normal motor control involving basal ganglia and cortex interaction (Albin, Young, & Penney, 1989; Alexander, Delong, & Strick, 1986; M. R. Delong, 1990) a model that has wide acceptance but is also not fully representative of all neural interconnections related to movement (Calabresi, Picconi, Tozzi, Ghiglieri, & Di Filippo, 2014). It has also been more recently demonstrated that basal ganglia connectivity extends to parietal association and the cerebellar regions (Bostan, Dum, & Strick, 2013; Schendan, Amick, & Cronin-Golomb, 2009). Nevertheless, the broadly acknowledged direct/indirect pathway model affords a basic encapsulation, albeit an oversimplification, of basal ganglia involvement in PD. As such it warrants brief review.

Movement is modulated through a looping circuit: cortex->basal ganglia->thalamus->back to the cortex. Ostensibly, and as posited by this model, the process is initiated at the cortex (green arrows, figure B), in the premotor and supplementary motor cortices (M. R. Delong, 1990). Cortical (excitatory glutaminergic) projections,

efferents, are sent to the basal ganglia striatum (principally the putamen) and the subthalamic nucleus (STN; the STN also receives primary motor cortex input). The direct pathway (dark blue arrows, figure B) courses from the putamen (mainly) to the pallidum internal segment (GPi) and substantia nigra pars reticula (SNr). The ultimate effect of the direct pathway is to disinhibit the thalamic nuclei. By contrast, the indirect pathway (orange arrows, figure B) has an opposing role of thalamic inhibition; hence, the direct and direct pathways normally yield a balanced state of activation. The indirect pathway courses from striatum to the external pallidum segment (GPe), then to the subthalamic nucleus STN, and the latter's excitatory projections ultimately, as in the direct pathway, reach the GPi and SNr. The latter two nuclei efferents terminate in the thalamic ventral anterior nucleus (VA), intralaminar nucleus (IL), and ventral lateral nucleus anterior aspect (VL). The thalamic VA projects to supplementary and premotor cortices; the VL projects to primary motor and premotor cortices, and the IL projects to a range of frontal and parietal cortical areas(Saalmann, 2014). There is also evidence that GPi and SNr efferents reach the pedunculopontine nucleus (PPN) of the dorsal pons as well as the midbrain superior colliculus (Devito & Anderson, 1982; Rubin, McIntyre, Turner, & Wichmann, 2012) (the PPN and superior colliculus are not shown in figure B). While only mentioned here (and below) but not reviewed, bear in mind that normal levels of inhibition of target thalamic, PPN, and superior colliculus structures involve particular firing rates of the GPi and SNr efferents (M. R. Delong, 1971, 1990; Rubin et al., 2012; Wichmann et al., 1999).

The striatum cells are almost entirely GABAergic (gama-aminobutyric acid [GABA] secreting), and inhibitory (striatum interneurons are also inhibitory and many contain acetylcholine). The STN, however, is glutaminergic. The STN is the sole excitatory nucleus in the basal ganglia. The GPi and SNr are the principal basal ganglia output nuclei; they are GABAergic, inhibitory neurons.

The SNc neurons are, as stipulated earlier, dopaminergic. Striatum receptors have been differentiated as D1 and D2 G protein-coupled types (but there are other striatum receptor types). The D1 receptors activate the direct pathway; the D2 receptors activate the indirect pathway (Kandel, Schwartz, & Jessell, 2000). In functional terms, each receptor type's response is antithetical when bound to dopamine, which is largely provided by the SNc. When bound to dopamine, D1 receptor probability of firing, and hence direct pathway probability of activation is increased. By contrast, when a D2 receptor is bound to dopamine the likelihood of firing and hence

the indirect pathway likelihood of activation is reduced (M. R. Delong, 1990; Keeler, Pretsell, & Robbins, 2014; Surmeier, Ding, Day, Wang, & Shen, 2007). In the presence of normal quantities of dopamine and normal receptor function, both receptor types therefore ultimately result in disinhibition or diminished inhibition of the nuclei targeted by the GPi and SNr, notably the thalamic (VL, VA, IL) nuclei. The presence of dopamine therefore promotes thalamocortical activity. Crucially, in PD, with SNc neuron degeneration, SNc dopamine input to the striatum is insufficient. The unchecked inhibitory output of the GPi and SNr nuclei from both the direct and indirect pathways diminish activity of the targeted nuclei, which includes thalamic (VL, VA and IL) nuclei. Hence, overall thalamocortical activity is diminished. In short, in the context of the model, the direct and indirect circuit of motor control is put into disequilibrium by a dopamine shortage.

Included in figure (A), are the thalamus (green) and the ventral intermediate thalamic nucleus (VIM, red). Thalamotomy and deep brain stimulation (DBS: surgically implanted electrodes that alter abnormal activity by altering the firing rate of neurons) of the VIM was pioneered over 30 years ago (Benabid, Pollak, Louveau, Henry, & Derougemont, 1987) to treat essential tremor. DBS has become much a more common treatment than thalamotomy. Moreover, about 1/3rd of PD patients, also afflicted with essential tremor, elect to discontinue pharmacological treatment (i.e. carbidopa/levodopa) often because of tremor that is intractable to drugs (Louis, Rios, & Henchcliffe, 2010). In treatment of drug-refractory or tremor dominant PD, VIM DBS intervention has more complications and lower efficacy compared to DBS targeting of the internal globus pallidus (GPi) or subthalamic nucleus (STN) (Benabid et al., 1987; Fasano et al., 2010; Siegfried & Lippitz, 1994). The purported tremor mitigating effectiveness related to altering GPi and STN firing rate (as indicated in the latter referenced research) lends support to the direct and indirect pathway model of PD (Albin et al., 1989; M. R. Delong, 1990), though, again, the model does not fully account for PD movement dysfunction (see (Calabresi et al., 2014). A recent, non-invasive, treatment of tremor in PD involves high intensity ultrasound guided by magnetic resonance imaging. This is an innovative approach that is capable of ablating problematic structures (B. R. Shah et al., 2020). However, it is underlined that pharmacological and surgical (traditional and new non-surgical) treatments only target disease symptoms.

Dopamine, a neurotransmitter crucial to basal ganglia modulation of movement and also part of the reward system, is removed from a basal-ganglia neuron synaptic cleft and returned to the pre-synaptic terminal cytosol by dopamine active transporter (DAT). DAT, an integral membrane protein of dopamine neurons, is located in striatal axon terminal membranes where the protein structure pumps dopamine from the cleft back into the pre-synaptic terminal. DAT is localized in the SNc in both pre-synaptic (dendritic) and post-synaptic (axon terminal) plasma membranes (Nirenberg, Vaughan, Uhl, Kuhar, & Pickel, 1996). DAT protein, in an apparent calcium channel-coupled process (Cameron, Solis, Ruchala, De Felice, & Eltit, 2015), is an important regulator of dopamine receptor stimulation. Parkinson's patients have substantial depletion of dopamine neurons and striatal DAT (Kaufman & Madras, 1991; Mackie et al., 2018; Niznik, Fogel, Fassos, & Seeman, 1991; Seibyl et al., 1995).

Single-photon emission computerized tomography (SPECT) imaging is almost routinely now used to assist diagnosis and monitoring of PD progression. Often used in conjunction with the Unified Parkinson's Disease Rating Scale (Goetz et al., 2008), SPECT scans can reveal *in vivo* neurodegenerative progression of PD pathology, and typically focus on the dopaminergic nigrostriatal system. SPECT scans utilize radioisotope ligands (e.g. Ioflupane **^{123}I)** that are introduced by injection; the ligands reversibly bind to DAT. Loss of DAT accompanies SNc dopamine neuron projection degeneration, and SPECT DAT imaging can reveal nigrostriatal dopamine neuron degeneration based on loss of DAT, which corresponds to dopaminergic neuron loss (Uhl, 1992). It warrants note, that while SPECT DAT imaging is, purportedly, relatively accurate discriminating between neurodegenerative (i.e. PD) and non-neurodegenerative (e.g. supranuclar palsy, multiple system atrophy, dementia with LBs) parkinsonism (Suwijn, Verschuur, Slim, Booij, & de Bie, 2019), it should be kept in mind that SPECT DAT accuracy is typically based on comparison to clinical examination results (as opposed to post-mortem findings). As noted at the outset, PD has a 10-20% misdiagnosis rate, which makes clinical diagnosis a less than ideal diagnostic comparative reference for SPECT DAT evaluation.

Scan without evidence of dopaminergic deficit (SWEDD)

Diagnosis of PD is further complicated by the existence of a PD-lookalike category that presents with clinical symptoms of PD (e.g. some extent of motor disruption), but normal SPECT DAT imaging results, and hence

normal SPECT DAT estimated striatal dopamine neuron density. Members of this category are designated scan without evidence of dopaminergic deficit (SWEDD). Research reviewing six PD drug trials, where participants had SPECT DAT (or positron emission tomography [PET]) scans found the SWEDD category constituted a range of approximately 4-20% of all participants (Erro, Schneider, Stamelou, Quinn, & Bhatia, 2016). However, typically, most of those assigned to the SWEDD category are misdiagnosed as PD; the category members actually encompass a range of pathologies other than PD (e.g. essential tremor, which is without striatal binding neuron density loss, dystonia, psychogenic illness, fragile X permutation) (Schneider et al., 2007; Schwingenschuh et al., 2010). A positive levodopa response in a small portion of those in the SWEDD category may indeed be indicative of PD (Erro et al., 2016). A 22-month longitudinal study included investigation of before and after change in SPECT DAT striatal binding ratio (a mean measure of striatal density based on SPECT DAT), UPDRS score, and change in diagnosis of SWEDD category participants. At study termination, 40/90 (44%) SWEDD were re-diagnosed from PD to be non-PD (e.g. essential tremor, dystonia, psychogenic illness, etc.) and UPDRS scores were unchanged. The study authors concluded SWEDD participants were unlikely to be idiopathic PD candidates (Marek et al., 2014).

The staging of PD, and PD non-motor prodromal variables

More efficient detection of incipient stage PD pathology is required to both screen-out other pathologies masquerading as PD and improve timely disease management of PD pathology. An established premotor phase of early preclinical (symptom) PD has been determined (Sharma et al., 2013) to which certain premotor clinical (e.g. olfactory function) and biomarker (e.g. cerebral spinal fluid α-synuclein) variables have shown detection sensitivity. Non-motor symptoms typically antedate motor symptoms by several years (Gaig & Tolosa, 2009; Sawle, Playford, Burn, Cunningham, & Brooks, 1994), and have been convincingly linked to later emerging motor symptom pathology (Braak et al., 2003).

In seminal postmortem work(Braak et al., 2003; Braak, Ghebremedhin, Rub, Bratzke, & Del Tredici, 2004), Braak and colleagues reported a six-stage progressive development of PD pathology based on the sequential appearance of LB inclusions. The latter spread sequentially over 1-6 stages, in prion (virus-like) fashion, using neuronal pathways. In an ascending pattern, the LB-synucleinopathy initial stage (stage I) begins in

the dorsal vagal motor and glossopharyngeal nuclei of the medulla, the olfactory bulb and olfactory nucleus. In stage II, α-synuclein pathology spreads rostrally and dorsally within the medulla to the raphe nuclei and then to pontine locus coeruleus. It is not until stage III that the midbrain region, and critically, the SNc are affected (Braak et al., 2003; Braak et al., 2004). Thereafter, in stages IV-VI, LB-synucleinopathy spreads dorsally and ultimately to the neocortex. Figure (C) depicts MNI152 sagittal, coronal and axial images as well as a Gray's Anatomy (Plate 793, Public Domain, https://commons.wikimedia.org/w/index.php?curid=526636) representation of the "wandering" vagus nerve. In the figure (C) left side pane, three sites of Braak and colleagues (2003) staging are localized and numerically labeled: the vagus nerve dorsal motor nucleus (1: stage I), raphe nuclei and the locus coeruleus (2: stage II) and the SNc (3: stage III). In the right side pane of figure (C), the representation of

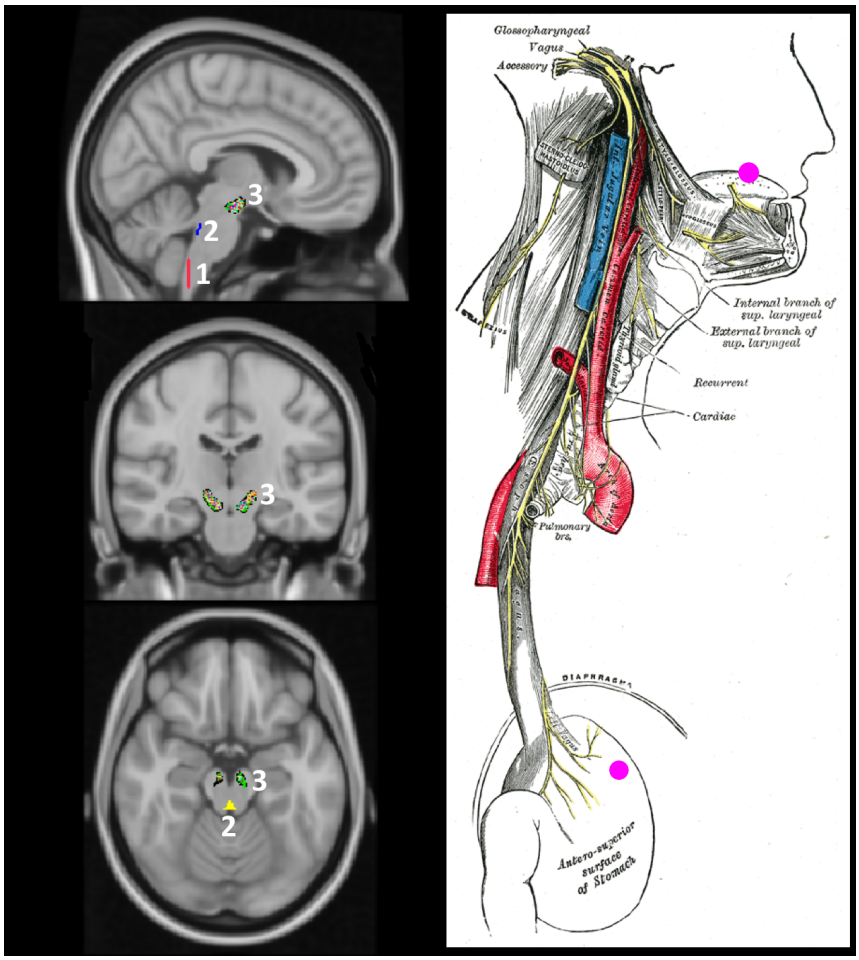


FIG C | *Braak staging sites and structures (Braak et al., 2003).* *Left pane: 1= stage I, vagus nerve dorsal motor nucleus; 2 = stage II, raphe nuclei (blue) and the locus coeruleus (yellow); 3 = stage III, SNc. MNI152 sagittal, coronal and axial images created using FSLeyes (McCarthy, 2020). Right: Gray's Anatomy (Plate 793, Public Domain, https://commons.wikimedia.org/w/index.php?curid=526636) representation of the "wandering" vagus nerve. Pink circles mark approximated pathology entry sites (olfactory and gastrointestinal) suggested by Braak and colleagues (Braak, de Vos, Bohl, & Del Tredici, 2006).*

the vagus nerve's brainstem to gastrointestinal path includes pink circles that mark approximated pathology entry sites (olfactory and gastrointestinal) suggested by Braak and colleagues (Braak et al., 2006); these sites are discussed briefly below. Note, to avoid an overly complex depiction, the figure (C) left pane does not include all staging structures specified by Braak et al (Braak et al., 2003). Also, while the locus coeruleus, raphe nuclei and SNc location and size are based on MNI152 atlases available to FSLeyes (McCarthy, 2020), the vagus dorsal motor nucleus location and proportions were estimated from a recent neuroanatomical text (Baker & Forshing, 2020).

Inconsistencies to this staging hypothesis have been demonstrated. One study, examined 71 PD tissue samples from the UK Parkinson's Society Tissue bank and while 38 (53%) had a pattern consistent with the Braak hypothesis, 5 cases did not have any indication of α-synuclein pathology and 47% did not conform to the Braak posited caudo-rostral spread of synucleinopathy (Kalaitzakis, Graeber, Gentleman, & Pearce, 2008). Other research has reported evidence that vagal dorsal motor nuclei involvement was not always present (Attems & Jellinger, 2008), and that after the age of 60, 8-12% of apparently normal individuals have incidental LB (iLBD) disease not associated with neuronal loss (Forno, 1996; Gibb & Lees, 1988) or PD symptoms.

Notwithstanding inconsistencies, there is overwhelming evidence that extranigral α-synuclein pathology occurs in PD. Furthermore, preclinical/premotor stages often occur, presenting with non-motor symptoms related to non-dopaminergic neuron dysfunction (Ahlskog, 2005) that roughly parallel Braak's staging of PD pathology. Conceding some extent of fallibility in the Braak staging theory, it nevertheless serves as a convenient framework to track the impact of α-synuclein pathology in PD. What follows is a short overview of the first three stages of Braak's PD staging hypothesis (Braak et al., 2003) reviewed in conjunction with a cursory overview of related brainstem and midbrain anatomy function and dysfunction.

Looking at the lower, more inferior systems corresponding to Braak's stage I, the olfactory pathway is an exposed point of entry for environmental pathogens. Indeed the inception sites of PD α-synuclein pathology proposed by Braak et al (Braak et al., 2003) are the nose as well as the gastric system. With respect to the olfactory system, at the level of the optic chasm, the olfactory tract (the olfactory nerve [CN 1], comprised of

efferents from nasal epithelium cells, becomes the olfactory tract at the olfactory bulb) bilaterally forms the lateral and medial olfactory striae (bands of axons); the former finds it way to the olfactory cortex, which includes a few separate structures: entorhinal and piriform cortices and the amygdala. The medial striae course via the anterior commissure to the contralateral olfactory bulb then to the olfactory cortex. The lateral olfactory striae fibers are mainly involved in olfactory transmission, while the medial striae principally mediate the autonomic salivation response. A portion of the medial olfactory striae prompting increased gastric and peristaltic activity, in response to food odors, interacts with the medullary dorsal vagal nucleus (Helwany M & B., 2020).

Damage, as inflicted by environmental insults or ostensibly by α-synuclein pathology (the latter may or may not be related to environmental pathogens), to cells of the olfactory epithelium, olfactory bulb, olfactory tract or striae, or fibers linking the olfactory network of neurons, could result in hyposmia (reduced sense of smell) or anosmia (loss of sense of smell). Subclinical reduction in DAT accompanied by hyposmia can precede clinical PD (Berendse et al., 2001). A large study (2263 participants) reported clinical PD was predated by impaired olfaction by a minimum of 4 years (Ross et al., 2008). Olfactory dysfunction while not exclusive to PD, has differentiated clinical PD across several studies with ≥ 81% sensitivity (Katzenschlager, Zijlmans, Evans, Watt, & Lees, 2004; M. Shah, Muhammed, Findley, & Hawkes, 2008; Suzuki et al., 2011; Wenning et al., 1995). Of note, research also suggests that olfactory cortex is involved in a variety of other processes including emotion and memory(Patel & Pinto, 2014).

Stage I pathology, in addition to the olfactory system, occurs, as already noted, in the brainstem vagus (CN X) nerve dorsal motor nucleus, which is located in the medulla ventral to the 4th ventricle. The vagus nerve, the core of the parasympathetic ("rest and digest") nervous system, has an exceptionally long brainstem-to-gastrointestinal range. It has several functions including governing lung and heart parasympathetic response and digestion. Braak et al. (Braak et al., 2006) suggest a possible inception of pathology leading to α-synucleinopathy that begins in the gastrointestinal area. Here, environmental pathogens crossing the epithelial stomach lining might spread via vagus nerve, in a retrograde manner, to the dorsal motor nucleus of the medulla (Braak et al., 2006). Pathogens, including accumulations of α-synuclein, can gain entry to the brain via the vagus nerve (Braak et al., 2003; Hawkes, Del Tredici, & Braak, 2009; Pan-Montojo et al., 2010; Ulusoy et al., 2013). Further,

remarkably, in a longitudinal study, truncal vagotomy has been demonstrated to reduce risk of recurrent post-procedure PD (Svensson et al., 2015), which lends support to argument that the vagus nerve is an early stage purveyor of PD-related pathology. Moreover, evidence also suggests vagus nerve involvement in mood disorders: vagal nerve stimulation can effectively treat mood disorders (anxiety and depression) (Breit, Kupferberg, Rogler, & Hasler, 2018). Behaviorally consistent with this initial early vagal involvement, constipation (Abbott et al., 2001; Ahlskog, 2005), mood disorder (major affective disorder) [(53)] including depression (Fukunishi, Hosokawa, & Ozaki, 1991), are known be common problems antedating motor symptoms of PD.

In Stage II, Braak et al. (Braak et al., 2003) reported sites affected by LB lesions include, as already stipulated, the raphe nuclei of the medulla and the locus coeruleus in the pons. Reciprocal connectivity between the raphe nuclei and suprachiasmatic nucleus (via the dorsomedial hypothalamic nucleus) alters serotonin levels and modulates circadian rhythm and hence wake and sleep states (Deurveilher & Semba, 2008). Disease spreads superiorly up to the locus coeruleus, a norepinephrine-producing nucleus of cells, in the pons. A component of reticular activation system, this structure is thought to play vital roles in numerous functions including the sleep-wake cycle, arousal, attention and memory, and cognitive control (Benarroch, 2009; Koehler, Baer, & Wagner, 2016; Malenka RC, Nestler EJ, & SE, 2009). The locus coeruleus is normally inactive during rapid eye movement (REM) sleep (Schwartz & Roth, 2008), and skeletal muscle atonia and REM state atonia is mediated by both gamma-amino butyric acid (GABA) and glycine (Brooks & Peever, 2012) inhibitory neurotransmitters. Inline with Braak's second stage induction of disease at the medullary raphe nuclei, and pontine locus coeruleus, is excessive daytime sleepiness predating heightened risk of PD (Abbott et al., 2005) as well as REM sleep behavior disorder (RBD); RBD has predated PD by three years in 52% (13/25) of cases and by 3.7 years in 38% (11/20) of cases (Olson, Boeve, & Silber, 2000; Schenck, Bundlie, & Mahowald, 1996).

Stage III is characterized, in part, by Lewy pathology appearance in the pedunculopontine nucleus (PPN) as well as the basal nucleus of Meynert. The PPN (a component of the reticular activation system) is located in the superior pons. It sends efferents to a variety of brain regions, notably the striatum and subthalamic nucleus. It has extensive interconnections with other basal ganglia nuclei including the internal and external pallidum, SNc and SNr (Martinez-Gonzalez, Bolam, & Mena-Segovia, 2011). The PPN participates in numerous functions: gait and

posture (Karachi et al., 2010), imagined gait and overall motor control (Tattersall et al., 2014; Weinberger et al., 2008), waking as well as REM sleep (Urbano et al., 2014). Given its broad motor involvement, the spreading of α-synucleinopathy to the PPN is consistent with the emergence PD motor symptoms associated with this stage. The basal nucleus of Meynert, inferior to the front of the thalmus and comprised mainly of cholinergic neurons, is implicated in prolonged attention (R. Liu et al., 2018) and memory, specifically long term memory (Ridley, Baker, Leow-Dyke, & Cummings, 2005). It is also implicated in aspects of visual perception(Smythies, 2009). Certainly, however, the main change at stage III, with respect to Braak's staging, is induction of disease at the dopaminergic midbrain SNc, which, as previously discussed, is involved in movement and reward functions.

A hiatus of 4 to 7 years (Gaig & Tolosa, 2009; Hilker et al., 2005) or longer (Savica, Rocca, & Ahlskog, 2010)is thought to occur between the start of SNc α-synucleinopathy associated neuropathology and actual presentation of dysfunctional motor symptoms. PD motor symptoms become manifest subsequent to insidious loss of 60-80% of nigrostriatal dopaminergic neurons (Cummings et al., 2011; Fahn et al., 2004), a pattern also demonstrated in non-human primate research(Brownell et al., 1998; Wullner et al., 1994).

As just reviewed, non-motor clinical symptoms (i.e. olfactory dysfunction, constipation, cognitive and memory dysfunction, depression, daytime sleepiness and RBD) can antedate PD motor symptoms, and appear sequentially, consistent with Braak's staging (Braak et al., 2003): first at initial lower induction sites then spreading upwards, in prion-like fashion, to gradually reach the motor involved SNc, where introduction of the α-synucleinopathy is regarded as central to PD motor dysfunction (Cummings et al., 2011; Fahn et al., 2004). Early stage PD pathology can be revealed not only by non-motor clinical symptoms but also by particular cerebral spinal fluid (CSF) biomarkers.

Paradoxically, while an abnormally high level of α-synuclein in the brain typifies PD pathology, reduced levels of α-synuclein and other biologics occur in the CSF of those with PD pathology. In drug-naïve early stage PD pathology, and relative to controls, CSF α-synuclein, amyloid-beta ($A\beta_{1\text{-}42}$), total tau (tTau), and tau phosphorylated tau_{181} (pTau) have demonstrated significantly lower levels, with reduced α-synuclein inversely related to heightened motor disruption (Kang et al., 2013). These biomarkers though, had mediocre early PD versus control diagnostic utility (area under the curve ≤ .80) (Kang et al., 2013). A follow-up study from the same

camp reported similar findings but also included SWEDD subjects: the SWEDD group had CSF biomarker levels that were intermediate between early PD and control levels (Kang et al., 2016). PD pathology has demonstrated higher levels of CSF A$\beta_{1\text{-}42}$, but lower levels of tTau and pTau relative to Alzheimer's disease (Shi et al., 2011). Moreover, it has been posited that the reduced CSF levels of α-synuclein, A$\beta_{1\text{-}42}$, tTau, and pTau in drug naïve early PD is related to an interaction among pathological levels of α-synuclein, amyloid-beta, tTau and pTau - an interaction that may play a role in PD pathogenesis (Murakami et al., 2019). Speculatively, it may also be that retention of biologics (i.e. α-synuclein, A$\beta_{1\text{-}42}$, tTau and pTau) in the PD brain neurons and microglia reduces dispersion of the biologics to the CSF, a postulate that parallels theoretical reasoning for altered levels of CSF biologics in Alzheimer's (Blennow, Hampel, Weiner, & Zetterberg, 2010; Tapiola et al., 2009). However, as stated at the outset of the current work, the initial cause of cerebral α-synuclein pathological accumulation, or whatever else contributes to the origin of the pathology (e.g. olfactory and/or enteric system pathogens (Braak et al., 2006; Braak et al., 2003)), has yet to be verified. In the absence of a definitive understanding of disease causation, new insights to disease inception may be gained by trying to predict PD from early stage pathology. Non-motor clinical variables (e.g. olfactory function and RBD) and CSF biomarker variables employed in predictive models will likely become increasingly useful in this regard.

Predictive models

Statistical predictive models have been widely used to ease the decision-making of physicians and more generally in medical research to help predict and determine risk or likelihood of pathology(L. X. Chen, 2020). The predictive binary (two-class, response variable) classification models germane to the current Parkinson's research are logistic regression, the general additive model (GAM)(S. N. Wood, 2004; S. N. Wood, 2008; S. N. Wood, 2011), decision tree (L. Breiman, Freidman, J., Olshen, R., & Stone, C., 1984; Therneau, 2018), random forest (L. Breiman, 2001) and XGBoost (T. Chen, Guestrin, C., 2016). These classification models, like most in general, yield both a discrete class prediction (i.e. a category: early PD vs. HC; early PD vs. SWEDD) as well as the predicted probability of class membership: these are probability estimates, ranging between 0 and 1.

Performance comparisons of logistic regression and tree models lend insight to model properties. Logistic regression has outperformed tree-based analysis across 32 smaller data-sets ($N < 1000$ observations)

(Lim, Loh, & Shih, 2000). Logistic regression and decision tree (L. Breiman, Freidman, J., Olshen, R., & Stone, C., 1984) have achieved a stable AUC utilizing far fewer instances per variable relative to random forest (van der Ploeg, Austin, & Steyerberg, 2014). By contrast, and in larger data-sets, random forest, has demonstrated higher performance (Guo et al., 2016). Similar outcomes were found for smaller versus larger data-sets, particularly data with higher levels of variance or noise: logistic regression performs best in smaller data-sets and data-sets distinguished by higher variability while random forest has superior performance in larger data-sets characterized by relatively high signal-to-noise ratio or low variability (Kirasich, 2018; Perlich, Provost, & Simonoff, 2004). Relative to logistic regression, random forest was found to attain a higher true positive rate (or sensitivity) but also a false positive rate that increased as variability in data increased (Kirasich, 2018). Narrowing discussion to data set size and performance, XGBoost (AUC .860) has out performed logistic regression (AUC .728) in a sample of $N = 6682$ (Zhang, Ho, & Hong, 2019), but in a sample of $N = 551$ logistic regression (AUC .873) out performed random forest (ROC AUC .854) and XGBoost (AUC .868) (Xiao et al., 2019). Moreover, while logistic regression and XGBoost have shown comparable performance (Chen, Lin, Yeh, Chai, & Weng, 2019; Hong, Haimovich, & Taylor, 2018) and logistic regression and random forest have modestly out performed XGBoost (Gao et al., 2018), a search of available online research (e.g. PubMed) ranks XGBoost quite consistently as the highest performing model, especially with larger data sets (Hernesniemi et al., 2019; Luo, Li, Liu, & Shen, 2019; Shimoda, Ichikawa, & Oyama, 2018; Tang et al., 2018). Further, without doubt, the XGBoost algorithm (T. Chen, Guestrin, C., 2016) has become the performance front-runner in machine learning challenges (Kaggle, 2018b). Yet, as articulated in a vetted on-line forum (Exchange, 2019, March 1), it is important to stress that despite the excellent track record of XGBoost, it is not guaranteed to always be the best model type in all settings.

Logistic regression classification models have become a mainstay to predict disease risk in clinical research circles, thereby facilitating the administration of appropriate patient care (Shipe, Deppen, Farjah, & Grogan, 2019). Logistic regression is a decades-old algorithm; the GAM and tree-type models have been more recently developed Logistic regression and XGBoost are the two most chronologically distanced models: logistic regression appeared in the mid 1800s (Cramer, 2002); a stable release of XGBoost (Chen, T., Guestrin, C., 2016) appeared in 2017. Many researchers favour a particular model type, yet it has been convincingly argued in the so-

called “no free lunch” theorem that in the absence of substantive insight regarding a particular analysis issue, modeling algorithm (A) will not necessarily perform better than modelling algorithm (B) (Cramer, 2002; Wolpert, 1996), and consequently assessing several different model types can be informative.

Relative to the GAM and tree models, logistic regression is simple to implement. This simplicity is coupled with the quantification of a predictor’s unique contribution to classification outcome. Certainly, its high usage over the past century to the present day underlines the overall across-discipline high regard for the logistic regression algorithm. The general additive model (GAM) is, a much newer, virtual halfway house between the parametric logistic regression model and non-parametric algorithms. The GAM, like logistic regression, is a generalized linear model (GLM) but it is non-parametric; it has the same assumptions as logistic regression (e.g. multicollinearity, independence of errors, absence of highly influential outliers) except, unlike logistic regression, the GAM does not have an assumption of linearity. In logistic regression this refers to the assumption of a linear relationship between continuous predictors and the logit of outcome. A GAM does not quantify predictor contribution as logistic regression does (via coefficients, z-scores and odds ratios). A GAM applies a smoothing function to a predictor typically transforming it to a non-parametric form (Gareth, 2013). But a predictor can also be specified to remain in parametric form.

Classification tree models also do not provide parameter or coefficient estimates quantifying a predictor’s unique contribution to model classification. Tree models are, along with the GAM, are non-parametric. Although the decision tree, GAM (smoothed variables), random forest and XGBoost (xgbtree-type) do not have parameters, they do have hyper-parameters. In general, hyper-parameters are model settings that need to be tuned to optimize model performance. Such tuning can be performed by trial and error or more automatically with software such as the caret package (Kuhn, M., 2013).

Tree models do not have any formal distribution assumptions, which make them particularly useful for data with non-normal patterns (e.g. multimodal and excessive skewness). Decision tree (Breiman, Freidman, J., Olshen, R., & Stone, C., 1984) models offer instant model visualization which is of considerable assistance to interpretation, but the decision tree model can is also prone to overfitting. Random forest (L. Breiman, 2001) uses decision tree ensembles and has an architecture that includes a randomized resampling procedure that mitigates

overfitting. XGBoost has properties that both control overfitting and uniquely advance the search for predictive relationships in a model (Chen, T., Guestrin, C., 2016). For additional information on all model types see Supporting information III (The models: Logistic regression, general additive, decision tree, random forest and XGBoost). It is worth noting two caveats that generally apply to models: models types can be biased by severely imbalance data, and while multicollinearity is not formally an issue with tree models, high predictor correlations can bias predictor or feature selection for all models. As a consequence, a cutoff (e.g. .75 to .85) sensitive to collinearity has been recommended (M Kuhn, 2013). As may be apparent from the preceding synopsis of the logistic regression, GAM, decision tree, random forest and XGBoost, each of these model types have differing properties, properties that collectively provide data analysis insights likely beyond that uncovered by any single model.

Prediction of Parkinson's pathology using non-motor clinical and biomarker variables

At the time of this writing there were only a few studies employing binary (two-class) predictive models aimed at improved early PD pathology detection incorporating non-motor clinical (Nalls et al., 2015) or combined non-motor clinical and biologic variables (Prashanth, Dutta Roy, Mandal, & Ghosh, 2016; Yu, Stewart, Aasly, Shi, & Zhang, 2018). The latter referenced, distinct, three studies all achieved approximately .90 or higher cross-validated AUC scores discriminating early PD from controls; all studies also trained models using Parkinson's Progressive Markers Initiative (PPMI) data (http://www.ppmi-info.org/). The Nalls et al. (Nalls et al., 2015) and Yu et al. (Yu et al., 2018) studies used logistic regression models and non-motor variables; Prashanth et al. (Prashanth et al., 2016) employed several models (logistic regression, Naïve Bays, booster trees, random forest and support vector machines) and used non-motor features as well as SPECT DAT imaging data. The Nalls et al. study included genetic risk for PD, and reported age, gender, family history, olfactory function and genetic risk as the features contributing most to classification accuracy (Nalls et al., 2015). Yu et al. investigated contribution of similar non-motor clinical features (excluding genetic risk) to model classification accuracy but also included CSF biomarkers α-synuclein, pTau, tTau and amyloid-beta predictors. The features reported most important to classification accuracy were age, gender, α-synuclein and olfactory function (Yu et al., 2018). In both the Nalls et al. and Yu et al. studies olfactory function proved to be the single most important predictor to model

classification. In the multiple model Prashanth et al. (Prashanth et al., 2016) research, the most important features to classification were putamen striatal binding ratio (from SPECT DAT imaging) followed by olfactory function.

Uniquely, the current work involved two separate binary classification analyses: discriminating early PD vs. age-matched healthy controls, and discriminating early PD vs. SWEDD. It was hypothesized five distinct model types (logistic regression, the general additive model, decision tree, random forest, and XGBoost) employing combined non-motor and biomarker features would classify early PD from controls with > 80% cross-validated AUC, but that the diverse nature of SWEDD would reduce the early PD versus SWEDD cross-validated classification AUC. This level of classification accuracy appeared inline with other predictive research results (Nalls et al., 2015; Prashanth et al., 2016; Yu et al., 2018) that, in common with the current work, used PPMI data. It was also posited that the rank of predictor importance to classification would vary among model types. Further, some difference in rank of predictor importance to classification between early PD/control and early PD/SWEDD analyses was also expected. Finally, the percentage of class prediction concordance of the two top performing models with longitudinal diagnoses was also determined. Moreover, to potentially assist in research and clinical settings, the R code and data for all models is available from the author (cslfalcon@gmail.com), and to facilitate better model understanding, the fundamentals of model algorithms are outlined (see Supplementary information III).

Chapter Two: Methods

2.0 Procedures

As already stipulated, classification performance was compared for logistic regression, general additive (GAM) (Wood, S., N, 2004; Wood, S., N, 2008; Wood, S. N, 2011), decision tree (Breiman, L., Freidman, J., Olshen, R., & Stone, C., 1984; Therneau, 2018), random forest (Breiman, L, 2001) and XGBoost (Chen, T., Guestrin, C., 2016) models in two separate analyses: early PD versus control and early PD versus SWEDD (scan without evidence of dopaminergic deficit). This amounted to building ten (5 x 2) classifiers. The AUC was the main performance metric. Sensitivity, specificity, general accuracy and the Kappa static were also determined. The general sequence of data analysis steps is depicted in Fig. 1. Also the two highest performing classifiers from the early PD versus control classification analyses were applied to SWEDD test data to assess conversion of SWEDD to PD. The case-wise percentage of model predicted SWEDD to PD conversion that conformed to (available) longitudinal PPMI curated 12-36 month diagnosis was then assessed. Further, the case-wise percentage of early PD versus SWEDD model sensitivity and specificity that conformed to (available) PPMI curated longitudinal 12-36 month diagnoses was also determined for the two highest performing early PD versus SWEDD classifiers. Longitudinal 12-36 month DAT scan mean putamen values provided an imaging measure of disease, but as noted in the preceding paragraph, model predictors consisted of only non-motor clinical and biomarker variables; imaging was not included among model predictors.

After screening, SWEDD minority class rate became 13% (43/338), and the random stratified training/validation data split further reduced the SWEDD training cohort to just 22 cases (and 148 early PD). To improve data symmetry early PD/SWEDD models were trained on SMOTE (synthetic minority oversampling technique) subsampled data.

It is underlined that to prevent leakage of test data information into training data, models were trained and features selected only from the training data. This mitigated overly optimistic model performance estimates on the test data. To ensure reproducibility, one specific seed value was set prior to partitioning of data and model execution. All data used can be obtained from https://github.com or the corresponding author.

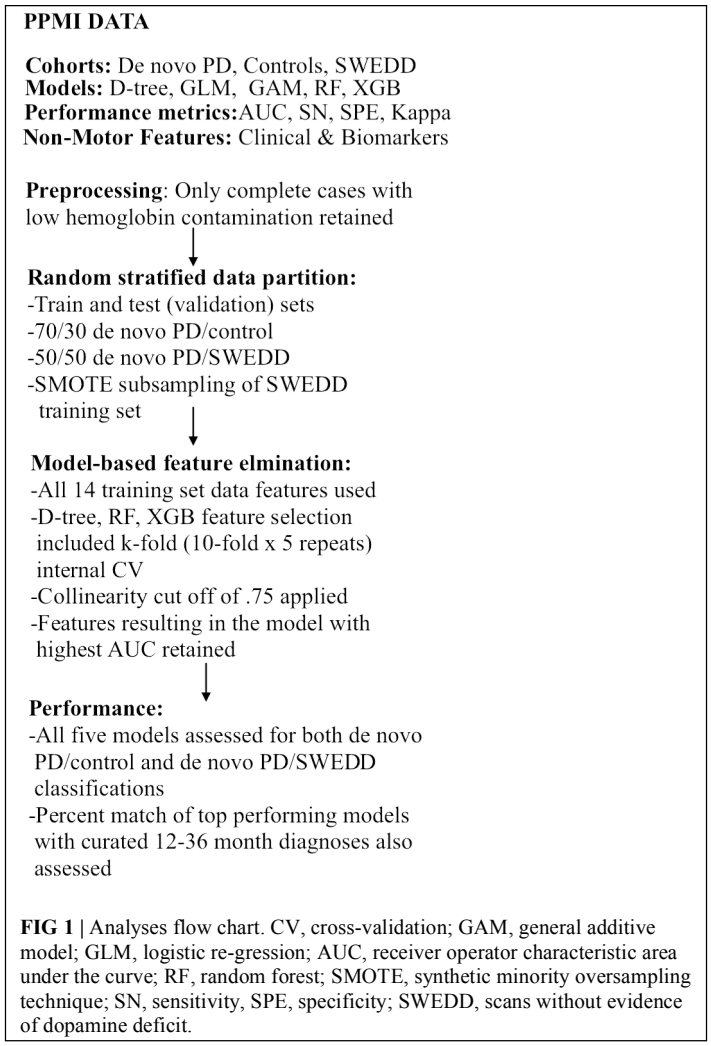


FIG 1 | Analyses flow chart. CV, cross-validation; GAM, general additive model; GLM, logistic re-gression; AUC, receiver operator characteristic area under the curve; RF, random forest; SMOTE, synthetic minority oversampling technique; SN, sensitivity, SPE, specificity; SWEDD, scans without evidence of dopamine deficit.

2.*1 Participant data*

Data used in the preparation of this article was obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. The PPMI is a landmark, multicenter, longitudinal research effort mandated to identify PD markers (4). It is a public-private partnership funded by the Michael J. Fox Foundation for Parkinson's Research and the funding partners, include AbbVie, Allergan, Amathus Therapeutics, Avid Radiopharmaceudicals, Biogen, BioLengend, Bristol-Myers Squibb, Celgene, Denali Therapeutics, GE Healthcare, Genentech, GSK, Lilly, Lundbeck, Merck, MesoScale Discovery, Pfizer, Piramal Imaging, Prevail Therapeutics, Roche, Sanofi Genzyme, Servier, Takeda, Teva, UBC, Verily, and Voyager.

Subject data was anonymized while also allowing individual subjects to be tracked across different studies. Acquired subject data (downloaded July 31, 2019) included three cohorts: 423 early (de novo) PD, 196 healthy controls (controls or HC), and 64 scans without evidence of dopamine deficiency (SWEDD). With respect to the models developed, only baseline data was used, and such baseline data was acquired within 2 years of PPMI project enrolment. It warrants mention that the PPMI early PD baseline cohort is drug naïve but symptomatic.

After screening for complete cases (for details see *2.4 Screening*), and then subsequent to random stratified partitioning of data in to train and validation (test) sets, data instances were as follows: for early PD/control classification, there were 207 early PD (135 male) versus 91 controls (60 male), the validation set was 88 early PD (57 male) and 39 controls (21 male); the two top early PD/control models data were also tested on all 43 SWEDD (25 male) and the 39 controls from the early PD/control test set.

For the early PD/SWEDD classification, a synthetic minority oversampling technique (SMOTE)-based training set was derived from a random sampled partition of 148 early PD (95 male) and 22 SWEDD (16 male). The SMOTE training data set consisted of a balanced 44 early PD (30 male) and 44 SWEDD (30 male), and it was used for all early PD/SWEDD models except the decision tree model. The early PD/SWEDD decision tree model used SMOTE data obtained during resampling (see 5.1 *Decision tree early PD versus SWEDD classification results (SMOTE-based model),* which resulted in 88 early PD (63 male) and 66 SWEDD (41 male). The validation or test set used for early PD/SWEDD (not altered or subsampled) consisted of 147 early PD (92 male) and 21 SWEDD (9 male). The PPMI curated data was used to assess extent of conformity of model predicted classification to diagnosis at 12-36 months post baseline.

2.2 Feature elimination and hyper-parameter tuning

Final features (predictors), selected only from training set data, were determined by model-based feature elimination coupled with the ROC AUC: models with features resulting in highest model AUC constituted the final models applied to the test data sets. For tree-based models, caret package (M. Kuhn, 2019, March, 3) internal cross-validation (10-fold, 5 repeats) resampling was used to tune hyper-parameters and arrive at the optimal feature set (see *Modeling and the caret package* Supporting information III). Stepwise regression employing the

Akaike information criterion (AIC)(Akaike, 1974) was used for logistic regression feature elimination. GAM models used the same features as logistic regression. In addition, as with the tree-based algorithms, final logistic regression model selection was narrowed to the model with highest AUC. For the logistic regression GAM, the intent was to use the GAM to supplement and corroborate GLM results but also as a distinct classification model. The logistic regression GAM was executed using the same stepwise regression features selected for logistic regression. The built-in measures of predictor importance, reviewed in Supporting information III, are as follows: *rpart* decision tree goodness of split; random forest mean decrease in Gini impurity; XBoost Gain index. Variable of importance criteria for logistic regression included predictor coefficient z-scores and p-values, coefficient odds ratios, as well as predictor impact on the model deviance statistic, and the Akaike information criterion. The importance of GAM features was assessed using the chi-squared statistic and associated p-values.

2.3 Clinical assessments and cerebral spinal fluid assays

Features, 14 in all, from prior research (see introduction) demonstrating promise discriminating early PD were assessed. Biological predictors, the biomarkers, were cerebral spinal fluid (CSF) levels of amyloid-beta $_{1\text{-}42}$ ($A\beta_{1\text{-}42}$), α-synuclein, tau phosphorylated at threonine 181 (pTau), and total tau (tTau). With respect to biomarkers, because hemoglobin contamination can influence the biologic measures, exclusion of samples with > 200 ng/ml has been recommended (Mollenhauer et al., 2011); this screening recommendation was adopted in the current work. The non-motor clinical measures were trait anxiety scores based on the 20-item trait anxiety (max 80 points or 4 points per question) scale from the State-Trait Anxiety Inventory (Spielberger, Gorsuch, & Lushene, 1970); depression score based on the 15-item (max 15) Geriatric Depression Scale (GDS) (Yesavage et al., 1983); cognition based on the 30 point Montreal Cognitive Assessment Inventory (MoCA) (Nasreddine et al., 2005), which has a maximum score of 30; olfactory acuity based on the 40-item (maximum score of 40) University of Pennsylvania Smell Identification Test (UPSIT)(Doty, Deems, & Stellar, 1988; Doty, Shaman, & Dann, 1984), which was reverse scaled (higher is proportional to lesser olfactory acuity); the 8-item (max score of 24) Epworth Sleepiness Scale (ESS) (Johns, 1991); the 10-item (max score of 13) rapid eye movement sleep behavior disorder questionnaire (RBDQ) (Stiasny-Kolster et al., 2007); and constipation, ranging from 0 to 4 (MDS-UPDRS-I, (Goetz et al., 2008). Sociodemographic measures also included under the rubric of non-motor clinical measures

were age, gender and years of education. The biomarkers are continuous variables, and the clinical variables are continuous or semi-continuous scales. Note, for all clinical measures except MoCA, and given UPSIT was reverse scaled, higher scores are generally suggestive of pathology while the reverse typically holds for the biologics; higher CSF biologic values suggest a more normal state.

2.4. Screening

The main screening criteria were complete records across all variables for a given subject's data as well as low hemoglobin blood contamination. Specifically, complete subject records for clinical non-motor and biomarker variables (itemized in 2.2 above) as well as complete imaging records (caudate and putamen SPECT DAT values) and complete records for the MDS-UPDRS III, (Goetz et al., 2008) scale. Data was also screened for low blood contamination as indicated by the CSF hemoglobin of < 200 ng/mL) (Mollenhauer et al., 2011). Strict adherence to the blood contamination criterion eliminated 131 cases, reducing the data-set to 151 controls, 328 early PD, and 47 SWEDD. Control group case number was further reduced by two missing UPDRS III scores (n = 149), 3 missing anxiety trait scores (n =146), two missing (n =144) MoCA scores, and 14 missing (n =130) striatal DAT values. For early PD, case number was further reduced by 15 cases (n =313) of incomplete RDBQ scores, two incomplete instances of (n = 311) GDS, one incomplete (n = 310) ESS score, three incomplete (n = 307) MoCA records, and 12 (n = 295) incomplete dopamine transporter (DAT) records. For SWEDD, there was one ESS (n = 46) missing record, one missing MoCA (n= 45) record and two missing olfaction (n = 43) records. Subsequent to this screening the final number of participants was 468 (130 controls; 295 early PD; 43 SWEDD).

2.5 Imaging

Because dopamine active transporter (DAT) values and clinical motor (MDS-UPDRS III) status measures are virtually ever-present in PD assessments, they were included as background indices to help quantify extent of pathology. Single Photon Emission Computed Tomography (SPECT) DAT values (i.e. striatal binding ratio) data was used as the striatal (dopamine) measure of neurodegenerative status. A complete technical specification and operations SPEC manual is provided by PPMI and is available at http://www.ppmi-info.org/wp-content/uploads/2017/06/PPMI-TOM-V8_09-March-2017.pdf.

2.6 Statistical analyses

The type I error rate was set at .05 ($\alpha = .05$). Statistical analysis was conducted in *R* (Team, 2018). The univariate distribution of all variables was initially examined for indications of relative data normality using descriptive statistics, density plots and numeric (Shapiro & Wilk, 1965) analyses. Gender proportion within groups was assessed with binomial tests; two-sample tests for equality of gender proportion were used to assess gender proportion between groups. Boxplots were used to show the range, or spread of variable data values for early PD, control and SWEDD groups. Bivariate variable relationships were assessed with correlation tests and scatterplots. The SPECT DAT values were included in these bivariate assessments to help link the broadly acknowledged disease indicator SPEC DAT with the non-motor clinical and biomarker predictors. Imaging values then, provided an indication of disease-relation to predictors (but, as noted already, imaging was not included in the classification analyses). Because the data was generally non-normally distributed, robust t-tests (Wilcox, 1990) were used to compare variables between groups. Models initially included (controlled for) age, education and gender. Model residuals reflected the non-linear patterns in the data (see Supporting information I: Figs S1-2 and S1-7).

Collinearity can make logistic regression coefficients unstable, less precise (Bagley, White, & Golomb, 2001; Field, 2012). It can result in GAM concurvity (a form of co-linearity where one smooth term approximates another) (S. N. Wood, 2019b). For tree models, however, concern for collinearity of variables is controversial (Classification, 2019). But considering random forest, for example, one of two or more correlated features can be randomly selected without preference; impurity removed by the selected feature potentially masks additional impurity that could have been removed by a correlated feature(s) (Exchange, 2019c). Indeed, with correlated features, less relevant features can take the place of more important features (Carolin Strobl, Boulesteix, Kneib, Augustin, & Zeileis, 2008) and feature ranking can be inaccurate (Tolosi & Lengauer, 2011). Because collinearity is certainly problematic for logistic regression, can potentially bias feature selection, ranking and hence classification of GAM and tree models, the current work adopted a multicollinearity of cut off of $r_s = .75$, which is relatively sensitive to pairwise correlations (M Kuhn, 2013). To prioritize unbiased feature selection and classification for all models, features exceeding the cut off were not combined within the same model. Note however, that for all models, all 14 non-motor clinical and biomarker features were included in initial model-

based feature elimination. The collinearity cut off was only applied to the final model feature set to increase reliability of feature importance ranking and classification.

The lone parametric model was logistic regression. To maintain a relaxed but reasonable observation to predictor ratio (no fewer than 5 cases per predictor) (Vittinghoff & McCulloch, 2007), no more than 7 predictors were entered in to any single regression model, which complied with the recommended observation to predictor ratio. Stepwise regression (backward) was used employing AIC (Venables, 2002). The Coefficients ($\widehat{\boldsymbol{\beta}}$), standard errors ($\boldsymbol{SE}\,\widehat{\widehat{\boldsymbol{\beta}}}$), z-values, and odds ratios were calculated. In addition, standardized versions of the predictor coefficients as well as rescaled versions of coefficients (i.e. with score range binarized) were also provided to lend an unbiased perspective to differentially scaled predicators (Cohen, Cohen, West, & Aiken, 2003). In general, predictors with narrower confidence intervals were regarded as more precise and stable (Poole, 2001). Measures of model fit to data included deviance, AIC (Akaike, 1974), and McFadden pseudo R^2 (McFadden, 1974). An additional indication of model fit used was the Hosmer and Lemeshow Goodness of Fit test (Hosmer, 2013). Assumption diagnostic tests used were the Durbin-Watson (1952) to assess independence of errors (or autocorrelation), multicollinearity to assess extent of predictor correlation using a version of the variable inflation factor (VIF) test adapted for logistic regression (Fox, 2011), and the Box-Tidwell (Box & Tidwell, 1962) test for linearity of the logit. Logistic regression is also impacted by outliers, which may have properties including excessive leverage, discrepancy or influence (for details see Logistic regression under *The models: Logistic regression, decision tree, random forest and XGBoost* in Supporting information III). Case-wise analyses and plots were used to assess the latter three measures. To measure how far case *i* observation was from the mean of the predictors, a cut-off provided by Belsely (Belsely, Kuh, & Welsch, 1980) was used based on $3M_h = 3(k + 1)/n$, where M_h is mean leverage for the predictor, *k* is the number of predictors and *n* is the number of cases. For discrepancy (based on external studentized residuals), the typically adopted cut-off value of ± 2 was used. Cook's distance (which combines leverage and discrepancy) was used to calculate influence, and the widely used cut-off of 1 was used.

As already noted, the logistic general additive model (GAM) had a twofold purpose: as a distinct classification model and to corroborate the logistic regression model. With respect to the latter, data characteristics and model assumptions noticeably impacted the logistic regression GLM: two semi-continuous variables (MoCA in the early PD/control analysis and years of education in the early PD/ SWEDD analysis) violated the linearity of the logit assumption. Transforms were attempted (e.g. the square root, log, cube root) with only minor improvement and such transforms incur the added interpretation complication. Consequently, the offending two variables were simply converted from continuous (numeric) to categorical variables (quartiles), and the categorical variables were used in the logistic regression GLMs. But while such categorization (here conversion of semi-continuous scales to quartiles) provides a remedy for variable non-linearity, information is likely lost; information that is retained when a GAM smoothing function is applied to the same continuous or semi-continuous variable (for an overview of Logistic regression and GAM models see *The models: logistic regression, decision tree, random forest and XGBoost* in Supporting information III). Accordingly, logistic regression GAMs in addition to being distinct classification models, also added perspective to logistic GLM output.

The GAM model thin plate smoother (tp) function (the default smooth function in the mgcv package (S. N. Wood, 2019, March 21) was the basis used. The restricted maximum likelihood (REML) function was chosen as the smoothness selection method governing the extent of wiggle in the wiggly parts of the tp basis function. The REML smoothing parameter estimator treated the basis tp as a random effects term. The REML method was used because it effectively penalizes overfitting (Wood, S., N, 2011). The degrees of freedom associated with a smoothed predictor, initially set by REML, were checked by ensuring the effective degrees of freedom (edf) of a given smoothed predictor was less than *k* (the upper limit on the degrees of freedom). In addition, a diagnostic qqplot specifically configured for the mgcv gam (qq.gam) was used to assess actual vs. theoretical quantiles. The GAM model output was not included in the Results section but was included, along with a diagnostic qqplot and scatterplots, in Supporting information II. Deviance, pseudo R^2 (McFadden, 1974) and explained deviance (which is GAM model pseudo R^2) for the logistic GLM and GAM models were compared. Pseudo R^2 is simply 1 – (model deviance / model null deviance).

2.7 Classification performance metrics

Early PD was the predicted class in the early PD/control classification; SWEDD was the predicted class in the early PD/SWEDD classification. The AUC, rather than simple misclassification error, was the central performance metric. The AUC was also employed in feature elimination and to select tree-model optimal hyper-parameters settings using the caret package [(3)]. With application of a predict function, all models provide discrete class prediction and class membership probabilities (ranging between 0 and 1). Confusion matrices, which summarize prediction results, are used extensively. The confusion matrices provide the predicted classes in 2 x 2 cross-tabulation format of observed and predicted outcomes. Model performance was based on model and, importantly, cross-validated model performance on test sets. Model classification results are provided in the form of confusion matrices and performance metrics: AUC, sensitivity, specificity, accuracy and Kappa. The confusion matrices and the performance metrics results are provided in Chapters 4 and 5. A summary of all model performance metric results is provided in Chapter Six, table 22.

At the default .50 cut-off classification threshold predictive classification probabilities > .50 are categorized as positive events: early PD rather than control; SWEDD rather than early PD. However, the default .50 cutoff often provides a less than ideal balance of confusion matrix performance metric values (see *Predictive classification model evaluation metrics* in Supporting information III). Therefore, for each model, sensitivity, specificity, Kappa and accuracy metrics were reported for two thresholds: the default.50 classification threshold and an optimized classification threshold. The optimized model threshold was selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a). There was one exception where the optimal threshold point of balanced of sensitivity and specificity was point closest to the AUC top left. Moreover, as already noted under *2.0 Procedures*, the AUC was used as the central performance metric. The AUC was chosen because it is a widely used measure and, unlike the accuracy metric, it is unaffected by class imbalances (Fawcett, 2006) between early PD and controls, and the more severe class imbalance between SWEDD and early PD (though SMOTE subsampling compensated for SWEDD early PD instance imbalance). Details of ROC AUC properties are provided under *Predictive classification model evaluation metrics* in Supporting information III. The ROC AUC non-parametric method (E. R. Delong, Delong,

& Clarkepearson, 1988; Sun & Xu, 2014) was used as implemented in pROC (Robin et al., 2011) because it has relaxed normality assumptions. In addition to ROC AUC values and graphs for each model, a roc test for correlated (referring to the same response variable used by different models) ROC curves (Robin et al., 2011) was used to determine if the two highest performing models from each of the classification analyses significantly differed. Bonferroni family-wise error correction was used for ROC AUC comparison among more than two models.

It warrants note, that while a AUC of ~ 70% has been reported as analogous to Cohen's d of .80 (a large effect) (Rice & Harris, 2005) and hence indicative of good classification performance, context should determine the AUC that constitutes a high level of classification performance. In general, in the current work dealing with patient data, AUC values ≥ 80% but < 90% were regarded as indicative of good classification performance; values ≥ 90% are regarded as an excellent level of classification performance. As mentioned under *2.2,* tree-based model hyper-parameters were tuned using 10-fold (5 repeats) cross-validation resampling (M. Kuhn, 2019, March, 3). However, all tree-based models were initially fit using the default settings. An outline of model algorithms, properties, and settings, is provided in Supporting information III. The actual R code and data is readily available from the author (cslfalcon@gmail.com).

Chapter Three: Results

The descriptive statistics section below is followed by model classification results in Chapters 4 and 5. As specified in the methods section, only cases satisfying data requirements (i.e., low blood contamination with complete clinical and biologic data for all predictors) were retained. Also reiterating as outlined in Methods, subsampling, not used in the early PD/control classification, was employed in the early PD/SWEDD classification in aid of addressing the SWEDD minority class rate of 13% (training set: 22/148). A summary of model performance is provided in Chapter Six.

3.1 Descriptive statistics and t-tests

Fig 2 density plots convey a lack of symmetry typical of the predictor variables: 3 of the clinical and 3 of the biological variables are shown.

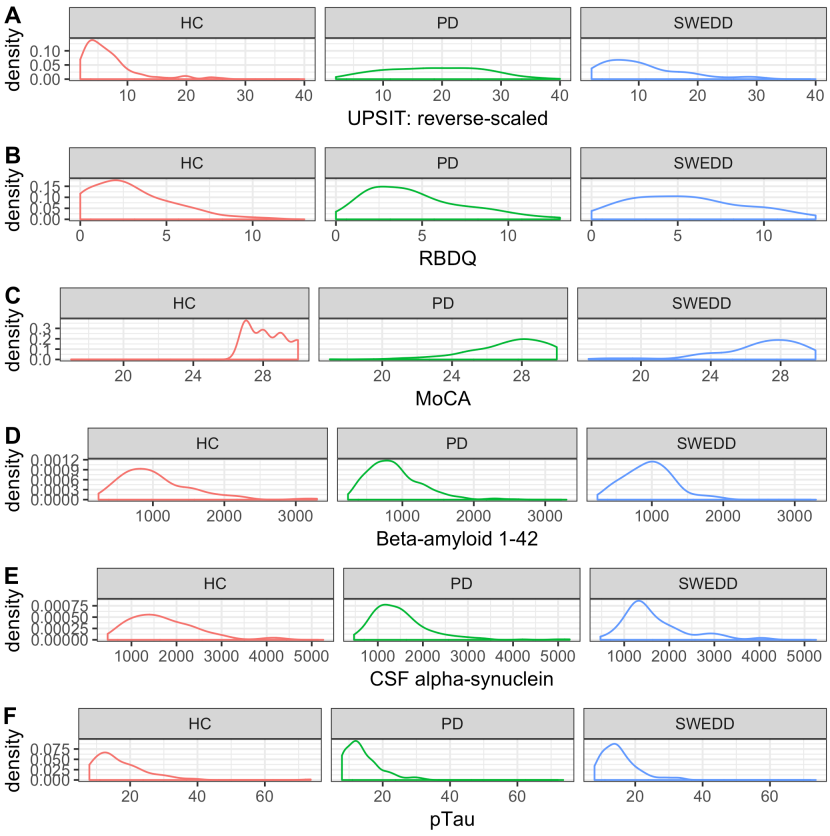


FIG 2 | *Density plots.* *UPSIT = University of Pennsylvania Smell Inventory Test, reverse-scaled; RBSQ = rapid eye- movement Behaviour disorder questionnaire; MoCA = Montreal Cognitive Assessment; HC = healthy controls; PD = early Parkinson's disease; SWEDD= scans without Evidence of dopamine deficit*

Tables 1, 2 provide variable descriptive statistics and pairwise tests (Wilcox, 1990) comparisons that further characterize all data. **Table 1** pertains to early PD versus control groups; **Table 2** pertains to early PD versus SWEDD. Nine of eleven (82%) clinical and biologics in **Table 1** significantly differed between early PD and control groups. By contrast, only 3/11 (27%) clinical and biologics significantly differed between early PD and SWEDD groups in **Table 2**, two of which, Epworth sleepiness scale (79) (ESS) and years of education, did not significantly differ between early PD and control groups (**Table 1**). The clinical variable University of Pennsylvania Smell Test (Doty et al., 1988; Doty et al., 1984), reverse-scaled in the current work (Upsit-rev), was significantly higher (higher reverse-scaled UPSIT is indicative of greater olfactory loss, more severe hyposmia) in early PD compared to controls as well as in early PD compared to SWEDD. The reverse-scaled UPSIT values are otherwise identical to standard (non-reverse scaled) UPSIT values. Gender representation for all data, not provided in **Tables 1 or 2**, was as follows: female controls, $n = 49$; female early PD, $n = 103$; male controls, $n = 81$; male early PD, $n = 192$; female SWEDD, $n = 18$; male SWEDD $n = 25$. A binomial test for controls revealed a proportion of .27 females, which significantly differed from the expected proportion of 50%, $p < .001$. Similarly, a binomial test for the early PD group indicated a proportion of .26 females, which significantly differed from the expected 50% (Dluzen & McDermott, 2000; Van Den Eeden et al., 2003), $p < .001$. Finally, a binomial test for the SWEDD group revealed a proportion of .30 females, which significantly differed from the expected 50%, $p < .01$. With respect to the proportion of male and female cases between groups there was not a significant difference between early PD and control groups, $\chi^2(1) = .303$, $p = .582$. Similarly, there was not a significant difference in gender proportion between early PD and SWEDD groups, $\chi^2(1) = .788$, $p = .375$.

TABLE 1 | Descriptive statistics and t-tests, early PD/controls

***N*= 425**	Early PD, *n*= 295: 192 male, 103 female						HC, *n*= 130: 81 male, 49 female						
	M	***sd***	***Mdn***	***min***	***max***	***skew***	***M***	***sd***	***Mdn***	***min***	***max***	***skew***	***t-test*[1]**
Clinical													
Anxiety	33	10	31	20	63	1	29	7	27	20	53	1	$p < .001$
CNST	0	1	0	0	3	2	0	0	0	0	3	4	$p < .001$
ESS	6	3	6	0	17	1	6	3	5	0	15	1	$p = .269$
GDS	5	1	5	1	11	1	5	1	5	1	15	3	$p = .270$
RBDQ	5	3	4	0	13	1	3	2	2	0	11	1	$p < .001$
Upsit-rev	22	8	22	5	43	0	7	5	6	2	26	2	$p < .001$
Biologics													
AB 1-42	885	379	835	239	2572	1	1043	526	941	239	3297	1	$p = .017$
CSF a-syn	1488	662	1374	472	5257	2	1698	756	1581	601	4271	1	$p = .004$

pTau	14	5	13	8	33	1	17	9	15	8	74	3	$p = .001$
tTau	164	56	154	79	345	1	192	81	170	79	581	1	$p = .004$
SocioDem													
Age	61	10	62	34	85	0	61	12	62	31	84	-1	$p = .868$
Yrs ed.	16	3	16	5	26	0	16	3	16	8	24	0	$p = .167$
MoCA	27	2	28	17	30	-1	28	1	28	27	30	0	$p = .007$
DAT values													
CaudL	2	1	2	0	4	0	3	1	3	1	5	0	$p < .001$
CaudR	2	1	2	0	4	0	3	1	3	1	5	0	$p < .001$
PutL	1	0	1	0	2	1	2	1	2	1	4	0	$p < .001$
PutR	1	0	1	0	3	1	2	1	2	1	4	0	$p < .001$
ave. Caud	2	1	2	0	4	0	3	1	3	1	5	0	$p < .001$
ave. Put	1	0	1	0	2	1	2	1	2	1	4	0	$p < .001$
Motor													
UPDRS III	22	10	21	5	62	1	1	2	0	0	10	2	$p < .001$

Anxiety, STAI trait subscale; CNST, MDS-UPDRS I NP1CNST: 0, none, 1, slight, 2, mild, 3, moderate, 4, severe; CaudL, left caudate; CaudR, right caudate; DAT, dopamine transporter; PutL, left putamen; PutR, right putamen; av. Caud = (left + right caudate)/ 2; av. Put, (left + right putamen)/2; UPDRS III, MDS-UPDRS III total; Upsit-rev (olfactory loss) = University of Pennsylvania Smell Identi-fication Test (a reverse scaled version); AB 1−42, beta-amyloid 1−42; CSF a-syn, cerebral spinal fluid a-synuclein; pTau, CSF phosphorylated Tau; tTau, CSF total tau; MoCA, Montreal Cognitive Assessment; Yrs. ed., years of education; PD, early Parkinson's patient data; HC, healthy control data; SocioDem, socio-demographic; a robust t-test based on Wilcox, 2005.

TABLE 2 | Descriptive statistics and t-tests, early PD/SWEDD

N= 338	Early PD, n= 295: 192 male, 103 female						SWEDD, n= 43: 25 male, 18 female						
	M	***sd***	***Mdn***	***min***	***max***	***skew***	***M***	***sd***	***Mdn***	***min***	***max***	***skew***	***t-test[1]***
Clinical (clinical)													
Anxiety	33	10	31	20	63	1	36	10	32	22	59	1	$p = .237$
CNST	0	1	0	0	3	2	0	1	0	0	4	2	$p = .956$
ESS	6	3	6	0	17	1	8	5	8	0	19	0	$p = .032$
GDS	5	1	5	1	11	1	6	2	5	2	11	1	$p = .299$
RBDQ	5	3	4	0	13	1	6	3	5	0	13	0	$p = .107$
Upsit-rev	22	8	22	5	43	0	10	7	9	2	29	1	$p < .001$
Biologics													
AB 1-42	885	379	835	239	2572	1	960	338	982	374	1897	0	$p = .044$
CSF a-syn	1488	662	1374	472	5257	2	1654	716	1370	488	4041	1	$p = .188$
pTau	14	5	13	8	33	1	15	5	14	8	33	1	$p = .157$
tTau	164	56	154	79	345	1	177	56	170	79	344	1	$p = .112$
SocioDem													
Age	61	10	62	34	85	0	60	10	62	39	79	0	p = .573
Yrs ed.	16	3	16	5	26	0	15	4	15	8	24	0	p = .034
MoCA	27	2	28	17	30	-1	27	3	27	18	30	-1	p = .563
DAT values													
CaudL	2	1	2	0	4	0	3	1	3	1	4	0	$p < .001$
CaudR	2	1	2	0	4	0	3	1	3	1	4	0	$p < .001$
PutL	1	0	1	0	2	1	2	1	2	1	3	0	$p < .001$
PutR	1	0	1	0	3	1	2	1	2	1	3	0	$p < .001$
ave. Caud	2	1	2	0	4	0	3	1	3	1	4	0	$p < .001$
ave. Put	1	0	1	0	2	1	2	1	2	1	3	0	$p < .001$
Motor													
UPDRS III	22	10	21	5	62	1	18	11	17	5	45	1	$p = .007$

Anxiety, trait subscale from the State-Trait Anxiety Inventory; CNST, constipation based on MDS-UPDRS I;CaudL, left caudate; CaudR, right caudate; PutL, left putamen; DAT, dopamine transporter; PutR, right putamen; av. Caud, (left + right caudate) / 2; av. Put, (left + right putamen)/2; UPDRS III, MDS-UPDRS III total; Upsit-rev (olfactory loss), University of Pennsylvania Smell Identification Test (a reverse scaled version); AB 1-42, beta-amyloid1-421-421-42 1-42; CSF a-syn, cerebral spinal fluid a-synuclein; pTau, CSF phosphorylated Tau; tTau, CSFtotal tau; MoCA, Montreal Cognitive Assessment; Yrs. ed., years of education; PD, early Parkinson's patient data; SWEDD, scans without evidence of dopaminergic deficit; SocioDem,

socio- demographic; a robust t-test based on Wilcox, 2005.

Boxplots in Fig 3 visually encapsulate properties (e.g. dispersion) of a few clinical predictors across groups: (A) is the University of Pennsylvania Smell Test (Doty et al., 1988; Doty et al., 1984) (UPSIT) that, in the current work was reverse-scaled (higher is indicative of greater olfactory loss, more severe hyposmia); (B) Anxiety refers to anxiety traits from the State-Trait Anxiety Inventory (Spielberger et al., 1970); (C) is rapid eye movement behaviour disorder questionnaire (Stiasny-Kolster et al., 2007) (RBDQ); (D) is the Epworth Sleepiness Scale (Johns, 1991) (ESS).

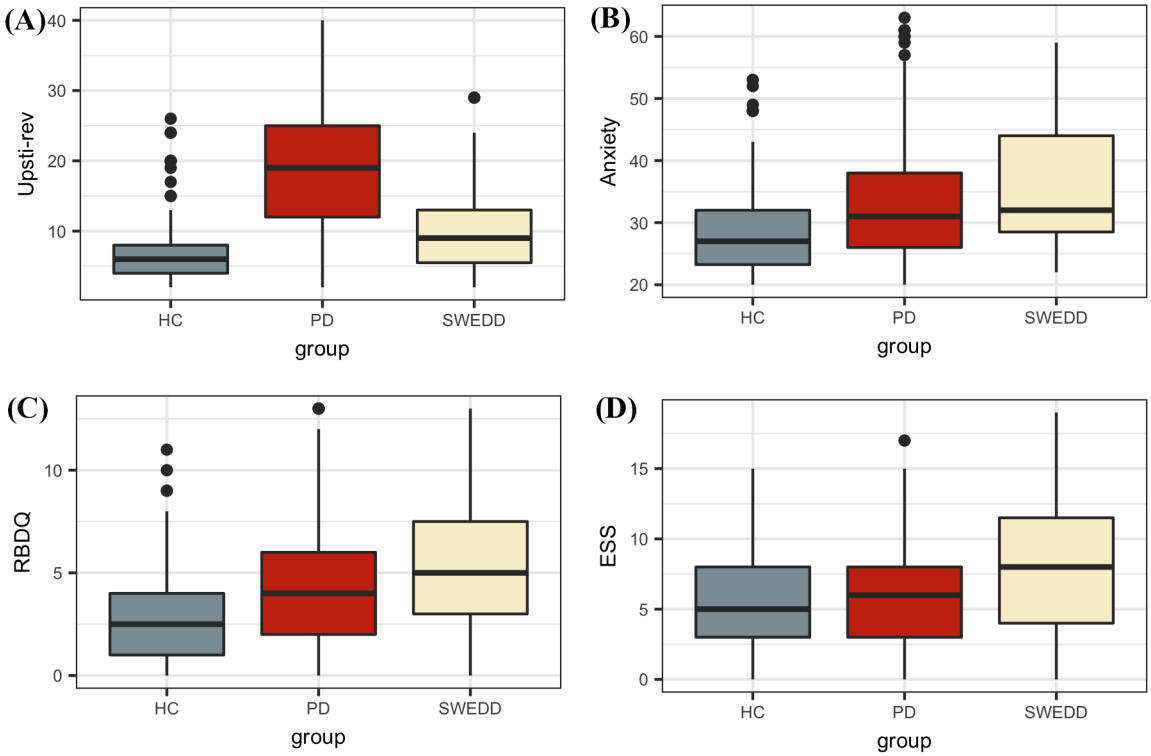


FIG 3 | *Predictor boxplots*: *(A) Upsit-rev = University of Pennsylvania Smell Inventory Test score reverse-scaled; (B) Anxiety = trait subscale from the State-Trait Anxiety Inventory; (C) RBDQ = rapid eye movement behaviour disorder questionnaire; (D) ESS = Epworth Sleepiness Scale; HC = healthy controls; PD = early PD; SWEDD = scan without evidence of dopaminergic deficit.*

Fig 4 boxplots characterized CSF biologic variables across groups: (A) is beta amyloid 1-42 ($A\beta_{1\text{-}42}$); (B) is CSF alpha-synuclein; (C) is pTau; and (D) is tTau.

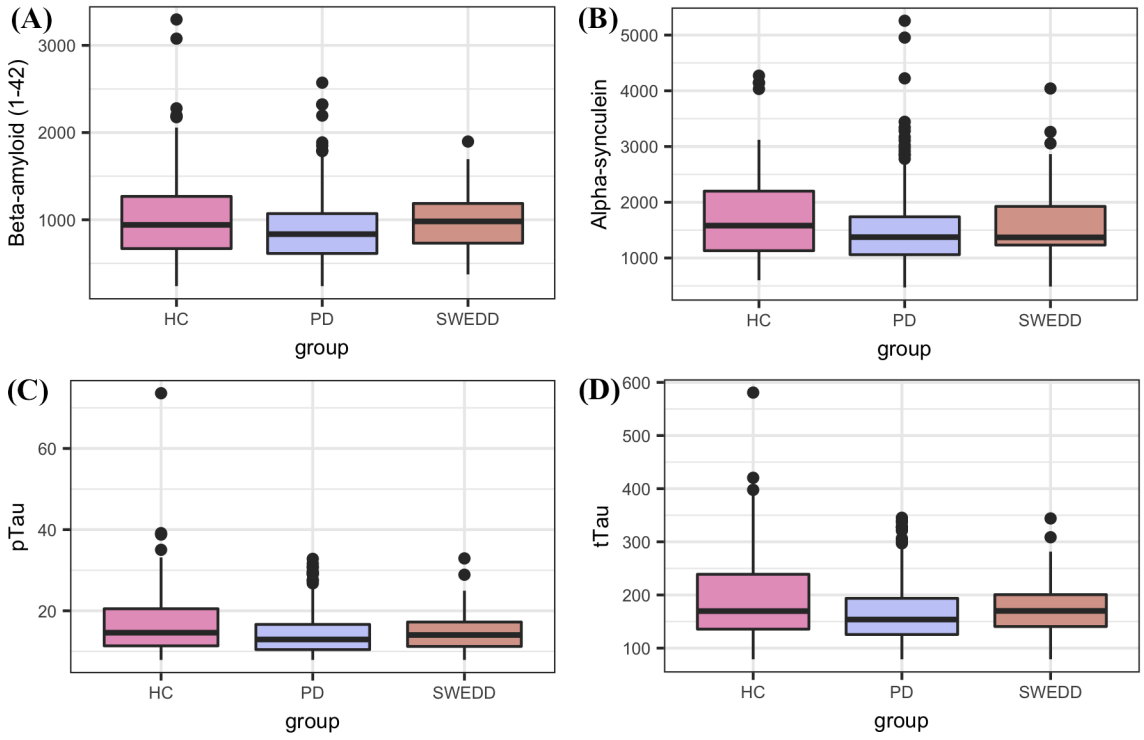


FIG 4 | *Biological predictor boxplots*: *(A) beta amyloid Aβ $_{1\text{-}42}$; (B) CSF alpha-synuclein; (C) pTau; and (D) tTau. HC = healthy controls; PD = early PD; SWEDD = scan without evidence of dopaminergic deficit. HC = healthy controls; PD = early PD; SWEDD = scan without evidence of dopaminergic deficit.*

3.2 Bivariate analyses

Because of the largely non-normal distribution of variables, non-parametric Spearman correlations were used rather than Pearson *r*. Again, as noted previously, imaging values were not used as model predictors in the classifier analyses following this section. However, their relation to non-motor predictors and biomarkers depicted in Fig 5 provides an indication of how neuropathology, as indexed by SPEC DAT mean caudate and mean putamen (also referred to striatal binding ratios) values, relates to the non-motor predictors and biomarkers. Fig 5 depicts variable correlations for all data ignoring groups. Circle size in Fig 5 is proportional to the Spearman correlation: larger circles reflect stronger correlations; correlations are color-coded, red indicating a negative correlation and blue indicating a positive correlation. For example, a strong negative association between hyposmia (reverse-scaled UPSIT, here named Upsit-rev) and DAT scan putamen values is evident; a strong negative association between hyposmia and DAT scan caudate values is also evident. Additionally, strong positive correlations exist among A$\beta_{1\text{-}42}$, α-synuclein, pTau and tTau.

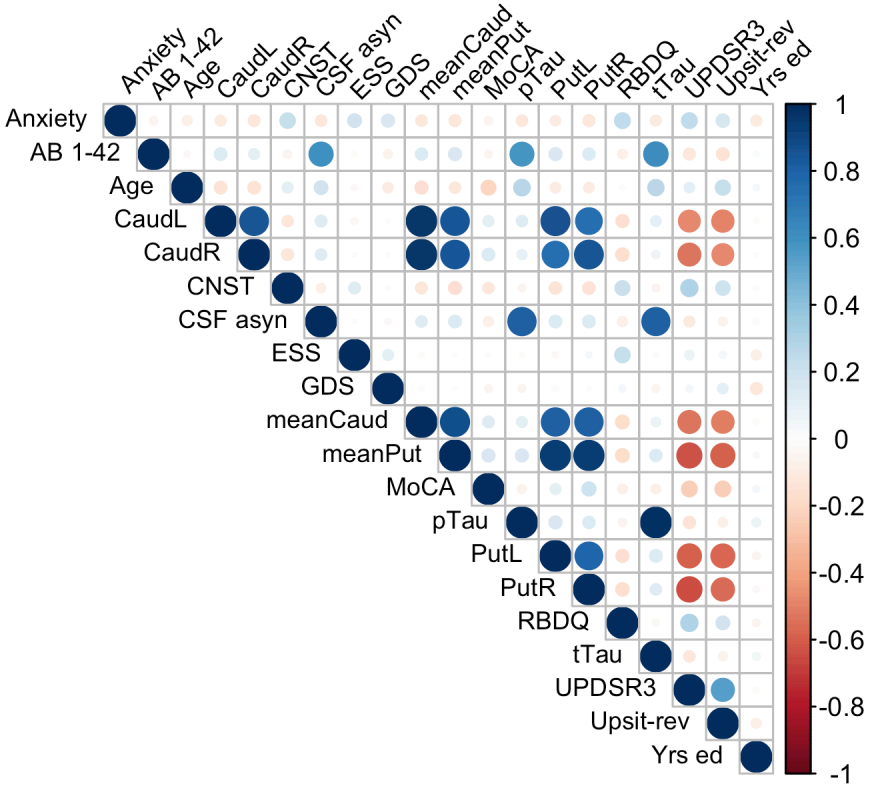


FIG 5 | *Predictor correlations (Spearman). Circle size is proportional to Spearman correlation; red indicates a negative correlation and blue a positive correlation; Anxiety, trait subscale from the State-Trait Anxiety Inventory; CNST, constipation based on MDS-UPDRS I; CaudL, left caudate; CaudR, right caudate; PutL, left putamen; PutR, right putamen; meanCaud, left + right caudate /2; meanPut, left + right putamen /2; UPDRS3, MDS-UPDRS III; Upsit-rev (hyposmia), University of Pennsylvania Smell Identification Test (a reverse scaled version); AB 1-42, beta-amyloid1-42; CSF a-syn, cerebral spinal fluid a-synuclein; pTau, CSF phosphorylated Tau; tTau, CSF total tau; MoCA, Montreal Cognitive Assessment; Yrs. ed., years of education. Note, CaudL, CaudR, PutL, PutR are dopamine transporter (DAT) DAT scan measures.*

Multicollinearity beyond the cutoff (.75) was found for pairwise combinations of CSF pTau, tTau and α-synuclein, as such these features were not combined in the same model (see Methods, 2.6 regarding collinearity). The correlation between pTau and tTau was r_s = .98. The correlations of α-synuclein and the tau proteins were r_s =.82 for pTau and α-synuclein, and r_s =.81 for tTau and α-synuclein. While several other predictors demonstrated significant correlations (details available on request) these correlations did not exceed .75. Fig 5 conveys the finding that both Upsit-rev and MDS-UPDRS III exhibited by far the strongest associations (negative associations) with SPECT DAT values (bilateral, mean caudate; bilateral, mean putamen DAT values). Again,

DAT values and MDS-UPDRS III, not modelled as predictors, are used here only as indices of disease. Fig 6, like Fig 5, indicates linkage of hyposmia and striatal DAT values. Again, olfactory loss or hyposmia is based on reverse-scaled UPSIT scores (i.e. higher values reflect greater hyposmia). The reverse-scaled UPSIT (hereafter referred as Upsit-rev) values are otherwise identical to standard (non-reverse scaled).

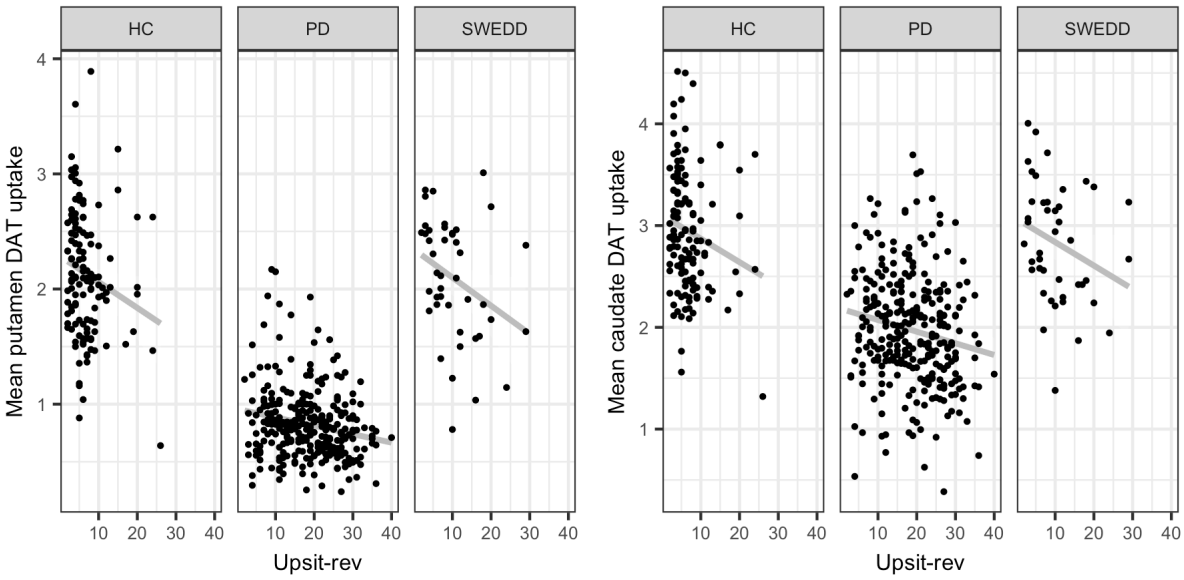


FIG 6 | ***Mean dopamine active transporter (DAT) values against hyposmia:*** *left, mean putamen DAT values against Upsit-rev = University of Pennsylvania Smell Test, reverse-scaled; right, mean caudate DAT values against Upsti-rev; HC= healthy controls; PD = early PD; SWEDD = scan without evidence of dopaminergic deficit*

Chapter Four: *Early PD versus control model classification*

As outlined under section *2.2*, final features, selected only from training set data, were determined by model-based feature elimination coupled with the AUC: models with features amounting to the highest model AUC constituted the final models applied to the test data sets. Caret package [3] internal cross-validation (10-fold, 5 repeats) resampling was used to tune tree-model (decision tree, Random forest, XGBoost) hyper-parameters and arrive at the optimal feature set. For more information regarding the caret package see *Modeling and the caret package* Supporting information III. Stepwise regression employing the Akaike information criterion (AIC)[120] was used for logistic regression feature elimination; the GAM model used the same features as the logistic regression model.

Model classification results in Chapters 4 and 5 (AUC, sensitivity, specificity, accuracy, Kappa) follow initial model-specific results (e.g. logistic regression coefficients and associated z-scores, tree-model hyper-parameter tuning settings, etc.) and features determined by model-based feature selection. For both early PD versus control (Chapter Four) and early PD versus SWEDD (Chapter Five) classification analyses, decision tree results are reported first followed by logistic regression and the general additive models. Subsequently, the random forest and the XGBoost model results are provided. A summary of classifier results for the early PD versus control and early PD versus SWEDD analyses is provided in Chapter Six.

With the exception of Chapter Six, the results for each algorithm are report in three tables. A table ranking predictor (feature) importance, and two additional tables: one showing a classification confusion matrix and the other revealing the calculated performance metrics. Tabulated feature-ranking results are provided for both built-in and caret package (M. Kuhn, 2019, March, 3) methods. The caret package has a generic 0-100% feature ranking scale, which conveniently ranks feature importance to classification on the same (0-100%) scale across different models.

The confusion matrices can be optionally used to verify sensitivity, specificity, Kappa and accuracy performance metrics. The performance metrics and the confusion matrices are provided at the default .50-classification threshold and the optimized classification threshold.

4.1 Decision tree classification

This section reviews the decision tree (early PD/control) classification. Of the original 14 predictors, the final feature set had 6 features of greatest importance to classification. The goodness of split ranking of features was largely in agreement with the generic feature elimination/ranking function in caret (M. Kuhn, 2019, March, 3). However, the latter included tTau rather than pTau among the top 6 predictors. The model with pTau had a marginally higher AUC; pTau was used rather than tTau. The final model predictors used are listed in **Table 3** in descending order of importance: e.g. Upsit-rev (reverse-scaled University of Pennsylvania Smell Identification Test) contributed most to model classification and age contributed the least. As noted in Methods and at the outset of Chapter Four, Scale 0-100%, in **Table 3**, refers features/predictors of importance ranked from 0 to 100% based on their importance to the model classification (M. Kuhn, 2019, March, 3). Goodness of split also ranks variables on a 0-100% scale of importance but is the *rpart* built-in variable of importance selection function. For further details on *rpart* Goodness of split and *rpar*t in general see Supporting information III (*The models: Logistic regression, general additive, decision tree, random forest and XGBoost*).

TABLE 3 | Predictor importance: decision tree model early PD/control

Predictors	*Scaled 0-100%*	*Goodness of split*
Upsit-rev	100.00	55.57
MoCA	49.78	28.78
RBDQ	16.30	10.92
pTau	13.49	9.42
Aβ_{1-42}	4.38	4.56
Age	0.00	2.22

$A\beta_{1-42}$ = beta amyloid $_{1-42}$; Upsit-rev = reverse scaled Upsit = University of Pennsylvania Smell Identification Test ; MoCA = Montreal Cognitive Assessment; pTau =tau phosphorylated at Thr181 at threonine 181 ($pTau_{181}$) ; RBDQ = rapid eye movement behavior disorder questionnaire; Goodness of split: number of times a variable or its surrogate is used to split the data; Scaled 0-100% = predictor importance to classification from 0-100%

Fig 7 provides a graphic of decision tree classification based on **Table 3** predictors. The single most important predictor is situated at the top of the tree (the root node) and predictors deemed of lesser importance by the algorithm are lower in the tree. Note, this version of the tree was only partially "pruned" to permit a more

complete view of the rpart decision tree classification process. A fully pruned tree is provided in Supporting information III (see Graphic 5).

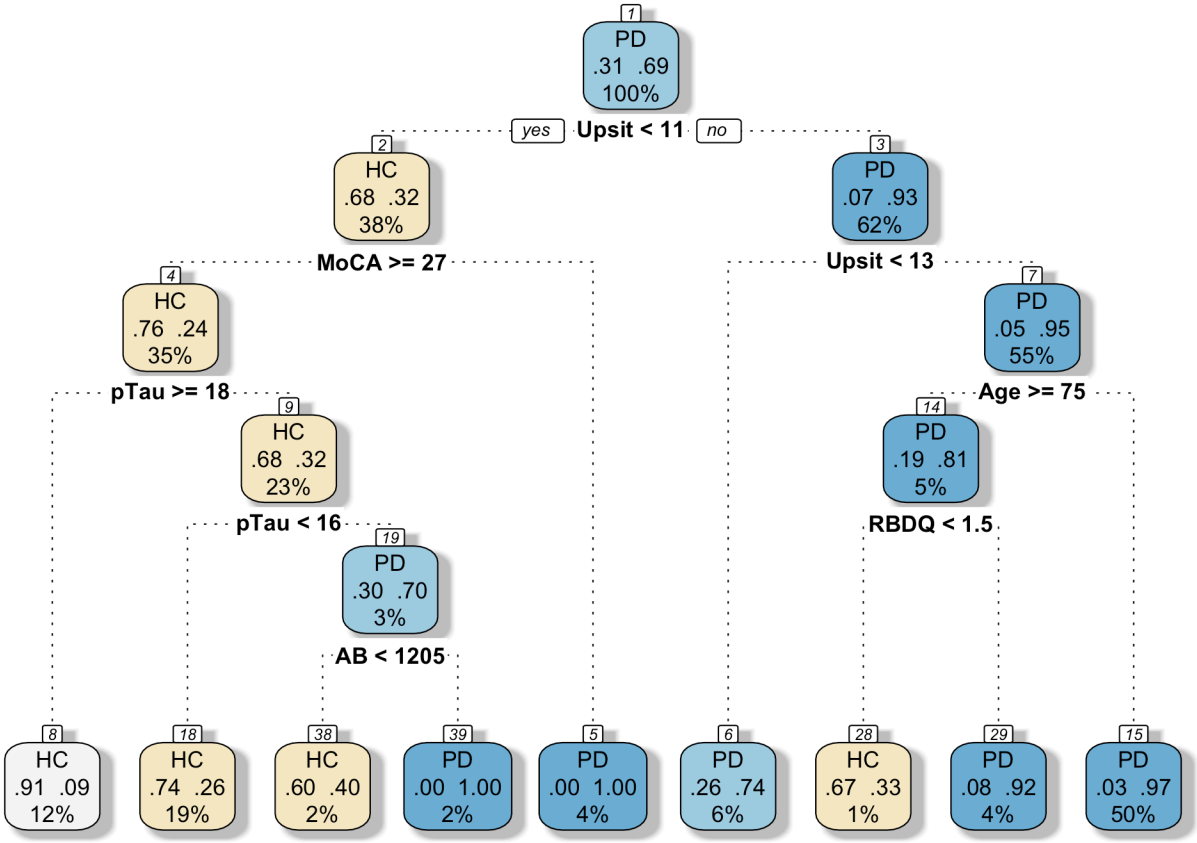


FIG 7 | *Tree plot, early PD versus controls*. *AB = beta-amyloid1-42 ($A\beta_{1-42}$); MoCA = Montreal Cognitive Assessment ; Upsit = reverse-scaled University of Pennsylvania Smell Inventory Test; pTau = tau phosphorylated at Thr181 at threonine 181 ($pTau_{181}$); RBDQ = rapid eye movement behaviour disorder. The top of each node displays the classification (and the very top number corresponds to the branch number), the mid-node shows the probability of the class at that node, and the lower node value is the percentage % of observations at the node. PD = early Parkinson's disease, HC = healthy controls*

The decision tree (*rpart*) default Gini index splitting rule was retained for model training. The pruned final tree model, that utilized these same top 6 predictors listed in **Table 3**, was based on a complexity parameter (*cp)* overfitting corrective measure of *cp*= 0.02197 (and minsplit = 9, minbucket = 3); the *cp* parameter (the amount by which splitting a given node reduced the relative error) was initially derived from the lowest cross-validated error (xerror) associated with the lowest number of splits in the tree model. Subsequently, and as with the other tree-based models (random forest and XGBoost) the AUC was the criterion used to select the optimal *cp*

value of the final model during k-fold (10 folds, 5 repeats) resampling. The resulting final model, with a k-fold tuned *cp* parmemter, also selected a *cp* = 0.0219. Supporting information III, again, provides added information on the rpart model.

The model AUC was .872 (95% CI: 0.831-0.913, DeLong; sensitivity .879; specificity .857). Table 4a confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. Table 4b provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided in Tables 4a and 4b. The optimal classification threshold was .586; it was selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a).

TABLE 4a | Decision tree confusion matrices, early PD/control

Class	OBS	Predicted, K-fold model, thr = .50		Predicted, K-fold model, thr = .586	
PD	88	TN = 35	FN = 16	TN = 35	FN = 16
HC	39	FP = 4	TP = 72	FP = 4	TP = 72

Note: *N = 127 (validation data); K-fold, where k =10 (repeated 5 times for 50 models); HC = healthy controls; OBS= observations; PD = early Parkinson's disease; Predicted; K-fold model = predictions rom k-fold model applied to validation data-set; thr = threshold, e.g. thr of .50 refers trained model class prediction on the validation data-set at the default threshold of .50; FN = false negative; FP = false positive; TN = true negative; TP = true positive*

TABLE 4b | Decision tree performance metrics, early PD/control

Threshold	Model	Resampling	Cross-validated performance metrics			
			ROC AUC (95% CI, DeLong)	*ACC*	*Kappa*	*SN*	*SP*
.50	Tree	10-fold, 5 repeats	.860 (95% CI: 0.799-0.922)	.842	.659	.818	.897
.586	Tree	10-fold, 5 repeats	.860 (95% CI: 0.799-0.922)	.842	.659	. 818	.897

Note: *Predicted results on the validation data-set.ACC = accuracy; Tree = decision tree; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = Sensitivity; Resampling = 10-fold, 5 repeats resampling of the model tuning parameter (the complexity parameter), whereby the optimal hyper-parameter setting was determined by the AUC*

The AUC metric was .860 (95% CI .79, .92), indicating an ~ 86% chance the model would correctly distinguish between early PD and controls. At the optimized classification threshold of .586 performance metrics remained unaltered. The ROC curve with the optimized classification threshold (.586) is provided in Fig 8.

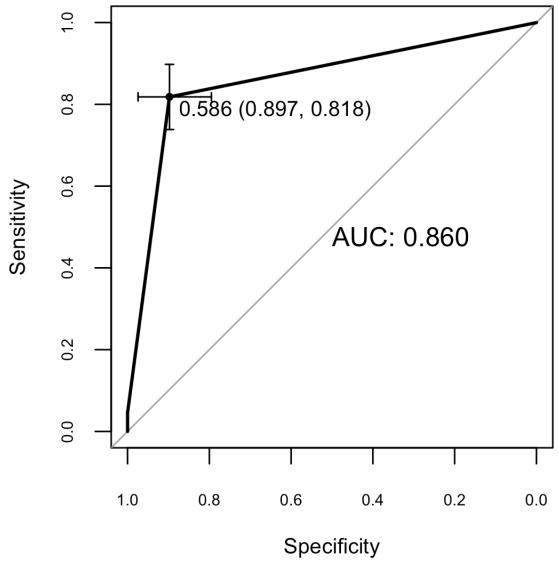


FIG 8 | ***Decision tree ROC AUC:*** *.860; specificity .897, sensitivity .818; error bars reflect variation in specificity and sensitivity threshold*

4.2 Logistic regression and general additive model classification

This section reviews the logistic regression (GLM) and general additive (GAM) (early PD/control) model-specific results, which includes feature elimination, model assumption diagnostics, and the tabulated logistic and GAM model results. This is followed by predictive performance results on the validation data-set. The logistic regression GAM was, as noted in Methods, used both to supplement the logistic regression GLM and as a distinct predictive classification model.

In **Table 5**, the final 6 predictors are ranked in order of greatest importance. The **Table 5** predictors are ranked on a 0-100% scale (M. Kuhn, 2019, March, 3), with the predictor of least importance being attributed zero value based on contribution to model classification; the predictors are also ranked by coefficient z-scores. The coefficient z-values are simply the coefficient z-values of the model with all 6 predictors included.

TABLE 5 | Predictor importance: logistic model

Predictors	*Scaled 0-100%*	*Coefficient z-values*
Upsit-rev	100	7.60
RBDQ	34.36	2.96
pTau	26.08	2.36

Age	20.43	1.96
MoCA median	20.42	1.96
CNST	15.40	1.60
MoCA high	0.00	0.51

Upsit-rev = reverse scaled UPSIT = University of Pennsylvania Smell Identification Test; pTau = CSF phosphorylated Tau; MoCA = the Montreal Cognitive Assessment; RBDQ = rapid eye movement behavior disorder questionnaire; Scaled = predictor importance to classification on a 0-100% scale (higher values indicate greater predictor importance)

Note that in **Table 5** MoCA is a categorical variable. It was originally a semi-continuous variable but violated linearity of the logit and was converted to a 3-level factor: the reference is the MoCA low range of scores (≤ 26). MoCA mid-range has a median score of 28, and MoCA high has a range of 29-30. While MoCA proved to be a variable of lesser importance it nevertheless contributed to the model and its conversion from continuous to categorical warrants brief explanation. The semi-contiguous MoCA (0-30 integer scale) violated linearity of the logit (Box & Tidwell, 1962). The remedial option used was to initially convert it to quartiles (e.g. https://rpubs.com/kaz_yos/logistic-linearity); quartiles were labeled Q1-Q4 constituted a 4-level categorical version of MoCA. There were only 3 cases equal to the base (Q1) score of 21 but there were 71 cases (all early PD) less than 27, which was the score marking the first quartile (.25). To reasonably represent the lower scores, the first two quantiles, Q1 and Q2, were merged, leaving a single low range for all instances with a score < 27. This resulted in a 3-level factor representing the MoCA low (21-26), midrange or median (28) and upper level scores (29-30).

The logistic regression analysis specifics of the predictors in **Table 5** are provided in **Table 6**. **Table 6** in addition to the coefficients ($\widehat{\boldsymbol{\beta}}$), standard errors ($\boldsymbol{SE}\,\widehat{\widehat{\boldsymbol{\beta}}}$), z-values and odds ratios, includes standardized predictor coefficients as well as rescaled coefficients (i.e. with score range binarized) (see 2.5 Statistics). Standardized and rescaled coefficients have been recommended (Cohen et al., 2003) to lend an unbiased perspective to differentially scaled predicators (e.g. olfactory deficit based on the University of Pennsylvania Smell test [UPSIT] has 0-40 scale, constipation based on the UPDRS I has a 0-4 scale), making such differentially scaled predictors comparable. Standardized predictor coefficients are provided for all predictors (except the categorical variable

MoCA); rescaled coefficients are provided for Likert-type scales UPSIT, rapid eye movement behaviour disorder (RBDQ), and constipation (CNST).

TABLE 6 | Logistic regression early PD/controls						**Standardized**		***Rescaled***	
Variable	$\widehat{\boldsymbol{\beta}}(\boldsymbol{SE})$	*z*	***95% CI***	***Odds***	***95% CI (Odds)***	$\widehat{\boldsymbol{\beta}}$*	***Odds****	$\widehat{\boldsymbol{b}}$	***p***
Intercept	-0.55(1.11)	-0.49							
Upsit-rev	0.27(0.03)	7.60	(.21, .35)	1.31	(1.23, 1.41)	2.48	11.95	10.82	< .0001***
RBDQ	0.23(0.07)	2.96	(.08, .40)	1.26	(1.09, 1.49)	0.66	1.93	3.05	= .003**
pTau	-0.08(0.03)	-2.36	(-.15, -.02)	.92	(0.86, .98)	-0.54	0.58	-	= .018*
Age	-0.03(0.01)	-1.96	(-.07, .00)	.97	(.93, 1.00)	-0.36	0.70	-	= .051
MoCA median	-0.93(0.47)	-1.96	(-1.87, -1.01)	.40	(.15, .99)	-	-	-	= .051
CNST	0.59(0.37)	1.60	(-.11, 1.36)	1.80	(.90, 3.91)	0.37	1.45	1.77	= .109
MoCA high	-0.22(0.43)	-0.51	(-1.06, .62)	.81	(.35, 1.86)	-	-	-	= .612

N = 298. CNST = MDS-UPDRS I NP1CNST; MoCA = Montreal Cognitive Assessment; Odds = odds ratio; Odds = standardized odds ratio; pTau = CSF phosphorylated Tau $_{181}$; RBDQ = rapid eye movement behaviour disorder questionnaire; Upsit-rev = reverse-scaled University of Pennsylvania Smell Identification Test;* $\widehat{\boldsymbol{\beta}}$** = standardized coefficient;* $\widehat{\boldsymbol{b}}$ *= rescaled coefficient (scaled variables have same 0 to 1 range); McFadden pseudo* R^2 *(*pR^2*) = .47; Residual deviance: 194.42 on 290 df; AIC= 210..42. P-values are rounded 3 decimal places; other values are rounded two decimal places.*

The biological predictors CSF α-synuclein, pTau, and tTau where highly correlated ($r_s > .75$). To avoid violation of assumption of multicollinearity these variables were not entered concurrently into the same model. **Table 6** specifies the final logistic model results derived from stepwise (backward, employing AIC) logistic regression, which controlled for age, years of education and gender. The pseudo R^2 (McFadden, 1974) of this final logistic model was .47, which suggests a very good model fit to the data (https://stats.stackexchange.com/questions/82105/mcfaddens-pseudo-r2-interpretation).

Elaborating briefly on **Table 6** standardized significant results, holding other variables constant, a one standard deviation increase in reduced olfactory acuity (Upsit-rev in **Table 6**) explains a 2.48 standard deviation increase in the log odds of early PD, an increase that significantly differed from zero, $z = 7.60$, $p < .0001$. For every standard deviation unit increase in reduced olfactory acuity the odds of early PD increased by a multiplicative factor of 11.95. Holding other variables constant, a one standard deviation increase in RDBQ was associated with a .66 standard deviation increase in the log odds of early PD, an increase that significantly differed from zero, $z = 2.96$, $p = .003$. Every standard deviation unit increase in RBDQ multiplies the odds of early PD by 1.93. Holding other variables constant, a one standard deviation increase in pTau was associated with a .54 standard deviation reduction in the log odds of early PD, a difference that significantly differed from zero, *z*

= -2.36, p = .02. Every standard deviation unit increase in pTau reduces the odds of early PD by about half (.58). The two most potent predictors, olfaction (reverse scaled UPSIT) and rapid eye-movement behaviour disorder (RBDQ), are depicted in effects plots in Fig 9, where the predicted probability of early PD group membership rises with higher predictor values.

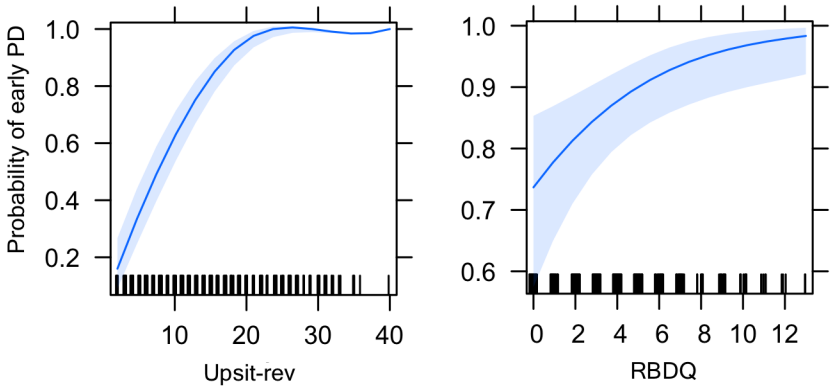


FIG 9 | *Effects plots*: *the probability of early PD membership (y-axis) increases with higher predictor values (x-axis); the y-axis is on the probability scale, shading around the lines reflects the 95% confidence interval. RBDQ = rapid eye movement behaviour disorder questionnaire; Upsit-rev = University of Pennsylvania Smell Identification Test (reverse-scaled).*

An insignificant (p = .55) Hosmer and Lemeshow Goodness of Fit test suggested that the model was an effective classifier; there was not a significant difference between observed and predicted values, $X^2(4) = 2.3$, p = .68. A final collinearity check using the variance inflation factor demonstrated no indications of high collinearity among predictors (the highest value was age at 1.24), and the Durbin-Watson test returned a value of 2.13, (p =. 22) indicating errors were not correlated. Logistic regression case-wise index plots of leverage, discrepancy and Cook's Distance, did not reveal any cases sufficiently extreme to bias model results (see Supporting information I case-wise plots Figs S1-S6). While 4% of cases (13/298) had leverage beyond the cut-off (.06 here) and 4 cases exceeded a discrepancy of > ± 2.5, there were no cases that exceeded both these leverage and discrepancy criteria. In short, none of the cases attained a Cook's Distance of 1 (the highest Cooks Distance value was .09). In addition, other than the MoCA variable already discussed, none of the other five predictors violated linearity of the logit.

The logistic regression GAM (referred to hereafter as just GAM) model specifics are provided in Supporting Information II, Table S2-1. The GLM included conversion of the semi-continuous MoCA in to

quartiles. The MoCA variable violated the assumption of linearity of logit (Box & Tidwell, 1962) and was converted in to quartiles. Although such conversion is common practice, it can result in a loss of information (Bennette & Vickers, 2012). The GAM model, which can retain all information, helped to further convey the relation of the MoCA variable to the probability of early PD classification (see *2.5 Statistical analyses*). The Box-Tidwell (Box & Tidwell, 1962) test for olfactory deficit (Upsit-rev), age, pTau, RDBQ and constipation all had insignificant Box-Tidwell *p*-values, though judging purely from the S1-2 plots these variables also had a degree of non-linear association with outcome. In short, although the Box-Tidwell test identified only MoCA was in violation of linearity of the logit, other predictors (notably age, pTau and olfactory deficit) had, arguably, a less than linear relationship with outcome on the logit scale. The utility of the logistic regression GAM to capture the non-linear characteristics of these variables was apparent; residual deviance was lowest and explained deviance (pseudo R^2) were highest when all predictors, not simply MoCa, were transformed into thin plate smoother splines. The GAM model deviance (161.90) was significantly lower than logistic model deviance (194.43), X^2 (2.66) = 32.52, $p < .0001$. GAM pseudo R^2 (55.9%) was also higher relative to logistic regression pseudo R^2 (47.0%). The GAM model output is provided in Supporting Information II, Table, S2-1. Interestingly, and as shown in Table S2-1, the MoCA variable was insignificant. In addition, of the six predictors, MoCA made the smallest contribution to the GAM model. The predictor ranking in descending order of importance was as follows based on p-values: Upsit-rev, RBDQ, age, pTau, constipation, and MoCA. With respect to GAM diagnostics, the estimated degrees of freedom (edf) did not closely approach predictor *k*-value. The default $k = 10$ (k-1= 9 maximum) produced good results but the following *k* values proved optimal to lower (residual) deviance and increase deviance explained (analogous to pseudo R^2): age, $k= 3$; olfaction, $k= 6$; RDBQ, $k = 4$; MoCa, $k = 5$; and pTau, $k=4$; constipation, $k=4$. The residuals in the (quantile-quantile) qq.gam plot (see Supporting Information II, Fig S2-1) adhered quite closely to the diagonal, suggesting the data set quantiles were relatively well behaved and approximated the theoretical quantiles. Scatterplots of all predictors are provided in Supporting Information II, Fig S2-2.

The logistic regression (GLM) model AUC was .920 (95% CI: 0.8877-0.9533, DeLong; sensitivity .934, specificity .797). Single predictor models using the top two predictors Upsit-rev (reverse-scaled UPSIT) and RBDQ had poor (RDBQ, AUC .674) to good (Upsit-rev AUC .889) AUC outcomes.

The GAM model ROC AUC was .946 (95% CI: 0.9217-0.9702, DeLong; sensitivity .923, specificity .850). The GLM and GAM model predictive results are summarized in **Tables 7a** and **7b**. The Table 7a confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. Table 7b provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided.

The optimized threshold for the GLM was .462; the optimized threshold for the GAM was .534. The optimized model thresholds were selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a). The cross-validated GLM ROC AUC metric was .907 (95% CI: 0.8493-0.9642 (DeLong)), indicating an ~ 91% chance the model would correctly distinguish between early PD and controls. The cross-validated GAM ROC AUC metric was .928 (95% CI: 0.8778-0.9783 (DeLong)), indicating an ~ 93% chance the model would correctly distinguish between early PD and controls. The GLM and GAM model ROC curves are shown in Fig 10. With the exception of specificity, performance metrics were higher in the GLM and GAM models relative to same metrics in the decision tree.

TABLE 7a | Logistic regression GLM and GAM confusion matrices

Model	Class	OBS	Predicted; thr = 0.5		Predicted; opt. * thr.	
GLM	PD	88	TN = 34	FN = 9	TN = 34	FN = 8
GLM	HC	39	FP = 5	TP = 79	FP = 5	TP = 80
GAM	PD	88	TN = 34	FN = 9	TN = 35	FN = 9
GAM	HC	39	FP = 5	TP = 79	FP = 4	TP = 79

*N= 127 (validation data-set); OBS= observations; PD = early PD; opt.= optimized threshold (Youden Index); thr = threshold, e.g. thr of .50 refers to trained model class predictions on validation data-set at the default thre-shold of .50.; HC = healthy control; FN = false negative; FP = false positive; TN = true negative; TP = true positive; * see TABLE 7b for optimized thresholds*

TABLE 7b | Logistic regression GLM and GAM performance metrics

Threshold	Model	Resampling	Performance metrics: validation data-set				
			ROC AUC (95% CI)	*ACC*	*Kappa*	*SN*	*SP*

.50	GLM	None	.907 (95% CI: 0.849-0.964, DeLong)	.890	.748	.898	.872
.462	GLM	None	.907 (95% CI: 0.849-0.964, DeLong)	.898	.764	.909	.872
.50	GAM	None	.928 (95% CI: 0.878-0.978, DeLong)	.890	.748	.898	.872
.534	GAM	None	.928 (95% CI: 0.878-0.978, DeLong)	.898	.768	.898	.897

Predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = sensitivity; Kappa = Cohen's Kappa; ROC AUC = receiver operating characteristic area under the curve.

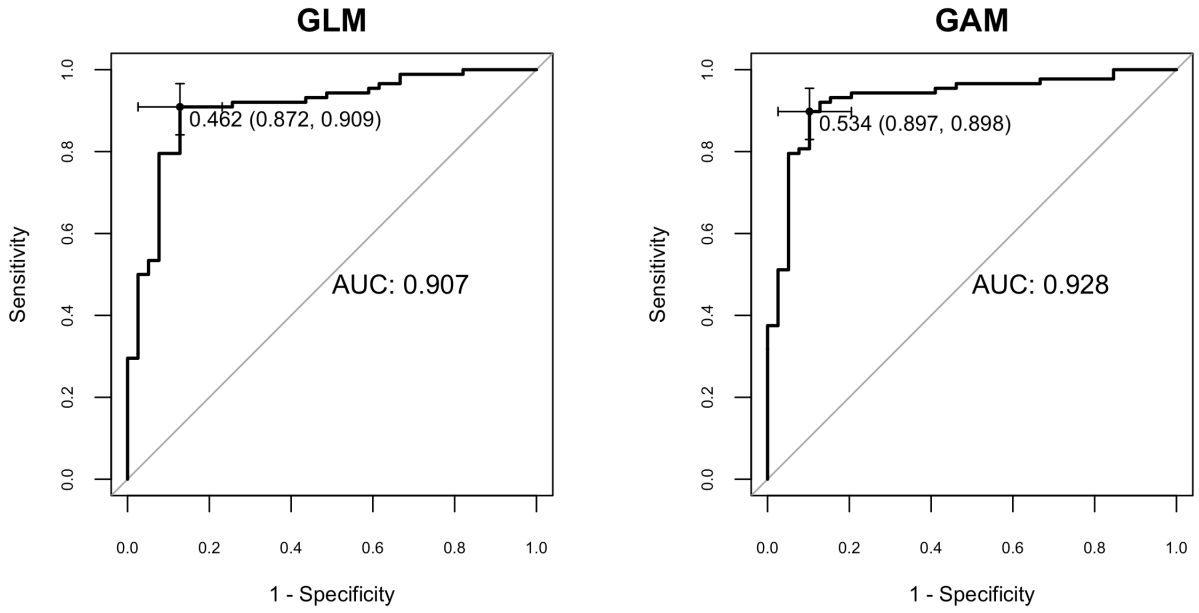


FIG 10 | *Logistic regression GLM and GAM AUC plots:* *GLM ROC AUC: .907 (95% CI: 0.849-0.964, DeLong), specificity .872, sensitivity .909; GAM ROC AUC .928 (95% CI: 0.878-0.978, DeLong), specificity .897, sensitivity .898; error bars reflect variation in specificity and sensitivity threshold*

4.3 Random forest classification

This section reviews the random forest (early PD/control) training data, model-specific results followed by predictive performance results on the validation data-set. From the 14 original predictors, the six most important predictors determined by feature elimination are listed in descending order of importance in **Table 8**. The predictor values listed under Scaled 0-100% in **Table 8** are ranked on a 0-100% scale (M. Kuhn, 2019, March, 3) based on contribution to the model. Predictor values in **Table 8** listed under Mean Gini decrease are ranked by the extent to which they contributed to the mean decrease in the Gini index (the decrease in impurity), across all trees. The Mean Gini decrease is the built-in random forest measure of predictor importance.

TABLE 8 | Predictor importance, random forest

Predictors	*Scaled 0-100%*	Predictors	*Mean Gini decrease*
Upsit-rev	100	Upsit-rev	43.75
MoCA	40.83	MoCA	14.91
RBDQ	23.83	RBDQ	14.39
CSF α-syn	1.59	$A\beta_{1-42}$	16.03
Age	.202	CSF α-syn	15.61
$A\beta_{1-42}$	0.00	Age	14.43

$A\beta_{1-42}$ = beta amyloid $_{1-42}$; CSF α-syn = CSF alpha synuclein; Mean Gini decrease = variable importance to average Gini (impurity) decrease; Upsit-rev= reverse scaled University of Pennsylvania Smell Identification; RBDQ = rapid eye movement behaviour disorder questionnaire; Vars Imp Scaled = predictor importance to classification on a 0-100% scale

The AUC criterion, used to select the optimal tuning parameters during the k-fold (10-fold x 5 repetitions) model training procedure, determined an optimal m_{try} = 1 hyper-parameter, where the m_{try} setting refers to the number of randomly sampled predictors used to spit data. The optimal model, based on the predictors in **Table 8**, also had an ntree (the number of trees grown) value of 3000 trees and a node size of 3. The out of bag (OOB) error rate was 14.09%. The OOB error is graphed in Fig 11. See Supporting information III for an explanation of OOB error in random forest.

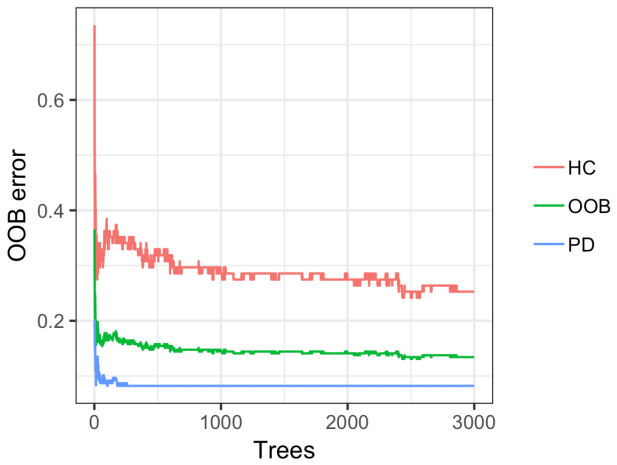


FIG 11 | *Out of bag error plot:* *HC = healthy controls; OOB = out of bag error; PD = early Parkinson's disease*

The model ROC AUC was .999 (95% CI: 0.9989-1, DeLong; sensitivity .990, specificity 1.00). Model predictive results (from applying the model to the validation/test data-set) are summarized in **Tables 9a** and **9b**.

The **Table 9a** confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. **Table 9b** provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided. The optimized threshold (.534) was selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a).

The predicted cross-validated ROC AUC was of .913 (95% CI: 0.8585-0.9672, DeLong) indicated an approximate 91% probability the model would correctly distinguish between early PD and controls. This was a higher AUC than the cross-validated (CV) .860 achieved by the decision tree model, comparable to the CV AUC of logistic model (rounded to two decimal places), and a little lower than the GAM .928 CV AUC.

TABLE 9a | Random forest confusion matrices

Class	OBS	Predicted, K-fold model, thr = 0.5		Predicted, K-fold model, thr = .541	
PD	88	TN = 31	FN = 5	TN = 34	FN = 8
HC	39	FP = 8	TP = 83	FP = 5	TP = 80

N= 127 (validation data-set); K-fold, where k =10 (repeated 5 times for 50 models); OBS= observations; K-fold model= predictions from k-fold model applied to validation data-set; thr = threshold, e.g. thr of .50 refers trainedmodel class prediction on validation data-set at the default threshold of .50FN = false negative; FP = false positive; TN = true negative; TP = true positive

TABLE 9b | Random forest performance metrics

Threshold	Model	Resampling	Cross-validated performance metrics				
			ROC AUC (95% CI, DeLong)	*ACC*	*Kappa*	*SN*	*SP*
.50	RF	10-fold, 5 repeats	. 913 (95% CI: 0.858-0.968)	.898	.754	.943	.795
.534	RF	10-fold, 5 repeats	. 913 (95% CI: 0.858-0.968)	.898	.764	.909	.872

Predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = sensitivity; Kappa = Cohen's Kappa; RF = random forest; Resampling = 10-fold, 5 repeats resampling of the model tuning parameters (i.e. mtry, but ntree, nodesize were also assessed), whereby the optimal hyper-parameter setting was determined by the AUC

A plot of the AUC, with sensitivity and specificity at the optimized threshold, is provided in Fig 12.

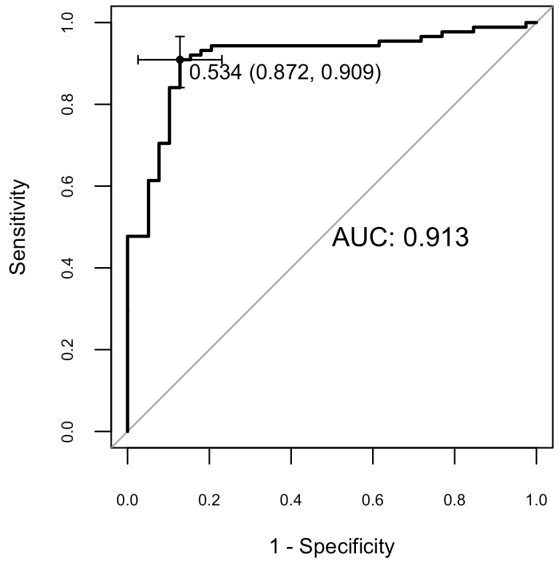


FIG 12 ***Random Forest AUC plot****: .913****;*** *specificity .872, sensitivity .909; error bars reflect variation in specificity and sensitivity threshold*

4.4 XGBoost classification

This section reviews the XGBoost (early PD/control) classification analysis. Feature selection results are followed by review of XGBoost optimized tuning parameters determined by resampling. Subsequently, the predictive performance outcome on the validation data-set is reviewed. The 12 most important predictors are listed in descending order of importance in **Table 10**. The predictors values listed under Scaled 0-100% in Table 10 are ranked on a 0-100% scale (M. Kuhn, 2019, March, 3) based on contribution to the model. Under the **Table 10** heading Gain, predictors are ranked by the extent to which they contributed to the XGBoost gain index. Gain is a predictor importance metric built-in to XGBoost that measures a feature's contribution to each tree in the model and hence overall prediction; higher values are better. Note the built-in and caret generic (0-100%) ranking of predictor importance to classification in **Table 10** differs after the top two predictors.

TABLE 10 | Predictor importance, XGBoost

Predictors	Scaled 0-100%	Predictors	Gain
Upsit-rev	100	Upsit-rev	0.5436
MoCA	27.23	MoCA	0.1404
$A\beta_{1-42}$	15.25	RBDQ	0.0535
α-syn	13.43	$A\beta_{1-42}$	0.0523

RBDQ	12.77	Anxiety	0.0486
Age	9.98	Age	0.0416
Anxiety	9.20	α-syn	0.0414
CNST	8.15	EDYRS	0.0303
EDYRS	5.97	CNST	0.0193
Depression	2.77	Depression	0.01932
Sex	2.58	ESS	0.0114
ESS	0.00	Sex	0.002

$A\beta_{1-42}$ = CSF beta amyloid $_{1-42}$; Anxiety = trait anxiety from the State-trait anxiety inventory; CNST = constipation based on the UPDRS I scale; α-syn = CSF alpha synuclein; EDYS = years of eduction; ESS = Epworth daytime sleepiness scale; Gain = variable importance to prediction, higher is better; GDS = Geriatric depression scale; MoCA = Montreal cognitive Assessment test; Upsit-rev= reverse scaled University of Pennsylvania Smell Identification; RBDQ = rapid eye movement behaviour disorder questionnaire; Vars Imp Scaled = predictor importance to classification on a 0-100% scale

A range of tuning parameter settings were assessed during resampling (10 fold, repeated 5 times) using the ROC AUC as the optimal parameter selection criterion. Tuning parameter guidelines adopted (O. Zhang, 2015) were endorsed by the creators of XGBoost (T. Chen, Guestrin, C., 2016). The optimal tuning parameters found were as follows: eta .02; nrounds 500; max_depth 5; min_child_weight 1; colsample_bytree .4; gamma 5; subsample setting of .5. For additional information regarding XGBoost hyper-parameters see *Supporting information III* (specifically see XGBoost with the section entitled *The models: Logistic regression, general additive, decision tree, random forest and XGBoost)*.

The model ROC AUC was .958 (95% CI: 0.9374-0.9795, DeLong; sensitivity .937; specificity .835). Model predictive results are summarized in **Tables 11a** and **11b**. The **Table 11a** confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. **Table 11b** provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided. The optimized model threshold (.660) was selected by the pROC package (Robin et al., 2011) where the optimal threshold was the point nearest to the top-left point of perfect sensitivity/specificity. A plot of the XGBoost AUC is provided in Fig 13.

The AUC statistic on the (external) validation data set was .923 (95% CI: 0.8747-0.9721, DeLong), indicating an approximate 92% chance the model would correctly distinguish between early PD and controls. This

was a higher AUC than the .860 achieved by the decision tree, similar to the random forest (.911) and logistic model (.907) AUCs but marginally less than the .928 AUC achieved by the GAM model.

TABLE 11a | XGBoost confusion matrices

Class	OBS	Predicted, K-fold model, thr = 0.5		Predicted, K-fold model, thr = .660	
PD	88	TN = 31	FN = 9	TN = 35	FN = 11
HC	39	FP = 8	TP = 79	FP = 4	TP = 77

N= 127 (validation data-set); K-fold, where k =10 (repeated 5 times for 50 models); OBS= observations; K-fold model = predictions based on k-fold model applied to validation data-set; thr = threshold, e.g. thr of .50 refers trainedmodel class prediction on validation data-set at the default threshold of .50.; FN = false negative; FP = false positive; TN = true negative; TP = true positive

TABLE 11b | XGBoost performance metrics

Threshold	Model	Resampling (k-fold)	Performance metrics: validation data-set			
			ROC AUC (95% CI, DeLong)	*ACC*	*Kappa*	*SN*	*SP*
.50	XGB	10-fold, 5 repeats	.923 (0.875-0.972)	.866	.708	.898	.795
.660	XGB	10-fold, 5 repeats	.923 (0.875-0.972)	.882	.736	.875	.897

K-fold model predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = sensitivity; Kappa = Cohen's Kappa; RF = random forest; ROC = receiver operating characteristic;Higher performance outcomes are in bold type; XGB = XGBoost model. Resampling = 10-fold, 5 repeats resampling of the model tuning parameters (e.g. eta, nrounds, etc), whereby the optimal hyper-parameter setting was determined by the AUC

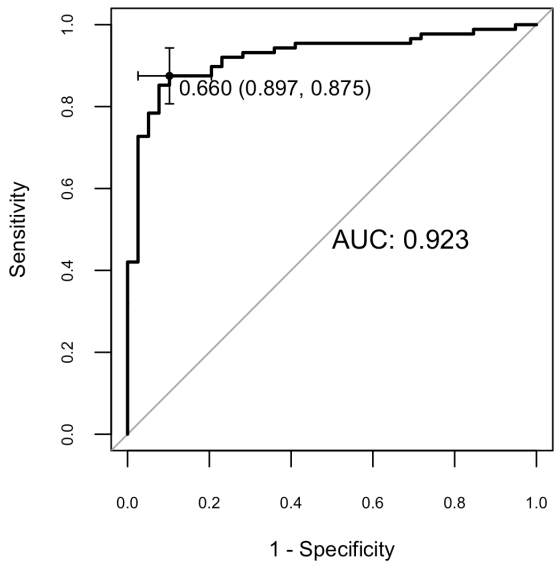


FIG 13 | *XGBoost AUC plot*: *.923 (95% CI: 0.875-0.972, DeLong); specificity is .897, sensitivity is .875; error bars reflect variation in specificity and sensitivity; optimized threshold was .660 determinded the point nearest o the top-left point of perfect sensitivity/specificity*

4.5 Long-term conversion of SWEDD to PD

In the preceding early PD/control classification analyses the GAM model achieved the highest AUC of .928 (sensitivity .898, specificity .897) when applied to the (early PD/control) validation data. The XGBoost model, when applied to the validation data, had a similar result (AUC .923, sensitivity .875, specificity .897). The GAM model (using features selected by stepwise regression), was more parsimonious, and used far fewer predictors than the XGBoost model (see sections *4.2* and *4.4* for details).

This section reports the results of these two higher performing early PD versus HC classifiers applied to the SWEDD data. Model performance on validation data is reviewed as is model utility predicting conversion of SWEDD to PD or PD-like at 12-36 months post baseline. The (early PD/control) classifiers map to SWEDD versus control. The results indicate the extent that model predicted SWEDD were classified as early PD – the estimated model predicted conversion of SWEDD to early PD. Note, again, that the early PD/control classifiers are being applied to SWEDD/control test data. As such, the diverse mix of PD-like (e.g. essential tremor, peripheral neuropathy, psychogenic illness, apparently normal etc.) as well as PD neuropathology that makes-up SWEDD is being shoehorned to fit early PD. Consequently, what the models classify as PD will be a mix of conditions, with the actual diagnosis being revealed in the longitudinal data (see below, *Longitudinal diagnosis vs. model predicted SWEDD to PD conversion*). The control data is the same control validation data used to validate the early PD/control models. The validation data set comprised of SWEDD (*n*= 43, 25 male) and control validation sets (*n* = 39 controls, 26 male) was isolated from the models during training.

Table 12 contains the performance metrics for the two early PD/control classifiers applied to the SWEDD/control validation data. **Table 12** includes results at the default .50 and optimal classification thresholds. The optimized model threshold was selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a). The ROC AUC plots of the early PD versus control models applied to SWEDD versus control validation data are provided in Fig 14.

TABLE 12 | XGBoost and GAM models applied to SWEDD versus control test data

Threshold	Model	Resampling (k-fold)	Performance metrics: validation data-set

			ROC AUC (95% CI, DeLong)	*ACC*	*Kappa*	*SN*	*SP*
.50	XGB	10-fold, 5 repeats	.831 (0.742-0.921)	.732	.466	.674	.795
.378	XGB	10-fold, 5 repeats	.831 (0.742-0.921)	.805	.608	.837	.769
.50	GAM	None	.863 (0.786-0.941)	.780	.564	.698	.872
.389	**GAM**	None	.863 (0.786-0.941)	.829	.658	.814	.846

*K-fold model predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = sensitivity**;** Kappa = Cohen's Kappa; RF = random forest; XGB = XGBoost model. Resampling = 10-fold, 5 repeats resampling of the model tuning parameters (e.g. eta, nrounds, etc), whereby the optimal hyper-parameter setting was determined by the AUC*

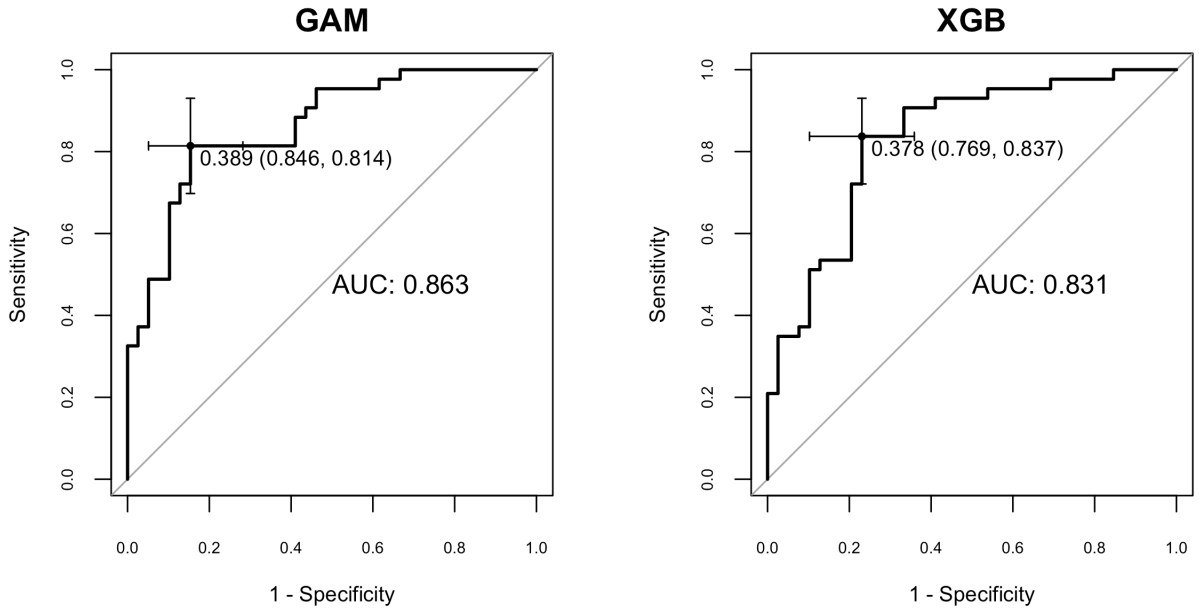


FIG 14 | ***GAM, XGBoost models AUC for SWEDD/control validation data.*** *Left**,** GAM = general additive model, ROC AUC .863 (95% CI: 0.786-0.941, DeLong), at the optimized threshold of .389 sensitivity = .814, specificity = .846); Right, XGBoost model, ROC AUC .831 (95% CI: 742-0.921, DeLong)**,** at the optimized threshold of .378 specificity =.769, sensitivity =.837. Error bars reflect variation in specificity and sensitivity threshold*

GAM and XGBoost model performance metrics were very similar (see **Table 12**). Not surprisingly, 39 of the validation data samples classified by the GAM model as PD or PD-like matched the 39/45 (87%) validation sample cases classified by the XGBoost model as PD or PD-like. In keeping with the similar model results, a modified (Robin et al., 2011) bootstrap (n = 2000) test (Hanley & McNeil, 1982) for correlated ROC curves indicated there was not a significant ROC AUC difference between the GAM and XGBoost models, D = 1.27, p = .21.

Reiterating the sensitivity and specificity results, at the optimized classification threshold (.389), the GAM model differentiated 35/43 (81.4%, sensitivity) SWEDD as PD-like. There were 33/39 (84.6%, specificity) subjects correctly classified as controls. For the XGBoost model, the optimized classification threshold (.378) differentiated 36/43 (83.7%, sensitivity) SWEDD as PD-like or early PD. There were 30/39 (76.9 %, specificity) subjects correctly classified as controls. Next, the PPMI longitudinal diagnoses were compared with model predicted classification to determine the percentage of actual SWEDD to PD conversions relative to cases that were PD-like but proved to not actually be PD pathology.

4.6 Longitudinal diagnosis vs. model predicted SWEDD to PD conversion

In the available longitudinal PPMI curated 12-24 month data, 12/38 (32%) GAM model (AUC .863; sensitivity .814; specificity .846) SWEDD cases predicted to be PD-like were diagnosed as idiopathic PD. However the majority of GAM model PD-like cases, 26/38 (68%), were not diagnosed as idiopathic PD longitudinally but rather were a mix of almost a dozen non-PD pathologies (e.g. essential tremor, psychogenic illness, etc.) and apparently normal cases. Four of those predicted by the model to convert to idiopathic PD (4/12) and re-diagnosed by 24 months as PD had DAT scan evidence of likely dopaminergic dysfunction as suggested by relatively low mean putamen DAT scan values (M = 1.24, SD= .73). The mean putamen DAT scan value of 26/38 cases predicted by the model to convert from SWEDD to PD but that converted to status other than PD was 2.01 (SD= .47), and the mean putamen DAT scan value of the model predicted control cohort was 2.14 (SD = .57). The 24-month time point diagnosis data also indicated that none of the 33 model predicted controls converted to PD.

In the available longitudinal PPMI curated 12-24 month data, 13/43 (30%) SWEDD predicted by the XGBoost model (AUC .831; sensitivity .837; specificity .769) to be PD-like had converted from SWEDD to idiopathic PD. But, as with the GAM model, the majority of XGBoost model PD-like data instances, 34/43 (79%), were not classified in the 12-36 month longitudinal data as idiopathic PD but diagnosed as a collective of diverse disorders similar to those noted for the GAM model. The XGBoost model predicted conversions also included the same four SWEDD subjects found by the GAM model with lower DAT scan values suggestive of dopaminergic dysfunction. The mean putamen DAT scan value of the remaining 20/34 (59%) cases predicted by

the model to convert to PD but that converted to status other than control or PD was 2.05 (*SD*= .49); the mean putamen DAT scan value of the controls was 2.09 (*SD* = .57). As with the GAM model, none of the model predicted controls were diagnosed at 12-24 months as PD.

Chapter Five: *Early PD versus SWEDD classification, SMOTE-based models*

This section reviews results for the early PD/SWEDD classification. Consistent with the early PD/control analyses, the early PD versus SWEDD model-specific results (e.g. tree model optimal tuning parameter settings, logistic regression coefficients and odds ratios, etc.) are followed by predictive performance results. Also, as in Chapter Four, both built-in and caret package (M. Kuhn, 2019, March, 3) (0-100% scale) feature-ranking methods are provided. In addition, classification results are presented in the same two-table format consisting of a confusion matrix table and a separate classification table of performance metrics. Further, a summary of all classifier results is provided in Chapter Six.

Unique to this chapter, subsampling SMOTE (i.e.. Synthetic Minority Oversampling Technique) was used to mitigate the SWEDD 13% minority class rate in the training data. As noted previously, resampling (10-fold, 5 repeats, distinct from subsampling) was applied to tree-based models (decision tree, random forest, XGBoost) to tune hyper-parameters but not to the GLM or GAM models. The GLM does not have hyper-parameters to resample (though the parameters/coefficients can be resampled but were not here), and at the time of this writing the resampling interface used (M. Kuhn, 2019, March, 3) did not permit specification of GAM formula particulars deemed important to model optimization.

Subsampling methods initially assessed were ROSE, and SMOTE. However, the ROSE model introduced negative synthetic values with each implementation; an occurrence reported elsewhere (Mani, 2015). Given all values in the data were positive, the injection of synthetic negative values was deemed unrealistic and the ROSE method was not employed in the analyses. Moreover, SMOTE subsampled models consistently had a higher AUC than both ROSE and non-subsampled models. As such, SMOTE subsampling was used to augment and balance the early PD SWEDD training data. Model results are presented in the following order: decision tree, logistic regression, general additive, random forest and XGBoost.

For all early PD/SWEDD models, except the decision tree model, the SMOTE training data set had a balance of 44 early PD (30 male) and 44 SWEDD (30 male). The early PD/SWEDD decision tree model used SMOTE data obtained inside of resampling (see section *5.1 below)* resulting in 88 early PD (63 male) and 66

SWEDD (41 male). The validation data set was left in its original class-asymmetric state: 147 early PD and just 21 SWEDD (with a gender division of 10 females to 11 males).

5.1 Decision tree early PD/SWEDD classification

An additional subsampling related finding occurred only with the decision tree model. Unlike the other tree models, decision tree SMOTE data was derived from subsampling inside k-fold (10 fold x 5 repeats) resampling (not possible at the time of this writing with the other models). A modified (Robin et al., 2011) bootstrap (n = 2000) test (Hanley & McNeil, 1982) for ROC curves indicated the model AUC from the SMOTE subsampled data, where the SMOTE data was obtained during resampling, had a significantly higher AUC (.932) compared to the AUC (.799) of the model based on SMOTE data that was not subsampled during k-fold resampling, D = 2.66,p = .008. An explanation and example of subsampling during (inside) resampling is readily available (M Kuhn, 2019). Consequently, the SMOTE data derived from subsampling during resampling was used (n = 88 early PD, 63 male; n = 66 SWEDD, 41 male). Again, this applied only to the decision tree model.

Of the original 14 predictors, built-in feature elimination (goodness of split) determined 10 predictors of importance from the SMOTE subsampled training data; these features are listed in **Table 13**. As shown in **Table 13**, both the *rpart* built-in variable of importance and the generic predictor ranking method (M. Kuhn, 2019, March, 3) rank variables on a 0-100% scale. Note that, with the exception of the top two predictors, the descending order of predictor rank differs between the generic and built-in (goodness of split) variable of importance methods.

TABLE 13 | Predictor importance: decision tree model, early PD/SWEDD

Predictors	*Scaled 0-100%*	*Predictors*	*Goodness of split*
Upsit-rev	100.00	Upsit-rev	24.64
ESS	89.69	ESS	16.20
Age	87.44	MoCA	10.53
EDYRS	79.76	tTau	10.52
tTau	56.77	Age	9.67
$A\beta_{1-42}$	51.20	RDBQ	9.03
MoCA	29.98	$A\beta_{1-42}$	7.22
Anxiety	5.59	EDYRS	3.10

RBDQ	3.81	Depression	2.71
Depression	0.00	Anxiety	1.83

Aβ $_{1\text{-}42}$ = beta amyloid $_{1\text{-}42}$; Depression = Geriatric depression scale; Upsit-rev = reverse scaled UPSIT = University of Pennsylvania Smell Identification Test ; MoCA =Montreal Cognitive Assessment; tau phosphorylated at Thr181 at threonine 181(pTau$_{181}$RBDQ = rapid eye movement behavior disorder questionnaire; Goodness of split: number of times a variable or its surrogate is used to split the data; Scaled = predictor importance to classification on a 0-100%

Fig 15 is a tree plot utilizing **Table 13** predictors. The single most important predictor to classification of group membership is situated at the top of the tree (the root node). The predictors deemed of lesser importance by the algorithm are lower in the tree. The decision tree (*rpart*) default Gini index splitting rule was retained for model training. For the pruned final model, the predictors listed in **Table 13** were based on a complexity parameter (*cp)* overfitting corrective measure of *cp*= .009 (minisplit = 20, minibucket= 7, maxdepth= 30). The AUC was the criterion used to select the optimal *cp* value of the final model during a k-fold model training procedure.

The model AUC was .932 (95% CI: 0.8936-0.9712, DeLong; sensitivity .891, specificity .900). Model (cross-validated) predictive results are summarized in **Tables 14a** and **14b**. The **Table 14a** confusion matrices specify the classification frequency predicted by the model applied to the validation data-set.

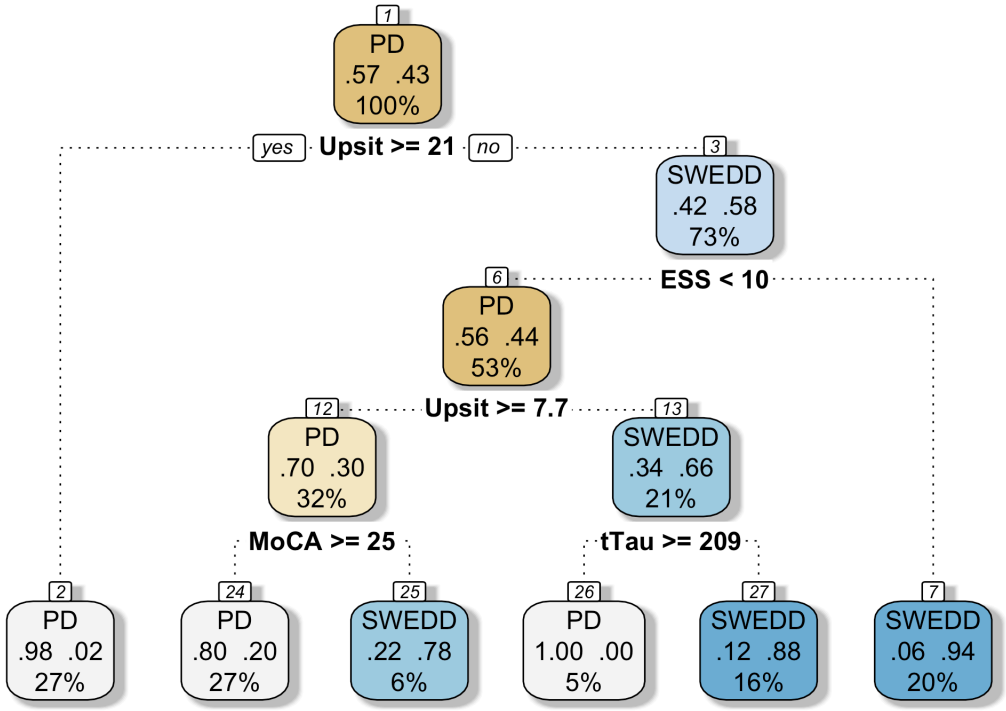


FIG | *15: Tree plot, Early PD/ SWEDD:* *MoCA = Montreal Cognitive Assessment; Olfaction = reverse-scaled University of Pennsylvania Smell Inventory Test; (Upsit-rev) tTau = total Tau; ESS = Epworth Daytime Sleepiness scale. The top of each node displays the classification (and the very top number corresponds to the branch number), the mid-node shows the probability of the class at that node, and the lower node value is the percentage (%) of observations at the node. PD = early Parkinson's disease, SWEDD = scans without dopamine deficit*

Table 14b provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided. The optimized model threshold (.486) was selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a).

TABLE | 14a: Decision tree confusion matrices, early PD/SWEDD

Class	OBS	Predicted, K-fold model, thr = 0.5		Predicted, K-fold model; opt. thr. = .486	
PD	147	TN = 120	FN = 27	TN = 120	FN = 7
SWEDD	21	FP = 7	TP = 14	FP = 27	TP = 14

Note: *N= 168 (validation data-set); K-fold, where k =10 (repeated 5 times for 50 models); OBS= observations; K-fold model = predictions based on k-fold model applied to validation data-set; thr = threshold, e.g. thr of .50 refers trainedmodel prediction on the validation data-set at the default threshold of .50.; FN = false negative; FP = false positive; TN = true negative; TP = true positive*

TABLE | 14b: Decision tree performance metrics, early PD/SWEDD

Threshold	Model	Subsampling	Resampling	Performance metrics: validation data-set				
				ROC AUC (95% CI, Delong)	*ACC*	*Kappa*	*SN*	*SP*
.50	Tree	SMOTE	10-fold, 5 repeats	.743 (95% CI: 0.617-0.869)	.798	.343	.341	.945
.486	Tree	SMOTE	10-fold, 5 repeats	.743 (95% CI: 0.617-0.869)	.798	.343	.667	.816

K-fold model predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC AUC = receiver operating characteristic area under the curve; SP = specificity; SN = sensitivity; Kappa = Cohen's Kappa; RF = random forest; Resampling = 10-fold, 5 repeats resampling of the model tuning parameter(s), whereby the optimal hyper-parameter setting was determined by the AUC; SMOTE = synthetic minority oversampling technique; Tree = decision tree.

The AUC metric of .743 (95% CI: 0.617-0.867, DeLong), indicated an approximate 74% chance the model would correctly distinguish between early PD and SWEDD. The AUC with the optimized classification threshold (.486) is provided in Fig 16.

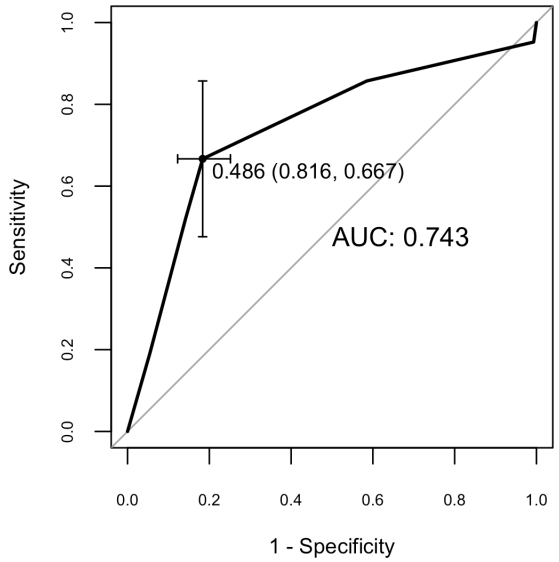


FIG 16 | *Decision tree AUC plot:* *specificity .816; sensitivity .667. Error bars reflect variation in specificity and sensitivity threshold*

5.2 Logistic regression, GAM, early PD/SWEDD

This section reviews the logistic regression model (GLM) and general additive model GAM results; feature elimination, model assumption diagnostics, and tabulated GLM results are provided. As in the early PD/control analysis, the logistic regression GAM was used here both to supplement the logistic regression GLM and as a distinct predictive classification model. The GLM results will be reviewed first followed by the GAM results.

As in the early PD versus control analysis, the logistic GLM model feature elimination used stepwise (backward) regression employing AIC to select predictors of greatest import to classification from a SMOTE subsampled training data-set. **Table 15** lists the 6 predictors ranked in descending order of importance that made the greatest contribution to the GLM model. **Table 15** ranks predictors on a 0-100% scale (M. Kuhn, 2019, March, 3), with the predictor of least importance being attributed zero. The predictors are also ranked by coefficient z-scores - the coefficient z-values of the model with all 6 predictors included. As explained below, one of the predictors was converted to quartiles, which increased the number of model variables to 8. This complied with the minimum of 5 variables per predictor (Vittinghoff & McCulloch, 2007), given there were 44 cases for

each class (*N*= 88). However, the variable gender was under represented, with each class having 14 females and 30 males.

The variable years of education (EDYRS in Table 15) was converted from an integer to a categorical variable comprised of 4 levels (EDYRS Q1-Q4). As occurred in the early PD/control GLM analysis but with a different variable, the variable years of education violated linearity of the logit (Box & Tidwell, 1962). Scatterplots (with smoothers) of all predictor values against the logit is provided in Fig S1-17. Remedially, the years of education variable was converted to (categorical) quartiles and (*k*-1, = 3) dummy variables. The reference was the lowest quartile, and the dummy variables were the lower quartile (EDYRS Q2), the upper quartile (EDYRS Q3), and those with years of education exceeding the upper quartile (EDYRS Q4). Of note, while 17 early PD exceeded the upper quartile of 18 or more years of education, just 7 SWEDD exceeded 18 years of education. The SWEDD cohort had significantly fewer years of education relative to the early PD cohort (see Table 2).

TABLE 15 | Predictor importance, logistic regression, early PD/SWEDD

Predictors	*Scaled 0-100%*	*Coefficient z-values*
Upsit-rev	100	3.90
Age	73.90	3.15
RBDQ	69.62	3.01
EDYRS Q4	67.94	2.96
EDYRS Q3	46.18	2.32
Sex	44.81	2.28
Depression	34.31	1.97
EDYRS Q2	0.00	.971

Coefficient z-scores = absolute value of coefficient z-scores; Depression = Geriatric depression scale; Upsit-rev= (UPSIT) University of Pennsylvania Smell Identification Test, reverse-scaled; Scaled = predictor importance to classification on a 0-100% scale; EDYRS Q2 = years of education below the lower quartile (25%); EDYRS Q3 = years of education below the upper quartile (75%); EDYRS Q4 = years of education above Q3; RBDQ = rapid eye movement behavior disorder questionnaire

The logistic regression analysis specifics are provided in **Table 16**. Coefficients ($\widehat{\boldsymbol{\beta}}$), standard errors ($\boldsymbol{SE}\ \widehat{\widehat{\boldsymbol{\beta}}}$), z-values and odds ratios, standardized predictor coefficients as well as rescaled coefficients (i.e. with the

score range binarized) (see 2.6 Statistics) are included. As noted previously, standardized and rescaled coefficients have been recommended (Cohen et al., 2003) to lend an unbiased perspective to differentially scaled predicators. Here, olfactory function based on the University of Pennsylvania Smell Test (UPSIT; reverse-scaled in the current work [Upsit-rev]), has 0-40 scale, gender, by contrast, is binary (0,1). Coefficients rescaled to a binary scale were (Likert-type variables) Upsit-rev, rapid eye movement behaviour disorder, and the geriatric depression scale. Standardized predictor coefficients are provided for all predictors (except for the categorical variable years of education).

TABLE 16 | Logistic regression early PD/SWEDD

						Standardized		***Rescaled***	
Variable	$\widehat{\boldsymbol{\beta}}(\boldsymbol{SE})$	*z*	***95% CI***	***Odds***	***95% CI (Odds)***	$\widehat{\boldsymbol{\beta}}$*	***Odds****	$\widehat{\boldsymbol{b}}$	***p***
Intercept	-0.8(3.43)	-2.40							
Upsit-rev	-0.36(0.09)	-3.90	(-.57, -.21)	.70	(.59, .84)	-2.79	0.06	-14.25	< .0001***
Age	0.13(0.04)	3.13	(.06, .22)	1.14	(1.05, 1.23)	1.34	3.81	-	= .002**
RBDQ	0.49(.16)	3.01	(.21, .86)	1.63	(1.19, 2.25)	1.50	4.50	6.39	= .003**
EDYRS Q4	-4.36(1.47)	-2.96	(-7.68, -1.82)	.01	(.00, .23)	-	-	-	= 003**
EDYRS Q3	-2.90(1.25)	-2.32	(-5.68, -0.65)	.05	(.00, .64)	-	-	-	= .020*
Sex (male)	2.54(1.11)	2.28	(.57, 5.06)	2.69	(1.43, 112.37)	1.18	3.26	-	= .022*
Depression	0.64(0.32)	1.97	(.07, 1.38)	1.90	(1.01, 3.58)	0.97	2.64	1.77	= .048*
EDYRS Q2	-1.22(1.26)	-0.97	(-3.88, 1.17)	.29	(.02, .3.48)	-	-	-	= .332

N = 88. EDYRS Q2= lower quartile for number of education years; EDYRS Q3 = upper quartile for number of education years; EDYRS Q4= number of education years of those exceeding the upper quartile. RBDQ = rapid eye movement behaviour disorder questionnaire; Upsit-rev = University of Pennsylvania Smell Identification Test (reverse-scaled); $\widehat{\boldsymbol{\beta}}$** = standardized coefficient;* $\widehat{\boldsymbol{b}}$ *= rescaled coefficient (scaled variables have same 0 to 1 range); McFadden pseudo* R^2 *(*pR^2*) = .56; Odds = odds ratio; Odds* = standardized odds ratio; Residual deviance: 54.09 on 79 df; AIC= 72.09. P-values are rounded 3 decimal places; other values are rounded two decimal places.*

A likelihood ratio test indicated the model was a significantly improved classifier relative to the null (no-predictor) model, $\chi^2(8) = 67.9$, $p = < .000$. The McFadden pseudo R^2 value of .56 suggests the model provided a good fit to the data and hence good predictive power to explain classification (Domencich & McFadden, 1996)(https://eml.berkeley.edu/~mcfadden/travel.html). The Hosmer-Lemeshow Goodness of Fit test was insignificant, $X^2(8) = 1.59$, $p = .95$, indicating there was not a significant difference between observed and predicted values. This suggests the model was an effective classifier.

It warrants reiterating again that olfaction was reverse-scaled (Upsit-rev), making higher values indicative of reduced olfactory acuity (e.g. in **Table 2** the SWEDD group has a much lower olfactory impairment score compared to early PD). With greater precision and stability accorded to variables with narrow confidence intervals (Poole, 2001), olfaction, age and RDBQ were the predictors of note. Interpretation of the results will be

confined to the results for olfaction, RBDQ, and years of education variable (EDYRS Q4), the latter representing those exceeding the upper quartile in years of education. Holding other predictors constant, a one standard deviation increase in reduced olfactory acuity explained a 2.79 standard deviation decrease in the log odds of SWEDD, a decrease that significantly differed from zero, $z = -3.90$, $p < .0001$. With other predictors partialed out, the odds of reduced olfactory acuity in SWEDD was 6% of the odds of reduced olfactory acuity in early PD. Holding other predictors constant, a one standard deviation increase in RBDQ explains a 1.50 standard deviation increase in the log odds of SWEDD, an increase that significantly differed from zero, $z = 3.01$ $p = .003$. With other predictors partialed out, a one standard deviation increase in RBDQ multiplies the odds of SWEDD by 4.50. Finally, holding other predictors constant, and relative to the base lowest level of years of education (first quartile), each year increase in education beyond the upper quartile reduced the log odds of SWEDD by 4.36, a reduction that significantly differed from zero, $z = -2.96$, $p = .003$. Relative to the lowest level of years of education, the odds of SWEDD having greater than 18 years of education (above the third quartile here) was 1% of the odds of early PD having greater than 18 years of education.

Effects plots for three highest ranked logistic regression GLM predictors, olfaction, age and RBDQ, are provided in Fig 17, where the predicted probability (y-axis) of SWEDD group membership decreases with higher olfactory deficit (increasing x-axis values) but increases with higher RBDQ score and age.

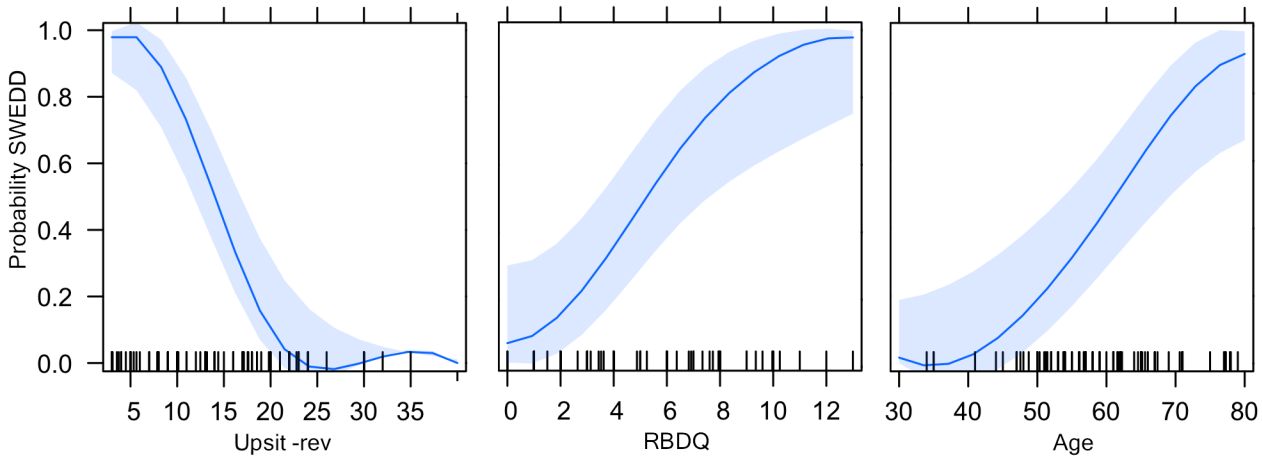


FIG 17 | *Effects plots*: *predicted probability (y-axis) of SWEDD group membership decreased with higher olfaction deficit (left) but increased with higher RBDQ scores and age (right). Upsit-rev = University of Pennsylvania Smell Inventory Test; (reverse-scaled); RBDQ = rapid eye movement behaviour disorder questionnaire; SWEDD = scan without evidence of dopaminergic deficit*

In a final check on collinearity, the variance inflation factor for logistic regression determined an absence of concern for explanatory variable collinearity, and the Durbin-Watson test (2.23, $p = .27$) indicated errors were not correlated. Case-wise diagnostics included index plots of leverage, discrepancy and Cook's Distance. There was no indication of cases sufficiently extreme to bias model results (see Supporting information 1 case-wise plots Figs S1 18-21). Approximately 3% of cases (3/88) had leverage beyond the cut-off (.31) and there were no cases that exceeded a discrepancy of $> \pm 2.5$. None of the cases attained a Cook's Distance of 1 (the highest Cooks Distance value was .42, and second highest was .12).

The GLM included conversion of years of education variable (which violated the assumption of linearity of logit) to quartiles. Because this conversion can result in a loss of some data (Bennette & Vickers, 2012; Greenland, 1995; Royston, 2000) the GAM model helped to further convey the relation of the years of education variable to the probability of SWEDD classification. A GAM model treating the years of education variable as a thin plate spline had significantly lower residual deviance (46.04) compared to the GLM (54.09), χ^2 (1.12)= 8.05, $p = .006$. GAM pseudo R^2 (62.3%) was also higher relative to logistic regression pseudo R^2 (56.0%). The GAM model output is provided in Supporting information II, Table, S2-2. As noted previously, the GAM model used the same features selected by the logistic GLM stepwise procedure. The GAM predictor ranking (not provided in Table 15) in descending order of importance was as follows based on *p*-values: Upsit-rev, RBDQ, age, years of education, gender, and depression. With respect to GAM diagnostics, the estimated degrees of freedom (edf) did not closely approach predictor *k*-value. The optimal k-value of the lone smoothed term, years of education, was $k= 7$. The residuals in the (quantile-quantile) qq.gam plot (see Supporting information II, Fig S2-3) adhered quite closely to the diagonal, suggesting the data set quantiles approximated the theoretical quantiles. A scatterplot of years of education (smoothed using a thin plat spline) is provided in Supporting information II, Fig S2-4.

The logistic regression (GLM) model AUC was .938 (95% CI: 0.8877-0.9533, DeLong; sensitivity .864; specificity .841). Single predictor AUC values were calculated only with logistic regression. Single logistic regression predictor models using the top 3 predictors had poor to moderate AUC values: age AUC .543; RBDQ AUC .701; hyposmia AUC .760. Based on modified (Robin et al., 2011) bootstrap ($n = 2000$) test (Hanley & McNeil, 1982) for correlated ROC curves, the 6-predictor integrated model (actually an 8-predictor model

counting the education quartiles modeled) had significantly improved classification compared to the top single predictor hyposmia model, $D = 3.62$, $p < .001$. The GAM model ROC AUC was .955 (95% CI: 0.916-0.9941, DeLong; sensitivity .886; specificity .909).

The GLM and GAM model predictive results are summarized in **Tables 17a** and **17b**. The **Table 17a** confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. **Table 17b** provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided. The optimized GLM threshold was .504; only marginally different from the default .50 threshold. As such, confusion matrices and performance metrics for the default and optimized threshold were identical. The optimized GAM threshold was .437. Optimized model thresholds were selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a). AUC graphs for the GLM and GAM cross-validated models are provided in Fig 18.

TABLE 17a | Logistic regression GLM, GAM confusion matrices, early PD/SWEDD

	Class	OBS	Predicted; thr = 0.5		Predicted; opt. thr.	
GLM	PD	147	TN = 111	FN = 7	TN = 111	FN = 7
GLM	SWEDD	21	FP = 36	TP = 14	FP = 36	TP = 14
GAM	PD	147	TN = 117	FN = 8	TN = 112	FN = 6
GAM	SWEDD	21	FP = 30	TP = 13	FP = 35	TP = 15

N= 168 (validation data-set); OBS= observations; GAM = general additive model; GLM = generalized linear model (logistic regression here); PD = early PD; opt. = optimized threshold (Youden Index); thr = threshold, e.g. thr of .50 refers to trained model class predictions on validation data-set at the default threshold of .50.; SWEDD= scan without evidence of dopaminergic deficit; FN = false negative; FP = false positive; TN = true negative; TP = true positive

TABLE 17b | Logistic regression GLM, GAM performance metrics, early PD/SWEDD

Threshold	Model	Subsampling	Resampling	Performance metrics: validation data-set			
				ROC (95% CI, Delong)	*ACC*	*Kappa*	*SN*	*SP*
.50	GLM	SMOTE	None	.779 (95% CI: 0.677-0.880)	.744	.265	.667	.755
.504	GLM	SMOTE	None	.779 (95% CI: 0.677-0.880)	.744	.265	.667	.755
.50	GAM	SMOTE	None	.787 (95% CI: 0.689-0.886)	.774	.286	.619	.796
.437	GAM	SMOTE	None	.787 (95% CI: 0.689-0.886)	.756	.299	.714	.762

Predicted results on validation data-set. ACC = accuracy; GAM = general additive model; GLM = generalized linear model (logistic regression here); ROC AUC = receiver operating characteristic area under the curve; SP = specificity; SN = sensitivity; Kappa = Cohen's Kappa; ROC = receiver opera-ting characteristic; SMOTE = synthetic minority oversampling technique.

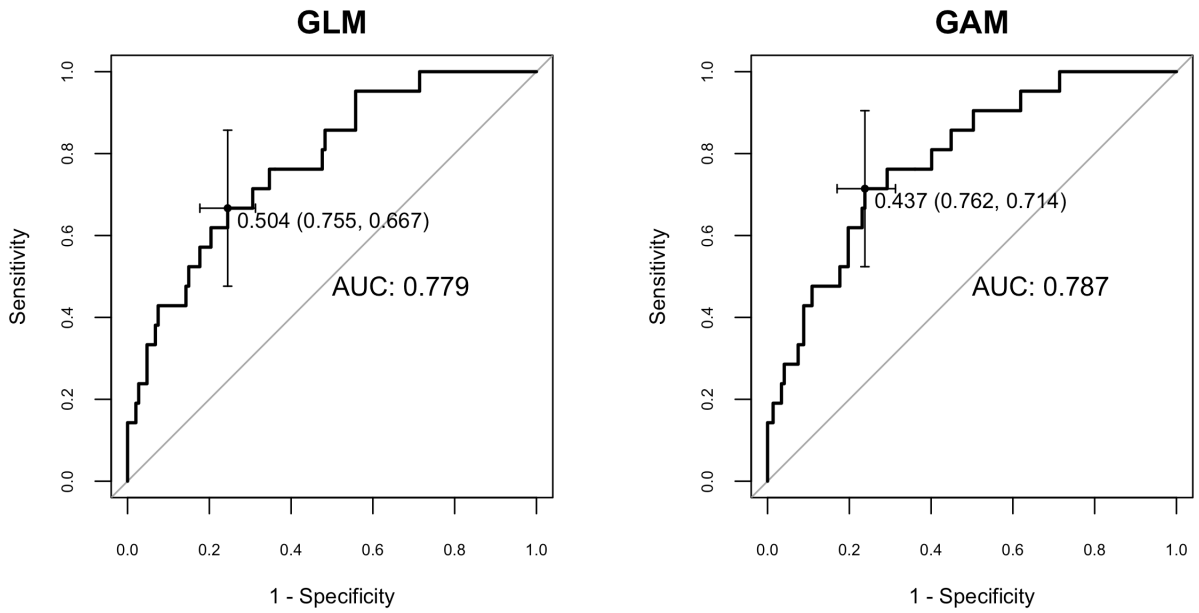


FIG 18 | *Logistic regression GLM, GAM AUC plots:* *GLM ROC AUC: .779 (95% CI: 0.677-0.880, DeLong), specificity .667, sensitivity .755; GAM ROC AUC .787 (95% CI: 0.689-0.886, DeLong), specificity .762, sensitivity .714; error bars reflect variation in specificity and sensitivity threshold*

5.3 Random forest classification early PD/SWEDD

This section reviews the random forest (early PD/SWEDD) analysis; feature elimination, and model specific findings are followed by model predictive performance results from application of the model to the validation data set. From the original 14 predictors of the SMOTE augmented training data (44 early PD, 30 male; 44 SWEDD, 30 male), 12 features were retained. The mean Gini decrease ranking of predictor importance (built-in to random forest) and the generic variable of importance predictor ranking (M. Kuhn, 2019, March, 3) are listed in **Table 18**; the ranking of predictors was identical up to and including the forth-ranked predictor age but differed thereafter. In **Table 18** under Scaled 0-100%, predictors are ranked on a 0-100% scale (M. Kuhn, 2019, March, 3) based on contribution to the model; under the heading Mean Gini decrease, predictors are ranked by the extent to which they contributed to the mean decrease in the Gini index (the decrease in impurity) across all trees.

The optimal m_{try} tuning parameter (selected by AUC criterion during 10-fold, 5 repetitions) resampling, was $m_{try} = 1$. The optimal model had an ntree (the number of trees grown) value of 3000 trees and a node size of 1. Fig 19 includes two random forest out-of-bag (OOB) error plots. Plot (A) shows the OOB error rate of the

original model, without subsampling. Plot (B) shows the OOB error rate for the SMOTE based model, which yielded substantive improvement in OOB error rate, particularly for SWEDD.

TABLE 18 | Predictor importance, random forest, early PD/SWEDD

Predictors	Scaled 0-100%	Predictors	Mean Gini decrease
Upsit-rev	100	Upsit-rev	5.27
ESS	74.16	ESS	4.41
RBDQ	60.37	RBDQ	3.89
Age	45.28	Age	3.61
Anxiety	44,16	$A\beta_{1-42}$	3.60
MoCA	39.46	Anxiety	3.55
EDYRS	35.19	EDYRS	3.36
CNST	30.02	CSF α-syn	3.19
$A\beta_{1-42}$	20.43	MoCA	3.12
Depression	19.53	Depression	2.76
CSF α-syn	9.59	CNST	1.60
Sex	0.00	Sex	.960

$A\beta_{1-42}$ = beta amyloid $_{1-42}$; Anxiety = State trait anxiety inventory; CNST = Constipation based on the UPDRS I; CSF α-syn = CSF alpha synuclein; ESS = Epworth daytime sleepiness scale; MeanGini decrease = variable importance to average Gini (impurity) decrease; MoCA = Montreal cognitive assessment; Upsit-rev= (UPSIT) University of Pennsylvania Smell Identification Test, reverse-scaled; RBDQ = rapid eye movement behaviour disorder questionnaire; Scaled = predictor importance to classification ona 0-100% scale.

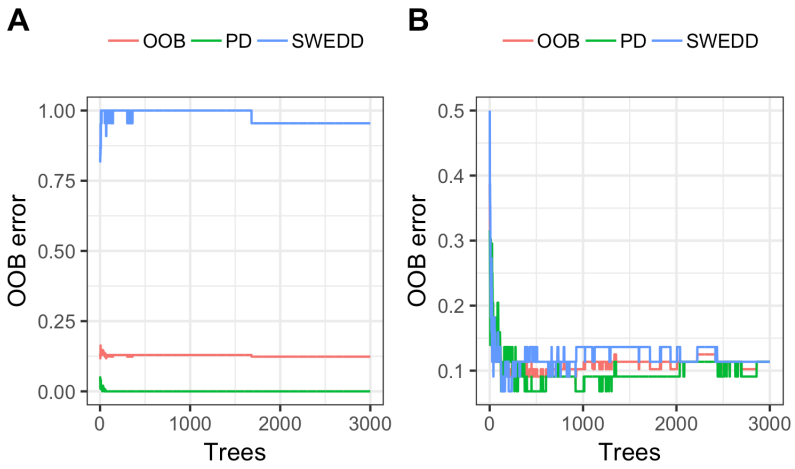


FIG 19 | *Out-of-bag (OOB) error rates compared: (A) Out-of-bag (OOB) error rate original non-subsampled data; (B) OOB rate Synthetic Minority Over-sampling Technique (SMOTE) data*

The model ROC AUC was 1.00 (95% CI: 1-1, DeLong; sensitivity 1.00; specificity 1.00), which is inevitable by design when the optimal nodesize =1, as occurred here. Model (cross-validated) predictive results are summarized in Tables 19a and 19b. **The Table 19a** confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. **Table 19b** provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided. The optimized threshold (.461) was selected by the pROC package (Robin et al., 2011) where the optimal threshold was the point nearest to the top-left point of perfect sensitivity/specificity. The AUC graph with the optimized threshold is provided in Fig 20.

TABLE 19a | Random forest confusion matrices, early PD/SWEDD

Class	OBS	Predicted, K-fold model; thr = 0.5		Predicted;, K-fold model, opt. thr. = .461	
PD	142	TN = 116	FN = 7	TN = 106	FN = 4
SWEDD	21	FP = 31	TP = 14	FP = 41	TP = 17

N= 168 (validation data-set); OBS= observations; PD = early PD; opt. = optimized threshold (closest t op-left); thr = threshold, e.g. thr of .50 refers to trained model class predictions on validation data-set at the default threshold of .50.; SWEDD= scan without evidence of dopaminergic deficit; FN = false negative; FP = false positive; TN = true negative; TP = true positive; K-fold, where k =10 (repeated 5 times for 50 models)

TABLE 19b | Random forest performance metrics, early PD/SWEDD

Threshold	Model	Subsampling	Resampling	Performance metrics: validation data-set			
				ROC (95% CI, Delong)	*ACC*	*Kappa*	*SN*	*SP*
.50	RF	SMOTE	10-fold x 5	.822 (95% CI: 0.746-0.899)	.774	.306	.667	.789
.461	RF	SMOTE	10-fold x 5	.822 (95% CI: 0.746-0.899)	.732	.302	.809	.721

K-fold model predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = sensitivity; RF = random forest; ROC AUC = receiver operating characteristic area under the curve; SMOTE = Synthetic Minority Oversampling Technique

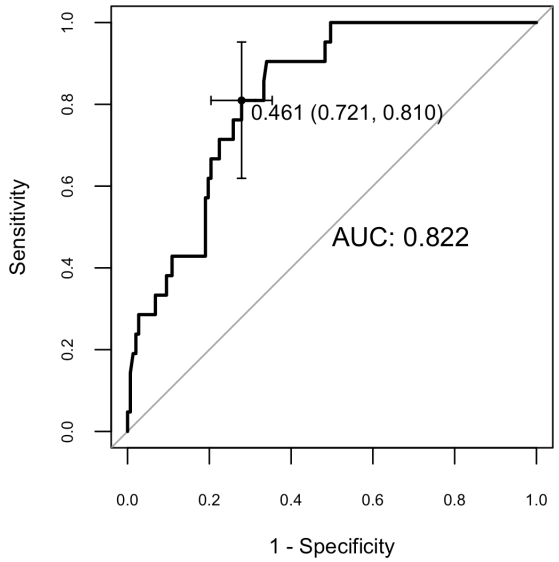


FIG 20 | *Random forest AUC plot*: *AUC = .822; sensitivity = .810, specificity = .721. Optimized threshold based on point closest to top-left plot region of perfect sensitivity/specificity; error bars reflect variation in specificity and sensitivity threshold*

5.4 XGBoost classification early PD/SWEDD

This section reviews the XGBoost (early PD/SWEDD) classification analysis. Feature selection results are followed by review of XGBoost optimized tuning parameters determined by k-fold (10 folds, 5 repeats) resampling. Subsequently, the predictive performance results on the validation data-set are reviewed. The SMOTE subsampled training data used had 44 instances of each class (30 male in each class).

From the original 14 predictors of the SMOTE augmented training data, the10 predictors were determined to be of greatest import to classification; these variables are listed in **Table 20**. The predictors listed under Scaled 0-100% in **Table 20** are ranked on 0-100% scale (M. Kuhn, 2019, March, 3) based on contribution to the model; predictors listed under Gain are ranked by extent of contribution to the XGBoost Gain index. Gain, an XGBoost built-in index of predictor importance, measures a feature's contribution to each tree in the model and hence overall prediction; higher values are better. Note that the built-in (gain-based) and generic based ranking of predictor importance was the same for the first 2 predictors and lowest ranked 8th-10th predictors but differed for other predictors.

TABLE 20 | Predictor importance, XGBoost

Predictors	Scaled 0-100%	Predictors	Gain
Upsit-rev	100.00	Upsit-rev	0.2241
ESS	91.55	ESS	0.1926
EDYRS	63.75	Age	0.1698
MoCA	42.37	$A\beta_{1-42}$	0.0805
RBDQ	37.98	EDYRS	0.0791
$A\beta_{1-42}$	28.08	MoCA	0.0719
Age	23.71	RBDQ	0.0684
Depression	12.57	Depression	0.0565
Anxiety	12.23	Anxiety	0.0383
CNST	0.00	CNST	0.0189

$A\beta_{1-42}$ = beta amyloid $_{1-42}$; Anxiety = trait anxiety from the State-trait anxiety inventory; EDYS = years of eduction; ESS = Epworth daytime sleepiness scale;Gain = variable importance to prediction, higher is better;Depression = Geriatric depression scale; MoCA = Montreal Cognitive Assessment test; Upsit-rev = reverse-scaled University of Pennsylvania Smell Identification; RBDQ = rapid eye movement behaviour disorder questionnaire; Scaled 0-100 %= predictor importance to classification on a 0-100% scale; Gain = predictors ranked by the extent of contribution to the XGBoost gain index

As in the XGBoost early PD/control analysis, a range of tuning parameter settings were assessed during resampling (10 fold, repeated 5 times), using the AUC as the optimal parameter selection criterion. Tuning parameter guidelines adopted [156] were endorsed by the creators of XGBoost (T. Chen, Guestrin, C., 2016). The optimal tuning parameters found were as follows: eta 0.2; nrounds 500; max_depth 4; min_child_weight 1; colsample_bytree 0.8; gamma 3; subsample setting of .75. For more information regarding XGBoost hyper-parameters see *Supporting information III* (*The models: Logistic regression, general additive, decision tree, random forest and XGBoost)*.

The model AUC was .993 (95% CI: 0.9927-1; sensitivity .977; specificity .954). Model predictive results are summarized in **Tables 21a** and **21b**. The **Table 21a** confusion matrices specify the classification frequency predicted by the model applied to the validation data-set. **Table 21b** provides performance metrics for the model applied to the validation data-set. Results at the default .50 and optimal classification thresholds are provided. The optimized model threshold (.542) was selected by the pROC package (Robin et al., 2011) utilizing a modified (Perkins & Schisterman, 2006) version of the Youden Index (Youden, 1950a).

TABLE 21a | XGBoost confusion matrices, early PD/SWEDD

Class	OBS	Predicted, K-fold model; thr = 0.5		Predicted;, K-fold model, opt. thr. = .542	
PD	142	TN = 104	FN = 2	TN = 110	FN = 2
SWEDD	21	FP = 43	TP = 19	FP = 37	TP = 19

N= 168 (validation data-set); OBS= observations; PD = early PD; opt. = optimized threshold based a modified version of the Youden Index; thr = threshold, e.g. thr of .50 refers to trained model class predictions on validation data-set at the default threshold of .50.; SWEDD= scans without evidence of dopamine deficit; FN = false negative; FP = false positive; TN = true negative; TP = true positive; K-fold, where k =10 (repeated 5 times for 50 models)

TABLE 21b | XGBoost performance metrics, early PD/SWEDD

Threshold	Model	Subsampling method	Resampling	Performance metrics: validation data-set				
				ROC (95% CI, DeLong)	*ACC*	*Kappa*	*SN*	*SP*
.50	XGB	SMOTE	10-fold x 5	.863 (0.777-0.948)	.732	.333	.905	.707
.542	XGB	SMOTE	10-fold x 5	.863 (0.777-0.948)	.768	.381	.905	.748

Note: *K-fold model predicted results on validation data-set. ACC = accuracy; Kappa = Cohen's Kappa; ROC = receiver operating characteristic; SP = specificity; SN = sensitivity; RF = random forest; SMOTE = Synthetic Minority Oversampling Technique*

The AUC cross-validated statistic (on the validation data set) was .863 (95% CI: 0.7774-0.9479, DeLong), indicating an approximate 86% chance the model would correctly distinguish between SWEDD and early PD. This was the highest AUC achieved in the early PD/SWEDD classification analyses. The ROC AUC of the XGBoost model applied to the validation data-set is plotted in Fig 21.

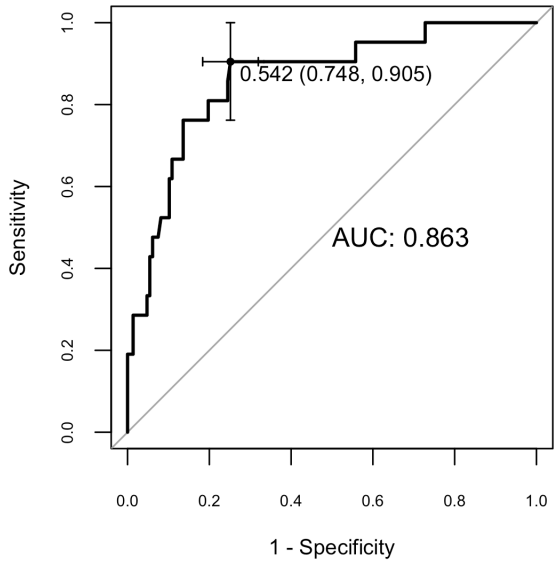


FIG 21 | *XGBoost AUC plot**: AUC .863 (0.777-0.948, DeLong; sensitivity*

.905 specificity .748; error bars reflect variation in specificity and sensitivity; threshold .542 determinded by the Youden Index (Youden 1950) modified by (Perkins and Schisterman 2006) and selected by the pROC package (Robin et al 2011)

5.5 Early PD/SWEDD: model Prediction and Long-Term Diagnosis

As in section *4.5,* PPMI longitudinal data (12-36 months) was again used to assess the extent that model class predictions matched longitudinal diagnosis; all models were trained on baseline data. While section *4.5* was largely concerned with early PD versus control model usefulness to identify SWEDD to PD conversions, this section describes early PD versus SWEDD model longitudinal accuracy in terms of sensitivity and specificity measures. As specified previously, the training data consisted of SMOTE resampled data (*N*= 88: 44 early PD, 30 male; 44 SWEDD, 30 male); the validation test data set was also previously specified (*N*= 168, 147 early PD, 92 male; 21 SWEDD, 11 male). The top performing early PD/SWEDD classifiers were random forest and XGBoost (see **Table 22** and Fig 25).

Longitudinal curated diagnoses available for the two top performing early PD/SWEDD classifiers, random forest and XGBoost, demonstrated, again, as in section *4.5*, the largely non-PD diversity of pathologies that constitute the SWEDD category. Here, however, the main interest was estimating long-term model accuracy or fidelity in terms of estimated model long-term sensitivity and specificity: long-term sensitivity was defined as the percentage of model classified non-PD SWEDD matching curated long-term diagnosis; long-term specificity was defined as the percentage of model classified PD that matched the curated long-term diagnosis.

At its optimal cutoff (.461), random forest long-term sensitivity amounted to 12/16 (75%) correctly predicted non-PD SWEDD cases that matched the 12-24 month curated diagnoses records available. The mean putamen DAT scan value averaged across all (non-PD) SWEDD at 12-36 months was 2.06 (*SD* = .50). Random forest model-approximated long-term specificity amounted to 92/128 (71.87%) cases (true negatives) correctly classified by the model as PD at 12-24 months. The mean putamen DAT scan value of these PD confirmed cases, averaged across all PD cases at 12-36 months, was .69 (*SD* = .27). For XGBoost, and at its optimal cutoff (.542), long-term model sensitivity to non-PD SWEDD amounted to 13/16 (81.25%) class predictions that correctly matched 12-24 month curated diagnoses. The mean putamen DAT scan value averaged across all (non-PD) SWEDD at 12-36 months was 2.0 (*SD* = .50). XGBoost long-term specificity amounted to 97/128 (75.78%) cases

(model true negatives) correctly classified by as PD. The mean putamen DAT scan value of the PD classified segment, averaged across all PD diagnosed cases at 12-36 months, was .70 (*SD* = .31).

Chapter Six: *Performance summary*

This section summarizes model analyses predictive findings. A review of model feature selection is followed model predictive results. With respect to features of importance, the latter were selected by model specific (e.g. stepwise feature elimination using AIC for regression) or built-in (e.g. mean Gini decrease in random forest) model-based feature elimination, with the final feature selection narrowed by features resulting in the highest model AUC. Figs 22 and 23 are based on a generic feature of importance ranking (M. Kuhn, 2019, March, 3) used throughout this work. Fig 22 pertains to the early PD/control model ranking of features of importance to classification; Fig 23 pertains to early PD/SWEDD model ranking of feature importance to classification. The scaling of model built-in feature ranking varies widely among model-types. The generic ranking of predictors (M. Kuhn, 2019, March, 3) was used in Figs 22 and 23 because it conveniently ranks the import of predictors of the different model types on the same common 0-100% scale. The model specific built-in feature ranking has been specified and juxtaposed with generic feature ranking in predictor importance tables throughout this work. Note, that while there was predictor-ranking agreement between generic and built-in ranking methods for top ranked features, lesser-ranked features often varied in rank to some extent between built-in and generic ranking methods. See the individual model predictor importance tables in Chapters 4 and 5 for details. The GAM features of importance are not shown in Figs 22 or 23. The GAM model used the same predictors as the logistic regression model (see GLM in Figs 22-23) but the rank of features, with the exception of hyposmia, differed. In descending order of importance, the rank of features to GAM early PD/control classification was hyposmia, RBDQ, age, pTau, constipation, and MoCA. In descending order of importance, the rank of features to GAM early PD/SWEDD classification was Upsit-rev, RDBQ, age, years of education, gender, and depression.

Overall, hyposmia was the top ranked predictor of importance and RBDQ and MoCA were, on average, of high rank for all models in both the early PD/control and early PD/SWEDD classification analyses. Otherwise there was variation in model feature selection and feature ranking between classification analyses, including variation within the same model types across the separate early PD/control and early PD/SWEDD analyses.

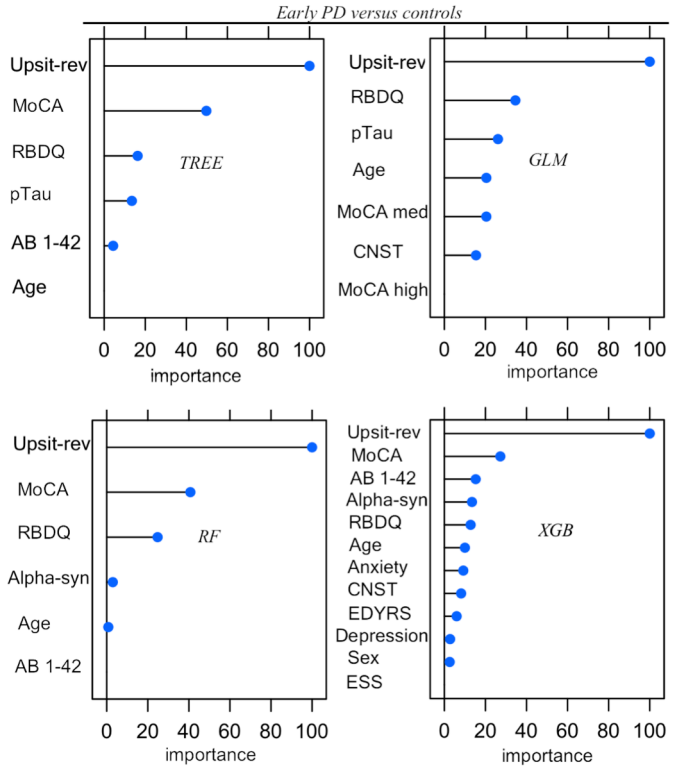


FIG 22 | Predictor importance ranking Early PD/controls. GLM, logistic regression; TREE, decision tree; RF, random forest; XGB, Extreme gradient boosting; Anxiety,state trait anxiety; AB 1-42, beta amyloid 1−42; Alpha-syn, CSF a-synuclein; CNST, constipation; ESS, Epworth sleepiness scale; Depression, Geriatric depressionscale; MoCA, Montreal Cognitive Assessment; MoCA med, MoCA median; Upsit-rev, reverse scaled Upsit-rev, (UPSIT) University of Pennsylvania Smell Identification Test, reverse-scaled; pTau, CSF phosphorylated Tau; RBDQ, rapid eye movement behavior disorder questionnaire; Sex, gender. Predictors ranked on a 0 to 100%scale of importance to classification.

The model performance results (from models applied to test validation data unseen by models during training) are summarized in **Table 22**. The AUC, accuracy, Kappa statistic, sensitivity and specificity outcomes are listed. **Table 22** superscript notation reflects tree-model k-fold resampling of tuning parameters and if subsampling (i.e. synthetic minority oversampling technique [SMOTE]) was used. SMOTE was used only to augment the early PD/SWEDD training data.

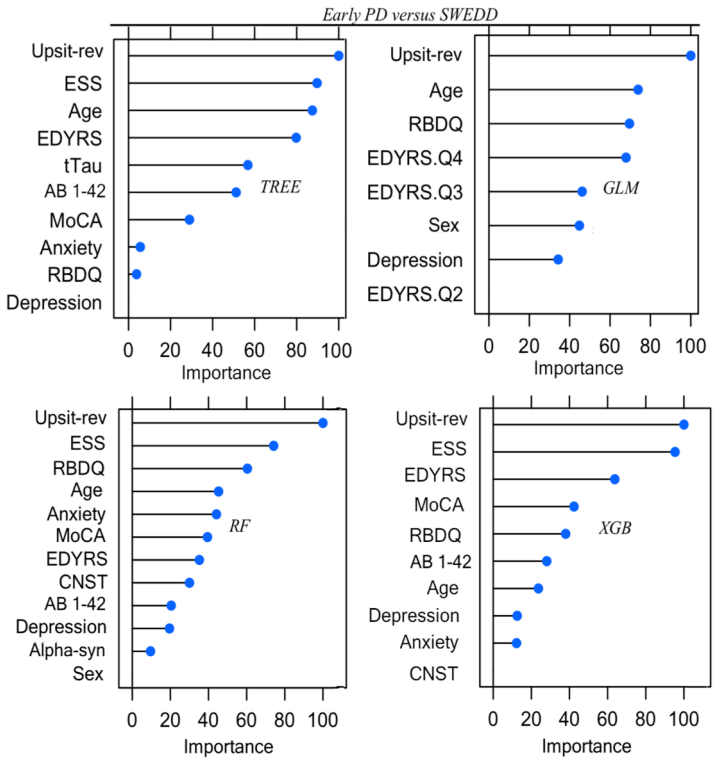


FIG 23 | Predictor importance ranking Early PD/SWEDD. GLM, logistic regression; TREE, decision tree; RF, random forest; XGB, Extreme gradient boosting; Anxiety, state trait anxiety; AB 1-42, beta amyloid 1–42; Alpha-syn, CSF a-synuclein; CNST, constipation; Depression, Geriatric depression scale; EDYRS Q2, lower quartile for number of education years; EDYRS Q3, upper quartile for number of education years; EDYRS Q4, number of education years of those exceeding the upper quartile; ESS, Epworth sleepiness scale; GDS, Geriatric depression scale; MoCA, Montreal Cognitive Assessment; Upsit-rev, University of Penn-sylvania Smell Identification Test, reverse-scaled;RBDQ, rapid eye movement behavior disorder questionnaire; Predictors ranked on a 0 to100% scale of importance.

TABLE 22 | Performance summary

Models	Train			Test (cross-validated)					
Metric	AUC (95% CI)	SN	SP	Opt.Thr.	AUC (95% CI)	ACC	Kappa	SN	SP
EARLY PD vs. HC									
GLM	.920 (0.888-0.953)	.934	.797	.462	.907 (0.849-0.964)	.898	.764	.909	.872
GAM	.946 (0.922-0.970)	.923	.850	.534	**.928** (0.878-0.978)	.898	.768	.898	.897
Tree [a]	.872 (0.831-0.913)	.857	.879	.586	.860 (0.799-0.922)	.842	.659	.818	.897
RF [a]	.999 (0.999-1.00)	.990	1.00	.534	.913 (0.858-0.968)	.898	.764	.909	.872
XGB [a]	.958 (0.937-0.979)	.898	. 901	.660	.923 (0.875-0.972)	.882	.736	.875	.897
EARLY PD vs. SWEDD									
GLM [b]	.938 (0.863-0.972)	.909	.841	.504	.779 (0.677-0.880)	.744	.265	.667	.755
GAM [b]	.955 (0.916-0.994)	.886	.909	.437	.787 (0.689-0.886)	.756	.299	.714	.762
Tree [a, b]	.932 (0.894-0.971)	.891	.900	.486	.743 (0.617-0.869)	.798	.343	.667	.816
RF [a, b]	1.00 (1.00-1.00)	1.00	1.00	.461	.822 (0.746-0.899)	.732	.302	.809	.721
XGB [a, b]	.997 (0.993-1.00)	.977	.954	.542	**.863** (0.777-0.948)	.768	.381	.905	.748

Superscript a, 10-fold, 5 repeats resampling of the model tuning parameter(s), whereby the optimal hyper-parameter setting was determined by the AUC; ACC, accuracy; superscript b, synthetic minority oversampling technique (SMOTE); AUC, receiver operating characteristic area under the curve; CI, DeLong confidence interval; Kappa, Cohen's Kappa; SP, specificity; SN, sensitivity; GAM, general additive model; GLM, logistic regression generalized linear model; RF, random forest; Tree, decision tree; XGBoost, Extreme gradient boosting; Thr, threshold; Bold model names, highest cross-validated AUC.

Reviewing the early PD/control results first, all models achieved an early PD/control classification AUC of > 80%. Three pairwise AUC tests were conducted, which was sufficient to gain a comparative perspective on model early PD/control cross-validated (CV) AUC scores. Using Bonferroni correction for family-wise error, and rounding down, α was set at .01 (.05/3= .0167) to control for family-wise error. A modified (Robin et al., 2011) bootstrap (n = 2000) test (Hanley & McNeil, 1982) was used for AUC pairwise comparisons of correlated ROC curves. First, the GLM AUC (.907) was higher than the decision tree AUC (.860), but this difference was not significant, D = 1.80, p = 0.071. The random forest AUC (.913) was higher than the decision tree AUC (.860), a difference that proved to be significantly different, D = 2.83, p = 0.005. It follows then, that the XGBoost AUC (.916) and GAM AUC (.928), which exceeded the random forest AUC (.913), were also significantly higher than the tree model AUC. Further, the GAM AUC (.928) was higher than the GLM AUC (.907) but this difference that was not significant, D = 1.83, p = 0.072. Accordingly, it follows that the random forest AUC (.913) and XGBoost AUC (.916), both higher than the GLM AUC (.907), also did not significantly differ from the GAM model AUC (.928). In short, all models except the GLM had significantly higher AUC values relative to the decision tree model, but there was not a significant AUC difference among the GAM, GLM, random forest and XGBoost models. The GLM and random forest optimized confusion matrices were identical, yet their AUCs are not identical. This underlines that while the AUC is a summary measure of specificity and sensitivity across a range of thresholds (Fawcett, 2006), the matrix values amount to a threshold point on the ROC curve through which different curves can pass. The GAM and XGBoost model were the highest performing early PD/control classifiers (see **Table 22**). The AUC of both models is graphed in Fig 24.

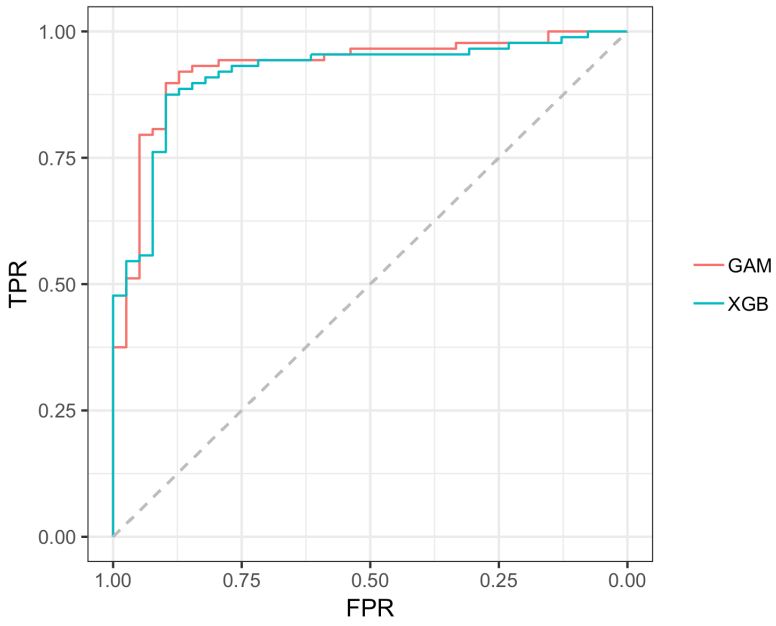


FIG 24 | *GAM, XGBoost AUC plots:* *GAM AUC = .928; XGB AUC = .916;GAM = general additive model; XGB = Extreme gradient boosting model; TPR = true positive rate (sensitivity); FPR = the False positive rate (1 – specificity).*

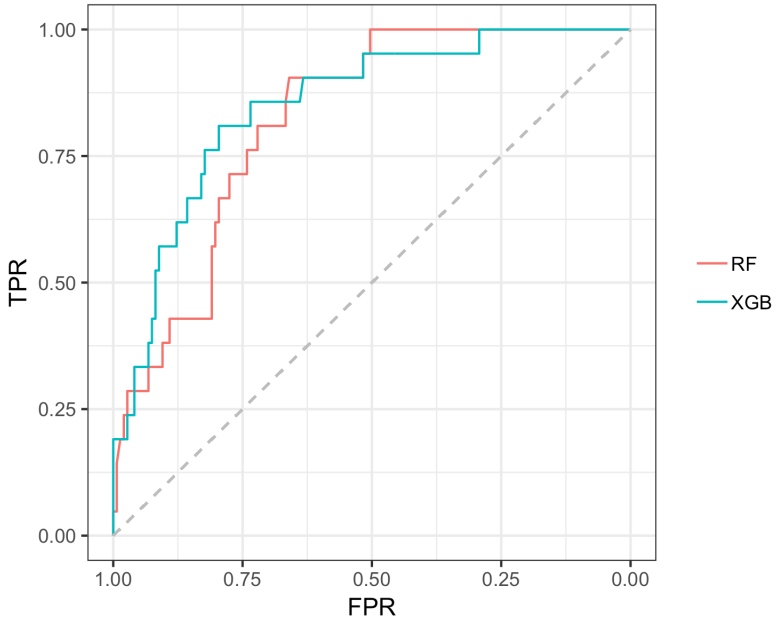


FIG 25 | *Random forest, XGBoost AUC plots:* *RF AUC = .822 (sensitivity .809; Specificity .721); XGB AUC = .863 (sensitivity .905; specificity .743);* *RF = random forest; XGB = Extreme gradient boosting model; TPR =true positive rate (sensitivity); FPR = the False positive rate (1 – specificity).*

In the early PD/SWEDD CV results, model classification performance metrics were lower relative to those in the early PD/control analysis. XGBoost and random forest were the most efficient early PD/SWEDD classifiers (see Table 22). A modified (Robin et al., 2011) bootstrap (n = 2000) test (Hanley & McNeil, 1982) was used for an AUC pairwise caparison of correlated ROC curves. The XGBoost AUC (.863), the highest AUC outcome in the early PD/SWEDD analysis, was not significantly different from the decision tree model AUC (.743), D = 1.89, p = 0.06. Among the five models, the decision tree, as in the early PD/control analysis, had the lowest AUC. Accordingly, it follows that the other models (GLM, GAM and random forest), which had higher AUC values relative to the decision tree but lower AUC values compared to XGBoost, were also not significantly different from either the decision tree or XGBoost AUC outcomes. The early PD/SWEDD AUC of the random forest and XGBoost models is provided in Fig 25.

Model prediction and long-term diagnosis

The GAM and XGBoost models were the best performing (highest AUCs) early PD/control classifiers, and these models were applied to the SWEDD/control (baseline) validation data to assess usefulness of the models to predict SWEDD to PD conversion. The GAM model discriminated SWEDD or PD-like cases from controls with AUC of .863 (optimal cut-off .389: sensitivity = .814; specificity = .846). In the available longitudinal PPMI (curated 12-24) month data, 12/38 (32%) GAM model SWEDD cases predicted to be PD-like were diagnosed as idiopathic PD. However the majority of GAM model PD-like cases, 26/38 (68%), were not diagnosed as idiopathic PD longitudinally but rather were a mix of almost a dozen non-PD pathologies and apparently normal cases. The 24-month time point diagnosis data also indicated that none of the 33 model predicted controls converted to PD. The XGBoost model applied to (baseline) SWEDD/control validation data discriminated SWEDD (or PD-like) cases from controls with an AUC of .831 (optimal cut-off .378: sensitivity = .837; specificity = .769). In the available longitudinal PPMI curated 12-24 month data, 13/43 (30%) SWEDD predicted by the XGBoost model to be PD-like had converted from SWEDD to idiopathic PD. The majority of XGBoost model PD-like data instances, 34/43 (79%), were not classified in the 12-36 month longitudinal data as idiopathic PD but diagnosed as a collective of diverse disorders similar to those noted for the GAM model. As with the GAM model, none of the model predicted controls was diagnosed at 12-24 months as PD.

The top performing early PD/SWEDD classifiers were random forest and XGBoost . With respect to early PD/SWEDD long-term sensitivity (percentage of model classification non-PD SWEDD matching curated long-term diagnosis) and specificity (percentage model classification of PD matching curated long-term diagnosis) of these classifiers, long-term sensitivity amounted to 12/16 (75%) correctly predicted non-PD SWEDD cases that matched the 12-24 month curated diagnoses records available. Random forest model-approximated long-term specificity amounted to 92/128 (71.87%) cases (true negatives) correctly classified by the model as PD at 12-24 months. XGBoost long-term model sensitivity to non-PD SWEDD amounted to 13/16 (81.25%) class predictions that correctly matched 12-24 month curated diagnoses. XGBoost long-term specificity amounted to 97/128 (75.78%) cases (model true negatives) correctly classified by as PD.

It warrants note that the results reported in this dissertation (and in the published version of this work online) were based on the Decision tree, GAM, Random forest, and XGBoost algorithm versions available at the original inception date of this research (January 2020). Newer updates of these algorithms, as well as the caret package (M. Kuhn, 2019, March, 3), were applied April 10, 2020. The newer software versions, with the original and unaltered model feature sets and hyper-parameter settings, resulted in the same overall outcome pattern but with consistently higher performance across all metrics (i.e. AUC, sensitivity, specificity, Kappa and accuracy). Model results with updated software (now as of October 1, 2020) is available online (https://www.researchgate.net/publication/342165274_Supporting_information_VIpdf) and from the author (cslfalcon@gmail.com). The much older GLM (logistic regression model) algorithm has not been altered for some time and the results for this model remain unaltered.

Chapter Seven: Discussion

Unique to the current work was the particular set of five classifiers used and the dual early PD/control and early PD/SWEDD analyses approach adopted. There is never a guarantee that one model type will outperform another (Wolpert, 1996). Comparing several classifiers (here five), determined that of these five, the optimal model differed for the early PD/control relative to the early PD/SWEDD classification analyses. The GAM was top performing early PD/control classifier, and the XGBoost model was the top performing early PD/SWEDD classifier. Overall, the XGBoost model had the most consistent classification performance, achieving the second highest performance in the early PD/control analysis and the highest early PD/SWEDD outcome (see **Table 22** for details). Moreover, as made apparent in Figs 22-23, an advantage of conducting the dual classification analysis was that particular features were revealed to have differential importance to early PD/SWEDD versus early PD/control discrimination. Notably, Epworth sleepiness scale (ESS) and years of education figured as prominent features of import to early PD/SWEDD classification but were of little to no consequence to early PD/control classification. Across both analyses hyposmia (based on the University of Pennsylvania Smell Test – reverse scaled) was inevitably the single most important feature to model classification. Rapid eye-movement behaviour disorder questionnaire (RBDQ) and the Montreal Cognitive Assessment (MoCA) were the next most common features of relatively high rank importance to classification for all models in both analyses. Biomarkers CSF α-synuclein and pTau were features of greater importance to early PD/control classification than to early PD/SWEDD classification, and age assumed greater importance in early PD/SWEDD classification.

Predictive model results from recent studies also using Parkinson's Progressive Markers Initiative (PPMI) data incorporating either clinical and genetic risk (Nalls et al., 2015) or clinical variables and biologics (Yu et al., 2018) achieved high early PD versus control AUC scores without including imaging (DAT scan) predictors: .923 (sensitivity 83.4%; specificity 90.3%) (Nalls et al., 2015); .927 (sensitivity 89.7%, specificity 80.4 %) (Yu et al., 2018). Predictive models in the current work, also incorporating PPMI clinical and biologics (not genetic risk) data, achieved similarly high model AUC scores discriminating early PD versus control. The two top performing models were the GAM and XGBoost classifiers. The GAM model had an AUC of .946 (sensitivity 91.3%; specificity 80.7%) and the XGBoost model an AUC of .958 (sensitivity 93.7%, specificity 83.5%). The Nalls et

al. (Nalls et al., 2015) and Yu et al. (Yu et al., 2018) studies both used logistic regression. Comparing apples to apples, the early PD/control logistic regression model in the current work had an AUC of .920 (sensitivity 93.4%, specificity 79.7%). The marginally lower logistic regression model AUC obtained was due in part to a smaller training set: the PPMI data was partitioned, using random stratification, into train and test sets while the above referenced studies used all the early PD/control PPMI data to train models and validated models in different cohort data sets. In addition, the stringent data filtering for only complete cases across 14 variables resulted in further data instance reduction. Moreover, while hyposmia (based on the UPSIT scale) and age were common features selected by the logistic regression stepwise process among the Nalls et al., Yu et al. and the current work, feature elimination in the current work otherwise resulted in a different final set of predictors. The Nalls study, which included genetic risk, not part of this study, used five features: hyposmia, genetic risk, family history, age and gender. The Yu et al. study used hyposmia, age, CSF α-synuclein and gender. The logistic model stepwise (using AIC) feature elimination procedure in the current work determined hyposmia, rapid eye-movement behaviour disorder questionnaire (RBDQ), pTau, age, MoCA and constipation as the most important features to early PD/control classification. Gender in the current study, and with respect to the SWEDD test data in particular, the random stratified split of SWEDD data into train and test sets left gender under represented. However, in the early PD/SWEDD as well as the early PD/control classifications gender was a feature of low or no importance. Further, Yu et al. commented that their model's outcome was similar whether or not gender was included.

Note, to avoid collinearity exceeding r_s .75, neither pTau jointly with tTau, nor α-synuclein with either pTau or tTau were used concurrently in the same model. The correlations of α-synuclein and pTau and α-synuclein and tTau were r_s =.82 and r_s =.81 respectively. The correlation between pTau and tTau was r_s = .98. The relatively low collinearity cut off of .75, sensitive to pairwise correlations (M Kuhn, 2013), was adopted to prioritize unbiased feature selection and classification consistently for all models (Carolin Strobl et al., 2008; C. Strobl, Boulesteix, Zeileis, & Hothorn, 2007; Tolosi & Lengauer, 2011) (see 2.6 Statistical analyses).

Another recent study also using PPMI data reported features important to early PD/control classification including hyposmia, RBDQ, CSF α-synuclein, pTau, tTau, and notably DAT scan values (Prashanth, Roy,

Mandal, & Ghosh, 2017). The DAT scan values (striatal binding ratios for the left and right putamen) made the greatest contribution to model performance and further heightened AUC scores to > .98 for all five models in the latter study. However, like the Nalls and Yu analyses, the current work did not include SPECT DAT scan values as predictors. SPECT imaging is not always accessible and a single scan can be costly (typically over $1800 in the US). Further, as noted in the introduction (Marek et al., 2014), SPECT DAT scan imaging results can be misleading; misdiagnosis of PD for SWEDD category members is an issue.

Importantly, the current work's early PD versus control classification models applied to validation data, unseen by models during training (the cross-validated [CV] outcome), achieved high classification accuracy. The highest performing model, the GAM, had a CV AUC of .928 (at the optimal threshold of .534, sensitivity 89.9%, specificity 89.7%). The second highest CV AUC from the XGBoost model was .923 (at the optimal threshold of .660, sensitivity 87.5%, specificity 89.7%). Overall, and as hypothesized, the non-motor clinical and biologic features used achieved > .80% AUC classification accuracy across all models (decision tree, logistic regression, general linear, random forest, and XGBoost), a level of constancy supporting the validity and reliability of these features to differentiate early stage PD pathology, from age-matched normal healthy subjects, with relatively high classification accuracy. This consistency, across all models adds to the growing body of research (Kang et al., 2013; Marek et al., 2018; Nalls et al., 2015; Prashanth et al., 2017; Yu et al., 2018) demonstrating the usefulness of non-motor clinical and biomarker features in early stage PD discrimination. In addition, the AUC of all models, with the exception of decision tree, were very similar. The decision tree model had a significantly lower AUC (.860) compared to the other four model types. In practice, the logistic regression model (GLM) offered the best blend of simplicity, parsimony of predictors and performance. In addition, as a parametric model, it had the benefit of quantifying predictor contribution to the model (e.g. coefficients). But in the event of a non-linear feature–logit relation, exemplified by the MoCA feature in the early PD/control classification, the GAM, random forest or XGBoost models may be more appropriate.

It was also posited that outcome of the second classification analysis involving early PD versus SWEDD discrimination would be less definitive and typified by lower AUC results for all models. This also proved true. Results for both early PD/control and early PD/SWEDD classification analyses are provided in **Table 22**. The

discrepancy of model performance between early PD/control and early PD/SWEDD classification is, at least in part, due to the wide range of disorders encompassed by the SWEDD category. The diversity of clinical entities within the SWEDD category, reported in other research (Erro et al., 2016; Nicastro, Garibotto, Badoud, & Burkhard, 2016; Stoessl, 2010; Wyman-Chick, Martin, Minar, & Schroeder, 2016), was evident in current study longitudinal findings where SWEDD proved to be largely a mix of almost a dozen clinical entities (Alzheimer's disease case, polyneuropathy, lateral sclerosis, essential tremor, psychogenic illness, apparently normal etc.). The heterogeneity of the SWEDD category, in general adds complexity and confusion to PD pathology differentiation. Indeed, removal of the term or category SWEDD, as currently conceptualized, has been recommended (Erro et al., 2016; Nicastro et al., 2016).

Developing a model (s) to disentangle non-PD SWEDD cases from actual cases of early PD pathology was one objective of this research. The two top performing early PD/SWEDD classification models, XGBoost (AUC .863, sensitivity .905, specificity .748) and random forest (AUC .822, sensitivity .809, specificity .721), were able to discriminate SWEDD from early PD with moderate sensitivity. From the random forest results 12/16 (75%) SWEDD non-PD predicted cases matched the SWEDD non-PD case diagnoses in 12-24 available records. From the XGBoost results 13/16 (81.25%) SWEDD non-PD predicted cases matched the SWEDD non-PD case diagnoses in 12-24 available records. The random forest long-term specificity to discriminate early PD, correctly, amounted to 92/128 (71.87%); XGBoost long-term specificity amounted to 97/128 (75.78%). These results suggest that either model could be useful to help differentiate non-PD SWEDD category patients from those with actual incipient PD pathology.

In a brief review of descriptive statistics (including the UPDRS III and DAT scan putamen and caudate values not used in models) more severe hyposmia, rapid eye-movement behaviour disorder (questionnaire-based [RBDQ]), anxiety traits, and constipation occurred in early PD compared to healthy controls. Montreal cognitive assessment (MoCA) scores as well as caudate and putamen DAT scan values were also lower for the early PD cohort compared to controls. As might be expected, UPDRS III scores were also much higher, typical of PD, compared to controls. Comparing SWEDD to early PD, hyposmia was more severe for early PD, Epworth sleepiness scale (ESS) was higher (worse) for SWEDD, and there were fewer years of education for SWEDD. T-

tests (Wilcox, 1990) demonstrated significant early PD/control and early PD/SWEDD differences of the afore mentioned variables (see **Tables 1-2**), findings consistent with prior research (Marek et al., 2018). Also in agreement with other research (Kang et al., 2016; Marek et al., 2018), there were significantly reduced cerebral spinal fluid biomarker values of $A\beta_{1\text{-}42}$, α-synuclein, pTau and tTau in early PD compared to healthy controls. In addition, there was significantly increased $A\beta_{1\text{-}42}$ in SWEDD compared to early PD, and while this agreed with findings from Marek et al (Marek et al., 2018), contrary to the latter study we did not find significantly differing α-synuclein between early PD and SWEDD (see **Table 2**). Finally, in agreement with other research (Kang et al., 2013; Llorens et al., 2016) moderate to high correlations (r_s >.75) were found among CSF α-synuclein, pTau and tTau.

The median age in the PPMI data used in the current work was 62, which along with other PPMI demographic data (education, ethnicity, and gender) is consistent with clinical trial demographics (Fahn et al., 2004; Investigators, 2007; Kordower et al., 2013). Age, though, poses the single highest risk factor for neurodegenerative diseases such as idiopathic PD (Lin & Beal, 2006). Further, as there is an age related increase in hyposmia (Hummel, Futschik, Frasnelli, & Huttenbrink, 2003) for instance, age is a variable with increasing confounding potential in more elderly cohorts (e.g. 85+). In the current work data, age was positively correlated with hyposmia (higher age was associated with more severe hyposmia), though the correlation was well under the .75 limitation set (r_s = .22, p < .001). While age in age-matched groups can be controlled for in the statistical sense (by inclusion in the model), classifiers trained on younger cohorts would, in general, help to isolate the importance of features (and their underlying physical properties) to detection of PD neurodegeneration independent of age.

As reported in the first paragraph of this discussion, Epworth sleepiness scale (ESS) in particular but also years of education were important features to early PD/SWEDD but not to early PD/control discrimination (see Table 22). Both features also significantly differed between early PD and SWEDD but not early PD and controls (see Tables 1-2). These findings, in concert with other PPMI data research (Jain, Park, & Comer, 2015; Marek et al., 2018), warrant further investigation. Is the difference in years of education, fewer years of education in SWEDD, just specific to the particular SWEDD cohort used? If not, how does more extensive education relate to

PD pathology? With respect ESS, an even more important early PD/SWEDD group differentiator, a question to be probed is how does dozing-off in certain situations (ESS measures dozing-off rather than fatigue) relate differently to the non-PD clinical entities of SWEDD compared to early PD?

It warrants note that hyposmia, the main model driver here as in other research (Nalls et al., 2015; Yu et al., 2018) is not specific to PD pathology (Abele, Riet, Hummel, Klockgether, & Wullner, 2003; Doty, Reyes, & Gregor, 1987; Doty et al., 1984; Galvez, Diaz, Hernandez-Castillo, Campos-Romo, & Fernandez-Ruiz, 2014; Schofield, Ebrahimi, Jones, Bateman, & Murray, 2012). It has been suggested that CSF α-synuclein, which is synucleinopathy-specific, may increase specificity for PD-type pathology when combined with other features (e.g. hyposmia) in a model (Yu et al., 2018). But if so, it is critical to first determine the species of α-synuclein specific to PD pathology. While variations of glia-to-glia, glia-to-neuron and neuron-to-neuron spread of α-synuclein are likely (Frost & Diamond, 2010; Jucker & Walker, 2013; Lee et al., 2010), the form of this toxic misfolded protein to be targeted for diagnostic and prognostic purposes remains to be established (e.g. α-synuclein monomers, oligomers [intermediate compounds between monomers and polymers] or the misfolded fibril?) A recent study demonstrated that α-synuclein fibrils injected into the mouse brain acted as agents recruiting monomeric endogenous α-synuclein and induced PD indicators including loss of substantia nigra pars compacta and striatal dopamine terminals as well as dysfunctional motor behaviour (Froula et al., 2019). However, the root cause may involve an oligomer pre-fibril state. For reviews on this subject see Mead et al (Meade, Fairlie, & Mason, 2019) and Xu and Pu (Xu & Pu, 2016).

Conclusion

This work undertook a unique investigation with dual early PD/control and early PD/SWEDD classification analyses. The overarching objective was to further assess the utility of non-motor clinical and biomarker features to discriminate early stage PD pathology. In agreement with other research, hyposmia, RBD, and CSF biomarkers distinguished early PD versus controls with high classification performance. Indeed, as a testament to the classification efficacy of features used, the current work demonstrated that five different models could achieve > .80% AUC cross-validated classification accuracy without imaging or motor predictors. Relative to early PD/control results, early PD/SWEDD model classification performance was lower (for all models), the optimally

performing model-type differed, and, with the exception of hyposmia, there was variation in feature selection or rank of features by models for early PD/control compared to early PD/SWEDD analyses- informative findings that justified the dual analysis approach. Moreover, data at 12-36 months from baseline indicated longitudinal model sensitivity of up to about 81% to distinguish non-PD SWEDD cases from PD pathology. The model may be useful to screen SWEDD category patients with actual incipient PD pathology from those non-PD SWEDD category patients. Without such screening, the heterogeneity in the SWEDD category will diminish the capacity of future models to detect and discriminate PD pathology.

Limitations

After filtering for only completed cases, and only cases meeting the screening criteria (i.e. exclusion of samples with > 200 ng/ml hemoglobin levels), data sets were quite small, particularly the SWEDD validation data-set. However, with respect to PD/SWEDD training data, SMOTE subsampling augmented training data instances while also balancing groups. It should be mentioned that ratios of biomarkers or (biomarkers and clinical variables) were not included in the current work, and would have added more depth to evaluations. In addition, a multinomial rather than binomial approach could have been used. However, in respect to the latter, most current classification research has used the binomial approach, which facilitates comparison among study outcomes.

Conflict of interest

None.

Funding
This research was supported by an NSERC Discovery grant to Joseph DeSouza. There were no financial interests relating to this research.

References

Abbott, R. D., Petrovitch, H., White, L. R., Masaki, K. H., Tanner, C. M., Curb, J. D., . . . Ross, G. W. (2001). Frequency of bowel movements and the future risk of Parkinson's disease. *Neurology, 57*(3), 456-462. doi:10.1212/wnl.57.3.456

Abbott, R. D., Ross, G. W., White, L. R., Tanner, C. M., Nelson, J. S., & Petrovitch, H. (2005). Excessive daytime sleepiness and the future risk of Parkinson's disease. *Movement Disorders, 20*, S101-S101.

Abele, M., Riet, A., Hummel, T., Klockgether, T., & Wullner, U. (2003). Olfactory dysfunction in cerebellar ataxia and multiple system atrophy. *Journal of Neurology, 250*(12), 1453-1455. doi:10.1007/s00415-003-0248-4

Adler, C. H. (2011). Premotor Symptoms and Early Diagnosis of Parkinson's Disease. *International Journal of Neuroscience, 121*, 3-8. doi:10.3109/00207454.2011.620192

Ahlskog, J. E. (2005). Challenging conventional wisdom: The etiologic role of dopamine oxidative stress in Parkinson's disease. *Movement Disorders, 20*(3), 271-282. doi:10.1002/mds.20362

Akaike, H. (1974). NEW LOOK AT STATISTICAL-MODEL IDENTIFICATION. *Ieee Transactions on Automatic Control, AC19*(6), 716-723. doi:10.1109/tac.1974.1100705

Albin, R. L., Young, A. B., & Penney, J. B. (1989). THE FUNCTIONAL-ANATOMY OF BASAL GANGLIA DISORDERS. *Trends in Neurosciences, 12*(10), 366-375. doi:10.1016/0166-2236(89)90074-x

Alexander, G. E., Delong, M. R., & Strick, P. L. (1986). PARALLEL ORGANIZATION OF FUNCTIONALLY SEGREGATED CIRCUITS LINKING BASAL GANGLIA AND CORTEX. *Annual Review of Neuroscience, 9*, 357-381. doi:10.1146/annurev.ne.09.030186.002041

Amit, Y., & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation, 9*(7), 1545-1588. doi:10.1162/neco.1997.9.7.1545

ATLAS. (2014). Machine Learning Wins the Higgs Challenge. Retrieved from https://atlas.cern/updates/atlas-news/machine-learning-wins-higgs-challenge

Attems, J., & Jellinger, K. A. (2008). The dorsal motor nucleus of the vagus is not an obligatory trigger site of Parkinson's disease. *Neuropathology and Applied Neurobiology, 34*(4), 466-467. doi:10.1111/j.1365-2990.2008.00937.x

Bagley, S. C., White, H., & Golomb, B. A. (2001). Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain. *Journal of Clinical Epidemiology, 54*(10), 979-985. doi:10.1016/s0895-4356(01)00372-9

Baker, E., & Forshing, L. (2020). *Neuroanatomy, Vagal Nerve Nuclei* Retrieved from https://www.ncbi.nlm.nih.gov/books/NBK545209/

Batista, G. P., R.; Monard, M.;. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter,, 6*(1), 20-29.

Belsely, D. A., Kuh, E., & Welsch, R. E. (1980). Regression diagnostics. *Wiley and Sons, New York, USA*.

Benabid, A. L., Pollak, P., Louveau, A., Henry, S., & Derougemont, J. (1987). COMBINED (THALAMOTOMY AND STIMULATION) STEREOTACTIC SURGERY OF THE VIM THALAMIC NUCLEUS FOR BILATERAL PARKINSON DISEASE. *Applied Neurophysiology, 50*(1-6), 344-346.

Benarroch, E. E. (2009). The locus ceruleus norepinephrine system Functional organization and potential clinical significance. *Neurology, 73*(20), 1699-1704. doi:10.1212/WNL.0b013e3181c2937c

Bengio, Y., & Grandvalet, Y. (2004). No unbiased estimator of the variance of K-fold cross-validation. *Journal of Machine Learning Research, 5*, 1089-1105.

Bennette, C., & Vickers, A. (2012). Against quantiles: categorization of continuous variables in epidemiologic research, and its discontents. *Bmc Medical Research Methodology, 12*. doi:10.1186/1471-2288-12-21

Berendse, H. W., Booij, J., Francot, C., Bergmans, P. L. M., Hijman, R., Stoof, J. C., & Wolters, E. C. (2001). Subclinical dopaminergic dysfunction in asymptomatic Parkinson's disease patients' relatives with a decreased sense of smell. *Annals of Neurology, 50*(1), 34-41. doi:10.1002/ana.1049

Blennow, K., Hampel, H., Weiner, M., & Zetterberg, H. (2010). Cerebrospinal fluid and plasma biomarkers in Alzheimer disease. *Nature Reviews Neurology, 6*(3), 131-144. doi:10.1038/nrneurol.2010.4

Bostan, A. C., Dum, R. P., & Strick, P. L. (2013). Cerebellar networks with the cerebral cortex and basal ganglia. *Trends in Cognitive Sciences, 17*(5), 241-254. doi:10.1016/j.tics.2013.03.003

Box, G. E. P., & Tidwell, P. W. (1962). Transformation of the Independent Variables (Vol. 4, pp. 531-550). http://www.jstor.org/stable/1266288: Technometrics.

Braak, H., & Braak, E. (1991). Demonstration of amyloid deposits and neurofibrillary changes in whole brain sections. *Brain Pathol, 1*(3), 213-216. doi:10.1111/j.1750-3639.1991.tb00661.x

Braak, H., de Vos, R. A. I., Bohl, J., & Del Tredici, K. (2006). Gastric alpha-synuclein immunoreactive inclusions in Meissner's and Auerbach's plexuses in cases staged for Parkinson's disease-related brain pathology. *Neuroscience Letters, 396*(1), 67-72. doi:10.1016/j.neulet.2005.11.012

Braak, H., Del Tredici, K., Rub, U., de Vos, R. A. I., Steur, E., & Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging, 24*(2), 197-211. doi:10.1016/s0197-4580(02)00065-9

Braak, H., Ghebremedhin, E., Rub, U., Bratzke, H., & Del Tredici, K. (2004). Stages in the development of Parkinson's disease-related pathology. *Cell and Tissue Research, 318*(1), 121-134. doi:10.1007/s00441-004-0956-9

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123-140. doi:10.1007/bf00058655

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. doi:10.1023/a:1010933404324

Breiman, L., Freidman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont: Wadsworth.

Breit, S., Kupferberg, A., Rogler, G., & Hasler, G. (2018). Vagus Nerve as Modulator of the Brain-Gut Axis in Psychiatric and Inflammatory Disorders. *Frontiers in Psychiatry, 9*. doi:10.3389/fpsyt.2018.00044

Brichta, L., & Greengard, P. (2014). Molecular determinants of selective dopaminergic vulnerability in Parkinson's disease: an update. *Frontiers in Neuroanatomy, 8*. doi:10.3389/fnana.2014.00152

Brooks, P. L., & Peever, J. H. (2012). Identification of the Transmitter and Receptor Mechanisms Responsible for REM Sleep Paralysis. *Journal of Neuroscience, 32*(29), 9785-9795. doi:10.1523/jneurosci.0482-12.2012

Brownell, A. L., Jenkins, B. G., Elmaleh, D. R., Deacon, T. W., Spealman, R. D., & Isacson, O. (1998). Combined PET/MRS brain studies show dynamic and long-term physiological changes in a primate model of Parkinson disease. *Nature Medicine, 4*(11), 1308-1312. doi:10.1038/3300

Brownlee, J. (2015, August 19). 8 tactics to combat imbalanced classes in your machine learning dataset. Retrieved from https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/

Burez, J. V. d. P., D. (2009). Handling Class Imbalance In Customer Churn Prediction. *Expert Systems with Applications, 36*(3), 4626-4636.

Calabresi, P., Picconi, B., Tozzi, A., Ghiglieri, V., & Di Filippo, M. (2014). Direct and indirect pathways of basal ganglia: a critical reappraisal. *Nature Neuroscience, 17*(8), 1022-1030. doi:10.1038/nn.3743

Cameron, K. N., Solis, E., Ruchala, I., De Felice, L. J., & Eltit, J. M. (2015). Amphetamine activates calcium channels through dopamine transporter-mediated depolarization. *Cell Calcium, 58*(5), 457-466. doi:10.1016/j.ceca.2015.06.013

Castaldi, P. J., Dahabreh, I. J., & Ioannidis, J. P. A. (2011). An empirical assessment of validation practices for molecular classifiers. *Briefings in Bioinformatics, 12*(3), 189-202. doi:10.1093/bib/bbq073

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321-357. doi:10.1613/jair.953

Chen, L. X. (2020). Overview of clinical prediction models. *Annals of Translational Medicine, 8*(4). doi:10.21037/atm.2019.11.121

Chen, T., & Guestrin, C. (2020). XGBoost. *XGBoost Parameters*. Retrieved from https://xgboost.readthedocs.io/en/latest/parameter.html - parameters-for-tree-booster

Chen, T., Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *SIGKDD*, 785-794. doi: http://dx.doi.org/10.1145/2939672.2939785

Chen, V. C., Lin, T. Y., Yeh, D. C., Chai, J. W., & Weng, J. C. (2019). Predicting chemo-brain in breast cancer survivors using multiple MRI features and machine-learning. *Magn Reson Med, 81*(5), 3304-3313. doi:10.1002/mrm.27607

Classification, S. E. (2019). Classification. Retrieved from https://stats.stackexchange.com/questions/266267/should-one-be-concerned-about-multi-collinearity-when-using-non-linear-models

Cohen, J. (1960). A Coefficient of Agreement for Nominal Data. *Educational and Psychological Measurement, 20*, 37-46.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied Multiple Regression Correlation Analysis for the Behavioral Sciences* (3r ed. ed.). New Jersey: Lawrence Erlbaum Associates, Inc.

Collins, L., Griffioen, P., Newell, G., & Mellor, A. (2018). The utility of Random Forests for wildfire severity mapping. *Remote Sensing of Environment, 216*, 374-384. doi:10.1016/j.rse.2018.07.005

Cramer, J. S. (2002). The origins of logistic regression. *Tinbergen Institute*, 167-178. doi:doi:10.2139/ssrn.360300

Cummings, J. L., Henchcliffe, C., Schaier, S., Simuni, T., Waxman, A., & Kemp, P. (2011). The role of dopaminergic imaging in patients with symptoms of dopaminergic system neurodegeneration. *Brain, 134*, 3146-3166. doi:10.1093/brain/awr177

Dauer, W., & Przedborski, S. (2003). Parkinson's disease: Mechanisms and models. *Neuron, 39*(6), 889-909. doi:10.1016/s0896-6273(03)00568-3

Delong, E. R., Delong, D. M., & Clarkepearson, D. I. (1988). COMPARING THE AREAS UNDER 2 OR MORE CORRELATED RECEIVER OPERATING CHARACTERISTIC CURVES - A NONPARAMETRIC APPROACH. *Biometrics, 44*(3), 837-845. doi:10.2307/2531595

Delong, M. R. (1971). ACTIVITY OF PALLIDAL NEURONS DURING MOVEMENT. *Journal of Neurophysiology, 34*(3), 414-&.

Delong, M. R. (1990). PRIMATE MODELS OF MOVEMENT-DISORDERS OF BASAL GANGLIA ORIGIN. *Trends in Neurosciences, 13*(7), 281-285. doi:10.1016/0166-2236(90)90110-v

Deurveilher, S., & Semba, K. (2008). Reciprocal connections between the suprachiasmatic nucleus and the midbrain raphe nuclei: A putative role in the circadian control of behavioral states. *Serotonin and Sleep: Molecular, Functional and Clinical Aspects*, 103-131. doi:10.1007/978-3-7643-8561-3_4

Devito, J. L., & Anderson, M. E. (1982). AN AUTORADIOGRAPHIC STUDY OF EFFERENT CONNECTIONS OF THE GLOBUS PALLIDUS IN MACACA-MULATTA. *Experimental Brain Research, 46*(1), 107-117.

Dickson, D. W., Fujishiro, H., DelleDonne, A., Menke, J., Ahmed, Z., Klos, K. J., . . . Ahlskog, J. E. (2008). Evidence that incidental Lewy body disease is pre-symptomatic Parkinson's disease. *Acta Neuropathologica, 115*(4), 437-444. doi:10.1007/s00401-008-0345-7

Dluzen, D. E., & McDermott, J. L. (2000). Gender differences in neurotoxicity of the nigrostriatal dopaminergic system: implications for Parkinson's disease. *The journal of gender-specific medicine : JGSM : the official journal of the Partnership for Women's Health at Columbia, 3*(6), 36-42.

Domencich, T., & McFadden, D. (1996). Urban Travel Demand: A Behavioral Analysis. Retrieved from https://eml.berkeley.edu/~mcfadden/travel.html

Doty, R. L., Deems, D. A., & Stellar, S. (1988). OLFACTORY DYSFUNCTION IN PARKINSONISM - A GENERAL DEFICIT UNRELATED TO NEUROLOGIC SIGNS, DISEASE STAGE, OR DISEASE DURATION. *Neurology, 38*(8), 1237-1244. doi:10.1212/wnl.38.8.1237

Doty, R. L., Reyes, P. F., & Gregor, T. (1987). Presence of both odor identification and detection deficits in Alzheimer's disease. *Brain Res Bull, 18*(5), 597-600. doi:10.1016/0361-9230(87)90129-8

Doty, R. L., Shaman, P., & Dann, M. (1984). DEVELOPMENT OF THE UNIVERSITY-OF-PENNSYLVANIA SMELL IDENTIFICATION TEST - A STANDARDIZED MICROENCAPSULATED TEST OF OLFACTORY FUNCTION. *Physiology & Behavior, 32*(3), 489-502. doi:10.1016/0031-9384(84)90269-5

Durbin, J., & Watson, G. S. (1951). TESTING FOR SERIAL CORRELATION IN LEAST SQUARES REGRESSION .2. *Biometrika, 38*(1-2), 159-178. doi:10.1093/biomet/38.1-2.159

Erro, R., Schneider, S. A., Stamelou, M., Quinn, N. P., & Bhatia, K. P. (2016). What do patients with scans without evidence of dopaminergic deficit (SWEDD) have? New evidence and continuing controversies. *Journal of Neurology Neurosurgery and Psychiatry, 87*(3), 319-323. doi:10.1136/jnnp-2014-310256
Everitt, B., S.; Skrondal, A. (2010). *Cambridge Dictionary of Statistics*: Cambridge University Press.
Ewald, B. (2006). Post hoc choice of cut points introduced bias to diagnostic research. *Journal of Clinical Epidemiology, 59*(8), 798-801. doi:10.1016/j.jclinepi.2005.11.025
Exchange, S. (2017). What's considered a good log loss? Retrieved from https://stats.stackexchange.com/questions/276067/whats-considered-a-good-log-loss
Exchange, S. (2018, August 3). Highly-correlated variables in random forest Retrieved from https://stats.stackexchange.com/questions/141619/wont-highly-correlated-variables-in-random-forest-distort-accuracy-and-feature
Exchange, S. (2019a). gamma parameter in xgboost. Retrieved from https://stats.stackexchange.com/questions/418687/gamma-parameter-in-xgboost
Exchange, S. (2019b). How to interpret the output of XGBoost importance? Retrieved from https://datascience.stackexchange.com/questions/12318/how-to-interpret-the-output-of-xgboost-importance
Exchange, S. (2019c). Multicollinearity. Retrieved from https://stats.stackexchange.com/questions/141619/wont-highly-correlated-variables-in-random-forest-distort-accuracy-and-feature
Exchange, S. (2019, March 1). Classification XGBoost vs Logistic Regression. Retrieved from https://stats.stackexchange.com/questions/394705/classification-xgboost-vs-logistic-regression
Fahn, S., Shoulson, I., Kieburtz, K., Rudolph, A., Lang, A., Olanow, C. W., . . . Parkinson Study, G. (2004). Levodopa and the progression of Parkinson's disease. *New England Journal of Medicine, 351*(24), 2498-2508.
Fasano, A., Romito, L. M., Daniele, A., Piano, C., Zinno, M., Bentivoglio, A. R., & Albanese, A. (2010). Motor and cognitive outcome in patients with Parkinson's disease 8 years after subthalamic implants. *Brain, 133*, 2664-2676. doi:10.1093/brain/awq221
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters, 27*(8), 861-874. doi:10.1016/j.patrec.2005.10.010
Field, A. P. (2012). *Discovering statistics using R*. London: Sage.
Flack, V. F., & Chang, P. C. (1987). FREQUENCY OF SELECTING NOISE VARIABLES IN SUBSET REGRESSION-ANALYSIS - A SIMULATION STUDY. *American Statistician, 41*(1), 84-86. doi:10.2307/2684336
Foody, G. M., Campbell, N. A., Trodd, N. M., & Wood, T. F. (1992). DERIVATION AND APPLICATIONS OF PROBABILISTIC MEASURES OF CLASS MEMBERSHIP FROM THE MAXIMUM-LIKELIHOOD CLASSIFICATION. *Photogrammetric Engineering and Remote Sensing, 58*(9), 1335-1341.
Forno, L. S. (1996). Neuropathology of Parkinson's disease. *J Neuropathol Exp Neurol, 55*(3), 259-272. doi:10.1097/00005072-199603000-00001
Fox, J. W., S. . (2011). *An {R} Companion to Applied Regression* (Vol. Second Edition). Thousand Oaks, CA: Sage.
Freedman, D. A. (1983). A NOTE ON SCREENING REGRESSION EQUATIONS. *American Statistician, 37*(2), 152-155. doi:10.2307/2685877
Freeman, E. A., & Moisen, G. G. (2008). A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Ecological Modelling, 217*(1-2), 48-58. doi:10.1016/j.ecolmodel.2008.05.015
Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189-1232. doi:10.1214/aos/1013203451
Frost, B., & Diamond, M. I. (2010). Prion-like mechanisms in neurodegenerative diseases. *Nature Reviews Neuroscience, 11*(3), 155-159. doi:10.1038/nrn2786

Froula, J. M., Castellana-Cruz, M., Anabtawi, N. M., Camino, J. D., Chen, S. W., Thrasher, D. R., . . . Volpicelli-Daley, L. A. (2019). Defining alpha-synuclein species responsible for Parkinson's disease phenotypes in mice. *Journal of Biological Chemistry, 294*(27), 10392-10406. doi:10.1074/jbc.RA119.007743

Fukunishi, I., Hosokawa, K., & Ozaki, S. (1991). DEPRESSION ANTEDATING THE ONSET OF PARKINSONS-DISEASE. *Japanese Journal of Psychiatry and Neurology, 45*(1), 7-11.

Gaig, C., & Tolosa, E. (2009). When Does Parkinson's Disease Begin? *Movement Disorders, 24*(14), S656-S664. doi:10.1002/mds.22672

Galvez, V., Diaz, R., Hernandez-Castillo, C. R., Campos-Romo, A., & Fernandez-Ruiz, J. (2014). Olfactory performance in spinocerebellar ataxia type 7 patients. *Parkinsonism & Related Disorders, 20*(5), 499-502. doi:10.1016/j.parkreldis.2014.01.024

Gao, C., Sun, H., Wang, T., Tang, M., Bohnen, N. I., Müller, M. L. T. M., . . . Dinov, I. D. (2018). Model-based and Model-free Machine Learning Techniques for Diagnostic Prediction and Classification of Clinical Outcomes in Parkinson's Disease. *Sci Rep, 8*(1), 7129. doi:10.1038/s41598-018-24783-4

Gareth, J., Witten, D., Hastie T., Tibshirani, rR. (2013). *An Introduction to Statisical Learning* G. Casella (Ed.)

Gibb, W. R., & Lees, A. J. (1988). The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J Neurol Neurosurg Psychiatry, 51*(6), 745-752.

Goetz, C. G., Tilley, B. C., Shaftman, S. R., Stebbins, G. T., Fahn, S., Martinez-Martin, P., . . . Movement Disorder Soc, U. (2008). Movement Disorder Society-Sponsored Revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS): Scale Presentation and Clinimetric Testing Results. *Movement Disorders, 23*(15), 2129-2170. doi:10.1002/mds.22340

Goldman, J. G., Vernaleo, B. A., Camicioli, R., Dahodwala, N., Dobkin, R. D., Ellis, T., . . . Simmonds, D. (2018). Cognitive impairment in Parkinson's disease: a report from a multidisciplinary symposium on unmet needs and future directions to maintain cognitive health. *Npj Parkinsons Disease, 4*. doi:10.1038/s41531-018-0055-3

Greenland, S. (1995). DOSE-RESPONSE AND TREND ANALYSIS IN EPIDEMIOLOGY - ALTERNATIVES TO CATEGORICAL ANALYSIS. *Epidemiology, 6*(4), 356-365. doi:10.1097/00001648-199507000-00005

Group, S. C. (2013, November 14). Logistic Regression (R). Retrieved from http://scg.sdsu.edu/logit_r/

Guo, F. T., Wang, G. Y., Su, Z. W., Liang, H. L., Wang, W. H., Lin, F. F., & Liu, A. Q. (2016). What drives forest fire in Fujian, China? Evidence from logistic regression and Random Forests. *International Journal of Wildland Fire, 25*(5), 505-519. doi:10.1071/wf15121

Habibzadeh, F., Habibzadeh, P., & Yadollahie, M. (2016). On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochemia Medica, 26*(3), 297-307. doi:10.11613/bm.2016.034

Hanley, J. A., & McNeil, B. J. (1982). THE MEANING AND USE OF THE AREA UNDER A RECEIVER OPERATING CHARACTERISTIC (ROC) CURVE. *Radiology, 143*(1), 29-36. doi:10.1148/radiology.143.1.7063747

Harrell, F., E. (2013). *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. New York: Springer-Verlag

Hastie, T., & Tibshirani, R. (1995). Generalized additive models for medical research. *Stat Methods Med Res, 4*(3), 187-196. doi:10.1177/096228029500400302

Hawkes, C. H., Del Tredici, K., & Braak, H. (2009). Parkinson's Disease The Dual Hit Theory Revisited. *International Symposium on Olfaction and Taste, 1170*, 615-622. doi:10.1111/j.1749-6632.2009.04365.x

Helwany M, & B., B. (2020). *Neuroanatomy, Cranial Nerve 1 (Olfactory)* FL: Treasure Island (FL): StatPearls Publishing; 2020

Hernesniemi, J. A., Mahdiani, S., Tynkkynen, J. A., Lyytikäinen, L. P., Mishra, P. P., Lehtimäki, T., . . . Oksala, N. (2019). Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome - the MADDEC study. *Ann Med*, 1-8. doi:10.1080/07853890.2019.1596302

Hess, C. W., & Okun, M. S. (2016). Diagnosing Parkinson Disease. *Continuum (Minneapolis, Minn.), 22*(4 Movement Disorders), 1047-1063. doi:10.1212/con.0000000000000345

Hilker, R., Schweitzer, K., Coburger, S., Ghaemi, M., Weisenbach, S., Jacobs, A. H., . . . Heiss, W. D. (2005). Nonlinear progression of Parkinson disease as determined by serial positron emission tomographic imaging of striatal fluorodopa F 18 activity. *Arch Neurol, 62*(3), 378-382. doi:10.1001/archneur.62.3.378

Ho, T. K. (1995). *Random decsion forests*. Paper presented at the Proceedings of the 3rd International Conference on Document Analysis and Recognition Montreal, QC.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *Ieee Transactions on Pattern Analysis and Machine Intelligence, 20*(8), 832-844.

Hong, W. S., Haimovich, A. D., & Taylor, R. A. (2018). Predicting hospital admission at emergency department triage using machine learning. *Plos One, 13*(7). doi:10.1371/journal.pone.0201016

Hosmer, D., W., Jr.;, Lemeshow, S.; Sturdivant, R., X. (2013). *Applied Logistic Regression.* Hoboken: John Wiley & Sons, Inc.

https://topepo.github.io/. (2019, March, 3). Variable Importance. Retrieved from http://topepo.github.io/caret/variable-importance.html

Hughes, A. J., Daniel, S. E., & Lees, A. J. (2001). Improved accuracy of clinical diagnosis of Lewy body Parkinson's disease. *Neurology, 57*(8), 1497-1499.

Hummel, T., Futschik, T., Frasnelli, J., & Huttenbrink, K. B. (2003). Effects of olfactory function, age, and gender, on trigeminally mediated sensations: a study based on the lateralization of chemosensory stimuli. *Toxicology Letters, 140*, 273-280. doi:10.1016/s0378-4274(03)00078-x

Investigators, P. S. G. P. (2007). Mixed lineage kinase inhibitor CEP-1347 fails to delay disability in early Parkinson disease. *Neurology, 69*(15), 1480-1490. doi:10.1212/01.wnl.0000277648.63931.c0

Jain, S., Park, S. Y., & Comer, D. (2015). Patterns of Motor and Non-Motor Features in Medication-Naive Parkinsonism. *Neuroepidemiology, 45*(1), 59-69. doi:10.1159/000437228

Jeatrakul, P. W., K.; Fung, C. (2010). Classification of Imbalanced Data By Combining the Complementary Neural Network and SMOTE Algorithm. *Neural Information Processing; Models and Applications*, 152-159.

Johns, M. W. (1991). A NEW METHOD FOR MEASURING DAYTIME SLEEPINESS - THE EPWORTH SLEEPINESS SCALE. *Sleep, 14*(6), 540-545. doi:10.1093/sleep/14.6.540

Jones, K., & Wrigley, N. (1995). GENERALIZED ADDITIVE-MODELS, GRAPHICAL DIAGNOSTICS, AND LOGISTIC-REGRESSION. *Geographical Analysis, 27*(1), 1-21.

Jucker, M., & Walker, L. C. (2013). Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature, 501*(7465), 45-51. doi:10.1038/nature12481

Kaggle. (2018a). What is XGBoost. *Learn Machine Learning series.* Retrieved from https://www.kaggle.com/dansbecker/xgboost

Kaggle. (2018b). XGBoost. Retrieved from https://www.kaggle.com/dansbecker/xgboost

Kalaitzakis, M. E., Graeber, M. B., Gentleman, S. M., & Pearce, R. K. B. (2008). The dorsal motor nucleus of the vagus is not an obligatory trigger site of Parkinson's disease: a critical analysis of alpha-synuclein staging. *Neuropathology and Applied Neurobiology, 34*(3), 284-295. doi:10.1111/j.1365-2990.2007.00923.x

Kandel, R., Schwartz, J. H., & Jessell, T. M. (2000). *Principals of Neural Science*. New York: McGraw-Hill.

Kang, J. H., Irwin, D. J., Chen-Plotkin, A. S., Siderowf, A., Caspell, C., Coffey, C. S., . . . Parkinson's Progression, M. (2013). Association of Cerebrospinal Fluid beta-Amyloid 1-42, T-tau, P-tau(181), and alpha-Synuclein Levels With Clinical Features of Drug-Naive Patients With Early Parkinson Disease. *Jama Neurology, 70*(10), 1277-1287. doi:10.1001/jamaneurol.2013.3861

Kang, J. H., Mollenhauer, B., Coffey, C. S., Toledo, J. B., Weintraub, D., Galasko, D. R., . . . Initiative, P. s. P. M. (2016). CSF biomarkers associated with disease heterogeneity in early Parkinson's disease: the Parkinson's Progression Markers Initiative study. *Acta Neuropathol, 131*(6), 935-949. doi:10.1007/s00401-016-1552-2

Karachi, C., Grabli, D., Bernard, F. A., Tande, D., Wattiez, N., Belaid, H., . . . Francois, C. (2010). Cholinergic mesencephalic neurons are involved in gait and postural disorders in Parkinson disease. *Journal of Clinical Investigation, 120*(8), 2745-2754. doi:10.1172/jci42642

Katzenschlager, R., Zijlmans, J., Evans, A., Watt, H., & Lees, A. J. (2004). Olfactory function distinguishes vascular parkinsonism from Parkinson's disease. *Journal of Neurology Neurosurgery and Psychiatry, 75*(12), 1749-1752. doi:10.1136/jnnp.2003.035287

Kaufman, M. J., & Madras, B. K. (1991). SEVERE DEPLETION OF COCAINE RECOGNITION SITES ASSOCIATED WITH THE DOPAMINE TRANSPORTER IN PARKINSONS-DISEASED STRIATUM. *Synapse, 9*(1), 43-49. doi:10.1002/syn.890090107

KDnuggetts. (2017). XGBoost. Retrieved from https://www.kdnuggets.com/?s=XGBoost

Keeler, J. F., Pretsell, D. O., & Robbins, T. W. (2014). FUNCTIONAL IMPLICATIONS OF DOPAMINE D1 VS. D2 RECEPTORS: A 'PREPARE AND SELECT' MODEL OF THE STRIATAL DIRECT VS. INDIRECT PATHWAYS. *Neuroscience, 282*, 156-175. doi:10.1016/j.neuroscience.2014.07.021

Kirasich, K. S., T.; Sadler, B. (2018). Random forest vs logistic regression: binary classification for heterogenous datasets. *SMU Data Science Review, 1*, 3.

Koehler, S., Baer, K.-J., & Wagner, G. (2016). Differential Involvement of Brainstem Noradrenergic and Midbrain Dopaminergic Nuclei in Cognitive Control. *Human Brain Mapping, 37*(6), 2305-2318. doi:10.1002/hbm.23173

Kordower, J. H., Olanow, C. W., Dodiya, H. B., Chu, Y. P., Beach, T. G., Adler, C. H., . . . Bartus, R. T. (2013). Disease duration and the integrity of the nigrostriatal system in Parkinson's disease. *Brain, 136*, 2419-2431. doi:10.1093/brain/awt192

Kuhn, M. (2013). *Applied Predictive Modeling*. In K. Johnson (Ed.): Springer.

Kuhn, M. (2019). Retrieved from https://topepo.github.io/caret/subsampling-for-class-imbalances.html - resampling

Kuhn, M. (2019, March, 3). Package caret. CRAN. Retrieved from https://www.rdocumentation.org/packages/caret/versions/6.0-80

Landis, J. R., & Koch, G. G. (1977). MEASUREMENT OF OBSERVER AGREEMENT FOR CATEGORICAL DATA. *Biometrics, 33*(1), 159-174. doi:10.2307/2529310

Lee, H. J., Suk, J. E., Patrick, C., Bae, E. J., Cho, J. H., Rho, S., . . . Lee, S. J. (2010). Direct Transfer of alpha-Synuclein from Neuron to Astroglia Causes Inflammatory Responses in Synucleinopathies. *Journal of Biological Chemistry, 285*(12), 9262-9272. doi:10.1074/jbc.M109.081125

Liaw, A. W., M. (2018, March 25). Package random forest. CRAN. Retrieved from https://www.stat.berkeley.edu/~breiman/RandomForests/

Lim, T. S., Loh, W. Y., & Shih, Y. S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning, 40*(3), 203-228. doi:10.1023/a:1007608224229

Lin, M. T., & Beal, M. F. (2006). Mitochondrial dysfunction and oxidative stress in neurodegenerative diseases. *Nature, 443*(7113), 787-795. doi:10.1038/nature05292

Liu, C. R., Frazier, P., & Kumar, L. (2007). Comparative assessment of the measures of thematic classification accuracy. *Remote Sensing of Environment, 107*(4), 606-616. doi:10.1016/j.rse.2006.10.010

Liu, R., Crawford, J., Callahan, P. M., Terry, A. V., Constantinidis, C., & Blake, D. T. (2018). Intermittent stimulation in the nucleus basalis of meynert improves sustained attention in rhesus monkeys. *Neuropharmacology, 137*, 202-210. doi:10.1016/j.neuropharm.2018.04.026

Llorens, F., Schmitz, M., Varges, D., Kruse, N., Gotzmann, N., Gmitterova, K., . . . Zerr, I. (2016). Cerebrospinal alpha-synuclein in alpha-synuclein aggregation disorders: tau/alpha-synuclein ratio as potential biomarker for dementia with Lewy bodies. *Journal of Neurology, 263*(11), 2271-2277. doi:10.1007/s00415-016-8259-0

Louis, E. D., Rios, E., & Henchcliffe, C. (2010). How are we doing with the treatment of essential tremor (ET)? *European Journal of Neurology, 17*(6), 882-884. doi:10.1111/j.1468-1331.2009.02926.x

Luo, L., Li, J., Liu, C., & Shen, W. (2019). Using machine-learning methods to support health-care professionals in making admission decisions. *Int J Health Plann Manage*. doi:10.1002/hpm.2769

Mackie, P., Lebowitz, J., Saadatpour, L., Nickoloff, E., Gaskill, P., & Khoshbouei, H. (2018). The dopamine transporter: An unrecognized nexus for dysfunctional peripheral immunity and signaling in Parkinson's Disease. *Brain Behavior and Immunity, 70*, 21-35. doi:10.1016/j.bbi.2018.03.020

Malenka RC, Nestler EJ, & SE, H. (Eds.). (2009). *Chapter 6: Widely Projecting Systems: Monoamines, Acetylcholine, and Orexin. Molecular Neuropharmacology: A Foundation for Clinical Neuroscience (2nd ed.)* (2nd ed.). New York: McGraw-Hill Medical.

Mani, A. (2015). Training and assessing classification rules with unbalanced data. Retrieved from https://stats.stackexchange.com/questions/166458/rose-and-smote-oversampling-methods

Marek, K., Chowdhury, S., Siderowf, A., Lasch, S., Coffey, C. S., Caspell-Garcia, C., . . . Initiative, P. s. P. M. (2018). The Parkinson's progression markers initiative (PPMI) - establishing a PD biomarker cohort. *Ann Clin Transl Neurol, 5*(12), 1460-1477. doi:10.1002/acn3.644

Marek, K., Seibyl, J., Eberly, S., Oakes, D., Shoulson, I., Lang, A. E., . . . Investigators, P. S. G. P. (2014). Longitudinal follow-up of SWEDD subjects in the PRECEPT Study. *Neurology, 82*(20), 1791-1797. doi:10.1212/WNL.0000000000000424

Martinez-Gonzalez, C., Bolam, J. P., & Mena-Segovia, J. (2011). Topographical organization of the pedunculopontine nucleus. *Frontiers in Neuroanatomy, 5*. doi:10.3389/fnana.2011.00022

McCarthy, P. (2020). FSLeyes. Retrieved from (https://git.fmrib.ox.ac.uk/fsl/fsleyes/fsleyes/

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, 105-142.

McKinley, J. W., Shi, Z., Kawikova, I., Hur, M., Bamford, I. J., Devi, S. P. S., . . . Bamford, N. S. (2019). Dopamine Deficiency Reduces Striatal Cholinergic Interneuron Function in Models of Parkinson's Disease. *Neuron, 103*(6), 1056-+. doi:10.1016/j.neuron.2019.06.013

Meade, R. M., Fairlie, D. P., & Mason, J. M. (2019). Alpha-synuclein structure and Parkinson's disease - lessons and emerging principles. *Molecular Neurodegeneration, 14*. doi:10.1186/s13024-019-0329-1

Menardi, G. T., N. (2014). Training and assessing classification rules with unbalanced data. *Data Min Knowl Disc, 28*(92). doi:https://doi.org/10.1007/s10618-012-0295-5

Molinaro, A. M., Lostritto, K., & van der Laan, M. (2010). partDSA: deletion/substitution/addition algorithm for partitioning the covariate space in prediction. *Bioinformatics, 26*(10), 1357-1363. doi:10.1093/bioinformatics/btq142

Mollenhauer, B., Locascio, J. J., Schulz-Schaeffer, W., Sixel-Döring, F., Trenkwalder, C., & Schlossmacher, M. G. (2011). α-Synuclein and tau concentrations in cerebrospinal fluid of patients presenting with parkinsonism: a cohort study. *Lancet Neurol, 10*(3), 230-240. doi:10.1016/S1474-4422(11)70014-X

Murakami, H., Tokuda, T., El-Agnaf, O. M. A., Ohmichi, T., Miki, A., Ohashi, H., . . . Ono, K. (2019). Correlated levels of cerebrospinal fluid pathogenic proteins in drug-naive Parkinson's disease. *Bmc Neurology, 19*. doi:10.1186/s12883-019-1346-y

Nalls, M. A., McLean, C. Y., Rick, J., Eberly, S., Hutten, S. J., Gwinn, K., . . . investigators, P. s. D. B. P. a. P. s. P. M. I. (2015). Diagnosis of Parkinson's disease on the basis of clinical and genetic classification: a population-based modelling study. *Lancet Neurol, 14*(10), 1002-1009. doi:10.1016/S1474-4422(15)00178-7

Nasreddine, Z. S., Phillips, N. A., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., . . . Chertkow, H. (2005). The montreal cognitive assessment, MoCA: A brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society, 53*(4), 695-699. doi:10.1111/j.1532-5415.2005.53221.x

Newman, D. J. H., S. ; Blake, C. L.; Merz, C. J. (1998). *UCI Repository of machine learning databases*. Retrieved from: http://www.ics.uci.edu/~mlearn/MLRepository.html

Nicastro, N., Garibotto, V., Badoud, S., & Burkhard, P. R. (2016). Scan without evidence of dopaminergic deficit: A 10-year retrospective study. *Parkinsonism Relat Disord, 31*, 53-58. doi:10.1016/j.parkreldis.2016.07.002

Nirenberg, M. J., Vaughan, R. A., Uhl, G. R., Kuhar, M. J., & Pickel, V. M. (1996). The dopamine transporter is localized to dendritic and axonal plasma membranes of nigrostriatal dopaminergic neurons. *Journal of Neuroscience, 16*(2), 436-447.

Niznik, H. B., Fogel, E. F., Fassos, F. F., & Seeman, P. (1991). THE DOPAMINE TRANSPORTER IS ABSENT IN PARKINSONIAN PUTAMEN AND REDUCED IN THE CAUDATE-NUCLEUS. *Journal of Neurochemistry, 56*(1), 192-198. doi:10.1111/j.1471-4159.1991.tb02580.x

Olofsson, P., Foody, G. M., Herold, M., Stehman, S. V., Woodcock, C. E., & Wulder, M. A. (2014). Good practices for estimating area and assessing accuracy of land change. *Remote Sensing of Environment, 148*, 42-57. doi:10.1016/j.rse.2014.02.015

Olson, E. J., Boeve, B. F., & Silber, M. H. (2000). Rapid eye movement sleep behaviour disorder: demographic, clinical and laboratory findings in 93 cases. *Brain, 123*, 331-339. doi:10.1093/brain/123.2.331

Overflow, S. (2016, October 5). Difference between varimp (caret) and importance (randomforest) for random forest. Retrieved from https://stackoverflow.com/questions/37888619/difference-between-varimp-caret-and-importance-randomforest-for-random-fores?rq=1

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing, 26*(1), 217-222. doi:10.1080/01431160412331269698

Pampel, F. C. (2000). *Logistic regression: a primer*. Thousand Oaks, CA: Sage.

Pan-Montojo, F., Anichtchik, O., Dening, Y., Knels, L., Pursche, S., Jung, R., . . . Funk, R. H. W. (2010). Progression of Parkinson's Disease Pathology Is Reproduced by Intragastric Administration of Rotenone in Mice. *Plos One, 5*(1). doi:10.1371/journal.pone.0008762

Patel, R. M., & Pinto, J. M. (2014). Olfaction: Anatomy, Physiology, and Disease. *Clinical Anatomy, 27*(1), 54-60. doi:10.1002/ca.22338

Perkins, N. J., & Schisterman, E. F. (2006). The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. *American Journal of Epidemiology, 163*(7), 670-675. doi:10.1093/aje/kwj063

Perlich, C., Provost, F., & Simonoff, J. S. (2004). Tree induction vs. logistic regression: A learning-curve analysis. *Journal of Machine Learning Research, 4*(2), 211-255. doi:10.1162/153244304322972694

Pontius, R. G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing, 32*(15), 4407-4429. doi:10.1080/01431161.2011.552923

Poole, C. (2001). Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology, 12*(3), 291-294. doi:10.1097/00001648-200105000-00005

Prashanth, R., Dutta Roy, S., Mandal, P. K., & Ghosh, S. (2016). High-Accuracy Detection of Early Parkinson's Disease through Multimodal Features and Machine Learning. *Int J Med Inform, 90*, 13-21. doi:10.1016/j.ijmedinf.2016.03.001

Prashanth, R., Roy, S. D., Mandal, P. K., & Ghosh, S. (2017). High-Accuracy Classification of Parkinson's Disease Through Shape Analysis and Surface Fitting in 123I-Ioflupane SPECT Imaging. *IEEE J Biomed Health Inform, 21*(3), 794-802. doi:10.1109/JBHI.2016.2547901

Provost, F. F., T.; Kohavi, R. (1998). *The case against accuracy estimation for comparing induction algorithms.* Paper presented at the International Conference on Machine Learning.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior, 29*(5), 615-620. doi:10.1007/s10979-005-6832-7

Ridley, R. A., Baker, H. F., Leow-Dyke, A., & Cummings, R. M. (2005). Further analysis of the effects of immunotoxic lesions of the basal nucleus of Meynert reveals substantial impairment on visual discrimination learning in monkeys. *Brain Research Bulletin, 65*(5), 433-442. doi:10.1016/j.brainresbull.2005.02.025

Rizzi, G., & Tan, K. R. (2017). Dopamine and Acetylcholine, a Circuit Point of View in Parkinson's Disease. *Frontiers in Neural Circuits, 11*. doi:10.3389/fncir.2017.00110

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., & Muller, M. (2011). pROC: an open-source package for R and S plus to analyze and compare ROC curves. *Bmc Bioinformatics, 12*. doi:10.1186/1471-2105-12-77

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. *Isprs Journal of Photogrammetry and Remote Sensing, 67*, 93-104. doi:10.1016/j.isprsjprs.2011.11.002

Ross, G. W., Petrovitch, H., Abbott, R. D., Tanner, C. M., Popper, J., Masaki, K., . . . White, L. R. (2008). Association of olfactory dysfunction with risk for future Parkinson's disease. *Annals of Neurology, 63*(2), 167-173. doi:10.1002/ana.21291

Royston, P. (2000). A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. *Statistics in Medicine, 19*(14), 1831-1847. doi:10.1002/1097-0258(20000730)19:14<1831::aid-sim502>3.0.co;2-1

Rubin, J. E., McIntyre, C. C., Turner, R. S., & Wichmann, T. (2012). Basal ganglia activity patterns in parkinsonism and computational modeling of their downstream effects. *European Journal of Neuroscience, 36*(2), 2213-2228. doi:10.1111/j.1460-9568.2012.08108.x

Saalmann, Y. B. (2014). Intralaminar and medial thalamic influence on cortical synchrony, information transmission and cognition. *Frontiers in systems neuroscience, 8*, 83-83. doi:10.3389/fnsys.2014.00083

Savica, R., Rocca, W. A., & Ahlskog, J. E. (2010). When Does Parkinson Disease Start? *Archives of Neurology, 67*(7), 798-801.

Sawle, G. V., Playford, E. D., Burn, D. J., Cunningham, V. J., & Brooks, D. J. (1994). SEPARATING PARKINSONS-DISEASE FROM NORMALITY - DISCRIMINANT FUNCTION-ANALYSIS OF FLUORODOPA F-18 POSITRON EMISSION TOMOGRAPHY DATA. *Archives of Neurology, 51*(3), 237-243. doi:10.1001/archneur.1994.00540150027011

Schapira, A. H. V. (2009). Neurobiology and treatment of Parkinson's disease. *Trends in Pharmacological Sciences, 30*(1), 41-47. doi:10.1016/j.tips.2008.10.005

Schenck, C. H., Bundlie, S. R., & Mahowald, M. W. (1996). Delayed emergence of a parkinsonian disorder in 38% of 29 older men initially diagnosed with idiopathic rapid eye movement sleep behavior disorder. *Neurology, 46*(2), 388-393. doi:10.1212/wnl.46.2.388

Schendan, H. E., Amick, M. M., & Cronin-Golomb, A. (2009). Role of a Lateralized Parietal-Basal Ganglia Circuit in Hierarchical Pattern Perception: Evidence From Parkinson's Disease. *Behavioral Neuroscience, 123*(1), 125-136. doi:10.1037/a0013734

Schneider, S. A., Edwards, M. J., Mir, P., Cordivari, C., Hooker, J., Dickson, J., . . . Bhatia, K. P. (2007). Patients with adult-onset dystonic tremor resembling parkinsonian tremor have scans without evidence of dopaminergic deficit (SWEDDs). *Mov Disord, 22*(15), 2210-2215. doi:10.1002/mds.21685

Schofield, P. W., Ebrahimi, H., Jones, A. L., Bateman, G. A., & Murray, S. R. (2012). An olfactory 'stress test' may detect preclinical Alzheimer's disease. *BMC Neurol, 12*, 24. doi:10.1186/1471-2377-12-24

Schwartz, J. R. L., & Roth, T. (2008). Neurophysiology of Sleep and Wakefulness: Basic Science and Clinical Implications. *Current Neuropharmacology, 6*(4), 367-378.

Schwingenschuh, P., Ruge, D., Edwards, M. J., Terranova, C., Katschnig, P., Carrillo, F., . . . Bhatia, K. P. (2010). Distinguishing SWEDDs Patients with Asymmetric Resting Tremor from Parkinson's Disease: A Clinical and Electrophysiological Study. *Movement Disorders, 25*(5), 560-569. doi:10.1002/mds.23019

Seibyl, J. P., Marek, K. L., Quinlan, D., Sheff, K., Zoghbi, S., Zeaponce, Y., . . . Innis, R. B. (1995). DECREASED SINGLE-PHOTON EMISSION COMPUTED TOMOGRAPHIC (123) I-BETA-CIT STRIATAL UPTAKE CORRELATES WITH SYMPTOM SEVERITY IN PARKINSONS-DISEASE. *Annals of Neurology, 38*(4), 589-598. doi:10.1002/ana.410380407

Shah, B. R., Lehman, V. T., Kaufmann, T. J., Blezek, D., Waugh, J., Imphean, D., . . . Chopra, R. (2020). Advanced MRI techniques for transcranial high intensity focused ultrasound targeting. *Brain*. doi:10.1093/brain/awaa107

Shah, M., Muhammed, N., Findley, L. J., & Hawkes, C. H. (2008). Olfactory tests in the diagnosis of essential tremor. *Parkinsonism & Related Disorders, 14*(7), 563-568. doi:10.1016/j.parkreldis.2007.12.006

Shapiro, S. S., & Wilk, M. B. (1965). 3-4.

Sharma, S., Moon, C. S., Khogali, A., Haidous, A., Chabenne, A., Ojo, C., . . . Ebadi, M. (2013). Biomarkers in Parkinson's disease (recent update). *Neurochemistry International, 63*(3), 201-229. doi:10.1016/j.neuint.2013.06.005

Shi, M., Bradner, J., Hancock, A. M., Chung, K. A., Quinn, J. F., Peskind, E. R., . . . Zhang, J. (2011). Cerebrospinal Fluid Biomarkers for Parkinson Disease Diagnosis and Progression. *Annals of Neurology, 69*(3), 570-580. doi:10.1002/ana.22311

Shimoda, A., Ichikawa, D., & Oyama, H. (2018). Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. *Comput Methods Programs Biomed, 163*, 39-46. doi:10.1016/j.cmpb.2018.05.032

Shipe, M. E., Deppen, S. A., Farjah, F., & Grogan, E. L. (2019). Developing prediction models for clinical use using logistic regression: an overview. *Journal of Thoracic Disease, 11*, S574-S584. doi:10.21037/jtd.2019.01.25

Siegfried, J., & Lippitz, B. (1994). BILATERAL CHRONIC ELECTROSTIMULATION OF VENTROPOSTEROLATERAL PALLIDUM - A NEW THERAPEUTIC APPROACH FOR ALLEVIATING ALL PARKINSONIAN SYMPTOMS. *Neurosurgery, 35*(6), 1126-1129. doi:10.1227/00006123-199412000-00016

Smythies, J. (2009). Philosophy, perception, and neuroscience. *Perception, 38*(5), 638-651. doi:10.1068/p6025

Spielberger, C. D., Gorsuch, R. L., & Lushene, R. E. (1970). Manual for the State Trait Inventory. Consulting Psychologists Press, Palo Alto, CA.

Stiasny-Kolster, K., Mayer, G., Schaeffer, S., Moeller, J. C., Gutenbrunner, M. H., & Oertel, W. H. (2007). The REM sleep behavior disorder screening questionnaire - A new diagnostic instrument. *Movement Disorders, 22*(16), 2386-2393. doi:10.1002/mds.21740

Stoessl, A. J. (2010). Scans Without Evidence of Dopamine Deficiency: The Triumph of Careful Clinical Assessment. *Movement Disorders, 25*(5), 529-530. doi:10.1002/mds.23138

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., & Zeileis, A. (2008). Conditional variable importance for random forests. *Bmc Bioinformatics, 9*. doi:10.1186/1471-2105-9-307

Strobl, C., Boulesteix, A. L., Zeileis, A., & Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *Bmc Bioinformatics, 8*. doi:10.1186/1471-2105-8-25

Sun, X., & Xu, W. (2014). Fast Implementation of DeLong's Algorithm for Comparing the Areas Under Correlated Receiver Operating Characteristic Curves. *Ieee Signal Processing Letters, 21*(11), 1389-1393. doi:10.1109/lsp.2014.2337313

Surmeier, D. J., Ding, J., Day, M., Wang, Z. F., & Shen, W. X. (2007). D1 and D2 dopamine-receptor modulation of striatal glutamatergic signaling in striatal medium spiny neurons. *Trends in Neurosciences, 30*(5), 228-235. doi:10.1016/j.tins.2007.03.008

Suwijn, S. R., Verschuur, C. V. M., Slim, M. A., Booij, J., & de Bie, R. M. A. (2019). Reliability of visual assessment by non-expert nuclear medicine physicians and appropriateness of indications of I-123 FP-CIT SPECT imaging by neurologists in patients with early drug-naive Parkinson's disease. *Ejnmmi Research, 9*. doi:10.1186/s13550-019-0537-2

Suzuki, M., Hashimoto, M., Yoshioka, M., Murakami, M., Kawasaki, K., & Urashima, M. (2011). The odor stick identification test for Japanese differentiates Parkinson's disease from multiple system atrophy and progressive supra nuclear palsy. *Bmc Neurology, 11*. doi:10.1186/1471-2377-11-157

Svensson, E., Horvath-Puho, E., Thomsen, R. W., Djurhuus, J. C., Pedersen, L., Borghammer, P., & Sorensen, H. T. (2015). Vagotomy and Subsequent Risk of Parkinson's Disease. *Annals of Neurology, 78*(4), 522-529. doi:10.1002/ana.24448

Tabachnick, B. G. F., L. S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.

Tang, C. Q., Li, J. Q., Xu, D. Y., Liu, X. B., Hou, W. J., Lyu, K. Y., . . . Xia, Z. F. (2018). [Comparison of machine learning method and logistic regression model in prediction of acute kidney injury in severely burned patients]. *Zhonghua Shao Shang Za Zhi, 34*(6), 343-348. doi:10.3760/cma.j.issn.1009-2587.2018.06.006

Tapiola, T., Alafuzoff, I., Herukka, S.-K., Parkkinen, L., Hartikainen, P., Soininen, H., & Pirttila, T. (2009). Cerebrospinal Fluid beta-Amyloid 42 and Tau Proteins as Biomarkers of Alzheimer-Type Pathologic Changes in the Brain. *Archives of Neurology, 66*(3), 382-389. doi:10.1001/archneurol.2008.596

Tattersall, T. L., Stratton, P. G., Coyne, T. J., Cook, R., Silberstein, P., Silburn, P. A., . . . Sah, P. (2014). Imagined gait modulates neuronal network dynamics in the human pedunculopontine nucleus. *Nature Neuroscience, 17*(3), 449-454. doi:10.1038/nn.3642

Team, R. C. (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. doi:https://doi.org/10.1016/j.aci.2018.08.003

Therneau, T. A., B.; Ripley, B. (2018). Recursive Partitioning and Regression Trees: CRAN.

Thiebaut, A. C. M., Kipnis, V., Chang, S.-C., Subar, A. F., Thompson, F. E., Rosenberg, P. S., . . . Schatzkin, A. (2007). Dietary fat and postmenopausal invasive breast cancer in the National Institutes of Health-AARP Diet and Health Study cohort. *Jnci-Journal of the National Cancer Institute, 99*(6), 451-462. doi:10.1093/jnci/djk094

Tolosi, L., & Lengauer, T. (2011). Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics, 27*(14), 1986-1994. doi:10.1093/bioinformatics/btr300

Tu, J. V., Austin, P. C., & Chan, B. T. B. (2001). Relationship between annual volume of patients treated by admitting physician and mortality after acute myocardial infarction. *Jama-Journal of the American Medical Association, 285*(24), 3116-3122. doi:10.1001/jama.285.24.3116

Uhl, G. R. (1992). NEUROTRANSMITTER TRANSPORTERS (PLUS) - A PROMISING NEW GENE FAMILY. *Trends in Neurosciences, 15*(7), 265-268. doi:10.1016/0166-2236(92)90068-j

Ulusoy, A., Rusconi, R., Perez-Revuelta, B. I., Musgrove, R. E., Helwig, M., Winzen-Reichert, B., & Di Monte, D. A. (2013). Caudo-rostral brain spreading of alpha-synuclein through vagal connections. *Embo Molecular Medicine, 5*(7), 1119-1127. doi:10.1002/emmm.201302475

Urbano, F. J., D'Onofrio, S. M., Luster, B. R., Beck, P. B., Hyde, J. R., Bisagno, V., & Garcia-Rill, E. (2014). Pedunculopontine nucleus gamma band activity-preconscious awareness, waking, and REM sleep. *Frontiers in Neurology, 5*. doi:10.3389/fneur.2014.00210

Van Den Eeden, S. K., Tanner, C. M., Bernstein, A. L., Fross, R. D., Leimpeter, A., Bloch, D. A., & Nelson, L. M. (2003). Incidence of Parkinson's disease: Variation by age, gender, and Race/Ethnicity. *American Journal of Epidemiology, 157*(11), 1015-1022. doi:10.1093/aje/kwg068

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *Bmc Medical Research Methodology, 14*. doi:10.1186/1471-2288-14-137

Van Hulse, J. K., T.; Napolitano, A. (2007). *Experimental Perspectives On Learning From Imbalanced Data.* Paper presented at the Proceedings of the 24th International Conference On Machine learning.

Venables, W. N. R., B. D. (2002). Modern Applied Statistics with S. . New York: Spriinger.

Vickers, A. J., Bianco, F. J., Serio, A. M., Eastham, J. A., Schrag, D., Klein, E. A., . . . Scardino, P. T. (2007). The surgical learning curve for prostate cancer control after radical prostatectomy. *Journal of the National Cancer Institute, 99*(15), 1171-1177. doi:10.1093/jnci/djm060

Vickers, A. J., Savage, C. J., Hruza, M., Tuerk, I., Koenig, P., Martinez-Pineiro, L., . . . Guillonneau, B. (2009). The surgical learning curve for laparoscopic radical prostatectomy: a retrospective cohort study. *Lancet Oncology, 10*(5), 475-480. doi:10.1016/s1470-2045(09)70079-8

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American Journal of Epidemiology, 165*(6), 710-718. doi:10.1093/aje/kwk052

Warren, J. D., Rohrer, J. D., Schott, J. M., Fox, N. C., Hardy, J., & Rossor, M. N. (2013). Molecular nexopathies: a new paradigm of neurodegenerative disease. *Trends in Neurosciences, 36*(10), 561-569. doi:10.1016/j.tins.2013.06.007

Wegman, E. J., & Wright, I. W. (1983). SPLINES IN STATISTICS. *Journal of the American Statistical Association, 78*(382), 351-365. doi:10.2307/2288640

Weinberg, C. R. (1995). HOW BAD IS CATEGORIZATION. *Epidemiology, 6*(4), 345-347.

Weinberger, M., Hamani, C., Hutchison, W. D., Moro, E., Lozano, A. M., & Dostrovsky, J. O. (2008). Pedunculopontine nucleus microelectrode recordings in movement disorder patients. *Experimental Brain Research, 188*(2), 165-174. doi:10.1007/s00221-008-1349-1

Wenning, G. K., Shephard, B., Hawkes, C., Petruckevitch, A., Lees, A., & Quinn, N. (1995). OLFACTORY FUNCTION IN ATYPICAL PARKINSONIAN SYNDROMES. *Acta Neurologica Scandinavica, 91*(4), 247-250. doi:10.1111/j.1600-0404.1995.tb06998.x

Wichmann, T., Bergman, H., Starr, P. A., Subramanian, T., Watts, R. L., & DeLong, M. R. (1999). Comparison of MPTP-induced changes in spontaneous neuronal discharge in the internal pallidal segment and in the substantia nigra pars reticulata in primates. *Experimental Brain Research, 125*(4), 397-409. doi:10.1007/s002210050696

Wilcox, R. R. (1990). COMPARING THE MEANS OF 2 INDEPENDENT GROUPS. *Biometrical Journal, 32*(7), 771-780.

Wolpert, D. H. (1996). The lack of A priori distinctions between learning algorithms. *Neural Computation, 8*(7), 1341-1390. doi:10.1162/neco.1996.8.7.1341

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association, 99*(467), 673-686. doi:10.1198/016214504000000980

Wood, S. N. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 70*, 495-518. doi:10.1111/j.1467-9868.2007.00646.x

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 73*, 3-36. doi:10.1111/j.1467-9868.2010.00749.x

Wood, S. N. (2019a). Generalized additive models with integrated smoothness estimation. Retrieved from https://astrostatistics.psu.edu/su07/R/library/mgcv/html/gam.html

Wood, S. N. (2019b). R: GAM concurvity measures. Retrieved from https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/concurvity.html

Wood, S. N. (2019, March 21). Package 'mgcv'.

Wullner, U., Pakzaban, P., Brownell, A. L., Hantraye, P., Burns, L., Shoup, T., . . . Isacson, O. (1994). DOPAMINE TERMINAL LOSS AND ONSET OF MOTOR SYMPTOMS IN MPTP-TREATED MONKEYS - A POSITRON EMISSION TOMOGRAPHY STUDY WITH C-11 CFT. *Experimental Neurology, 126*(2), 305-309. doi:10.1006/exnr.1994.1069

Wyman-Chick, K. A., Martin, P. K., Minar, M., & Schroeder, R. W. (2016). Cognition in Patients With a Clinical Diagnosis of Parkinson Disease and Scans Without Evidence of Dopaminergic Deficit (SWEDD): 2-Year Follow-Up. *Cognitive and Behavioral Neurology, 29*(4), 190-196. doi:10.1097/wnn.0000000000000107

Xiao, J., Ding, R., Xu, X., Guan, H., Feng, X., Sun, T., . . . Ye, Z. (2019). Comparison and development of machine learning tools in the prediction of chronic kidney disease progression. *Journal of translational medicine, 17*(1), 119-119. doi:10.1186/s12967-019-1860-0

Xu, L. J., & Pu, J. L. (2016). Alpha-Synuclein in Parkinson's Disease: From Pathogenetic Dysfunction to Potential Clinical Application. *Parkinsons Disease*. doi:10.1155/2016/1720621

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., Adey, M., & Leirer, V. O. (1983). DEVELOPMENT AND VALIDATION OF A GERIATRIC DEPRESSION SCREENING SCALE - A PRELIMINARY-REPORT. *Journal of Psychiatric Research, 17*(1), 37-49. doi:10.1016/0022-3956(82)90033-4

Youden, W. J. (1950a). INDEX FOR RATING DIAGNOSTIC TESTS. *Biometrics, 6*(2), 172-173.

Youden, W. J. (1950b). INDEX FOR RATING DIAGNOSTIC TESTS. *Cancer, 3*(1), 32-35. doi:10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3

Yu, Z., Stewart, T., Aasly, J., Shi, M., & Zhang, J. (2018). Combining clinical and biofluid markers for early Parkinson's disease detection. *Ann Clin Transl Neurol, 5*(1), 109-114. doi:10.1002/acn3.509

Zhang, O. (2015). . Retrieved from https://www.slideshare.net/OwenZhang2/tips-for-data-science-competitions

Zhang, Y., & Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics, 187*(1), 95-112. doi:10.1016/j.jeconom.2015.02.006

Zhang, Z. (2016). Model building strategy for logistic regression: purposeful selection. *Annals of Translational Medicine, 4*(6). doi:10.21037/atm.2016.02.15

Zhang, Z., Ho, K. M., & Hong, Y. (2019). Machine learning for the prediction of volume responsiveness in patients with oliguric acute kidney injury in critical care. *Crit Care, 23*(1), 112. doi:10.1186/s13054-019-2411-z

Appendices

Supporting information I (S1)

Note, FIGS S1-12-14 show the higher AUC (specificity and sensitivity) model outcomes achieved using SMOTE subsampling rather than non-subsampled (original) data for the early PD and SWEDD analysis. FIG S1-12 also includes ROSE subsampling, which was very competitive with SMOTE. But as noted in the text of this work, ROSE introduced negative values where none existed in the original data. As such, ROSE subsampling was abandoned in favour or SMOTE.

Figure Legends

FIG S1-1: early PD and control MoCA scores (histogram)
FIG S1-2: Linearity of logit plot, early PD/controls predictors
FIG S1-3: Composite: leverage (hat-values), discrepancy (studentized residuals), Cook's Distance
FIG S1-4: Leverage values; early PD/control
FIG S1-5: Discrepancy values; early PD/control
FIG S1-6: Cook's Distance; early PD/control
Expression 1: caret trainControl() settings
FIG S1-7: Linearity of logit plot, early PD/SWEDD continuous predictors
FIG S1-8: early PD/ SWEDD diagnostic plots
FIG S1-9: early PD/SWEDD leverage
FIG S1-10: early PD/SWEDD discrepancy
FIG S1-11: early PD/SWEDD Cook's Distance
FIG S1-12: Decision tree, early PD/SWEDD ROC AUC, non-sampled vs. SMOTE data models
FIG S1-13: Random forest, early PD/SWEDD ROC AUC, non-sampled vs. SMOTE data
FIG S1-14: GLM (logistic) early PD/SWEDD ROC AUC, non-subsampled vs. SMOTE model
FIG S1-15: XGBoost early PD/SWEDD ROC AUC, non-sampled vs. SMOTE data models
FIG S1-16: early PD/SWEDD years of education (histogram)
FIG S1-17 early PD/SWEDD linearity of logit plot
FIG S1-18: Composite: leverage (hat-values), discrepancy (studentized residuals), Cook's Distance values.
FIG S1-19: Leverage values; early PD/SWEDD
FIG S1-20 Discrepancy values; early PD/SWEDD
FIG S1-21 Cook's Distance; early PD/SWEDD
FIG S1-22 Logistic regression plots
FIG S1-23 early PD/SWEDD Epworth sleepiness scale (histogram)

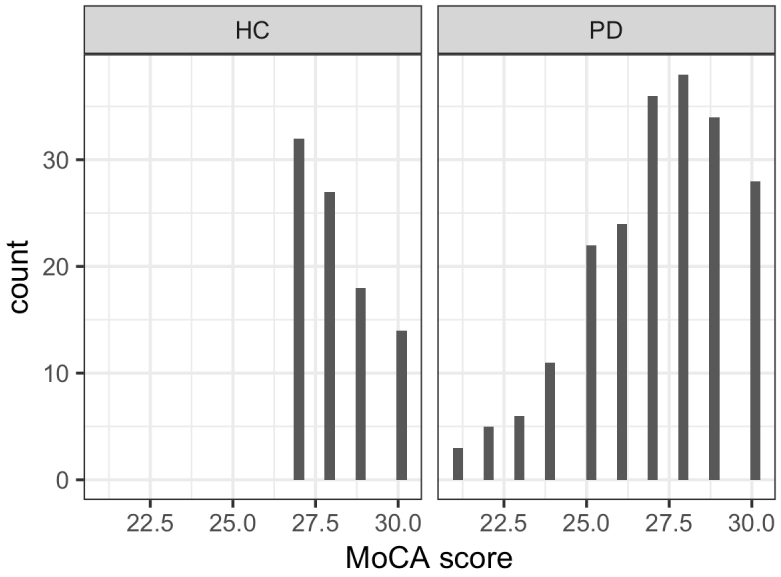


FIG S1-1: early PD and HC (controls) MoCA

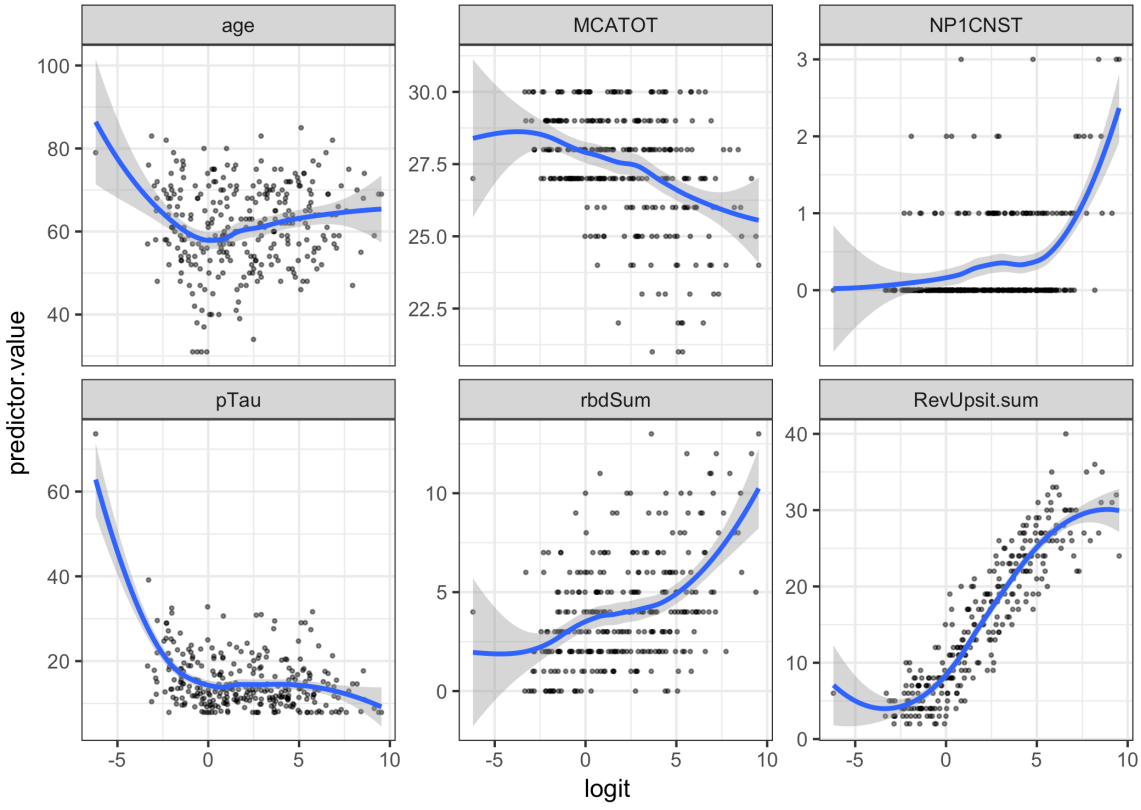


FIG S1-2 Linearity of the logit scatterplots with smoothers:, early PD/controls MCATOT = Montreal cognitive assessment test;NP1CNST = constipation; rbdSum = rapid eye movement behaviour disorder questionnaire; RevUpsit.sum = University of Pennsylvania smell test

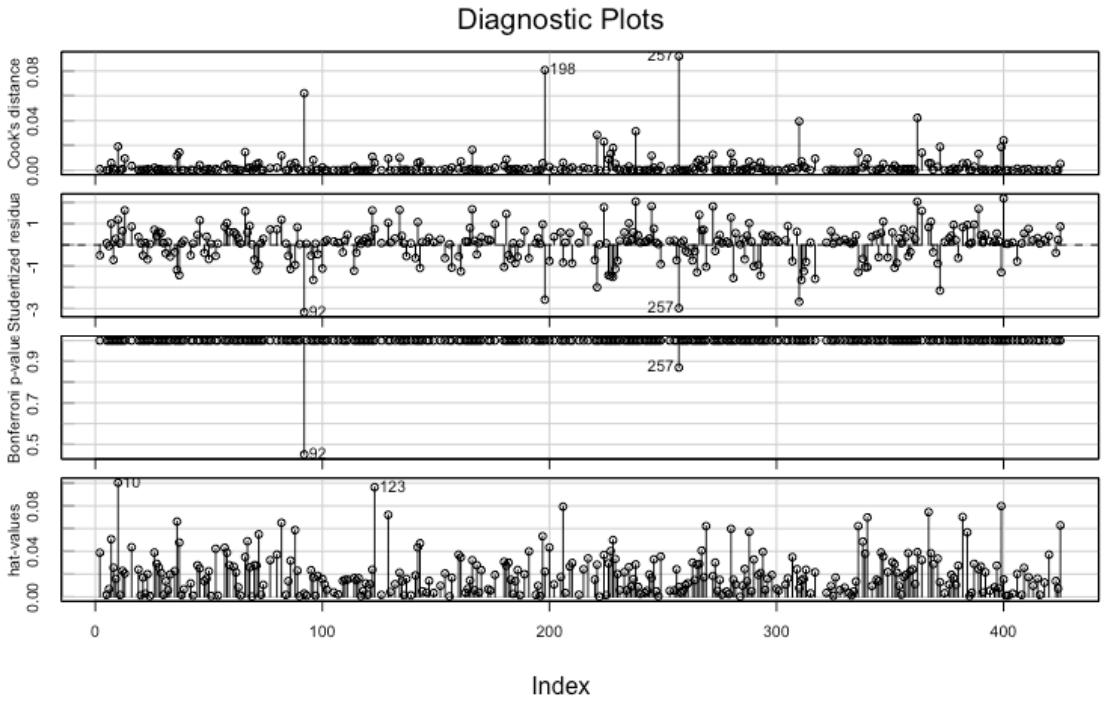


FIG S1-3: early PD/ control diagnostic plots

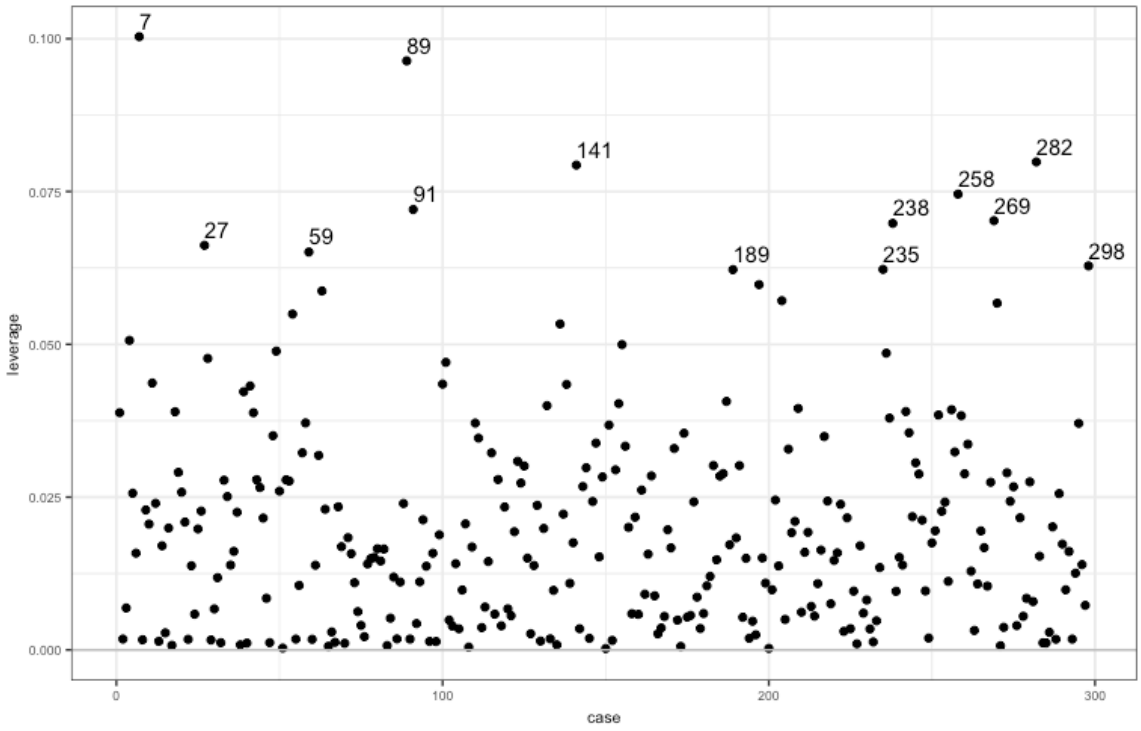


FIG S1-4 Leverage: 4% (13/298) with > .06 leverage

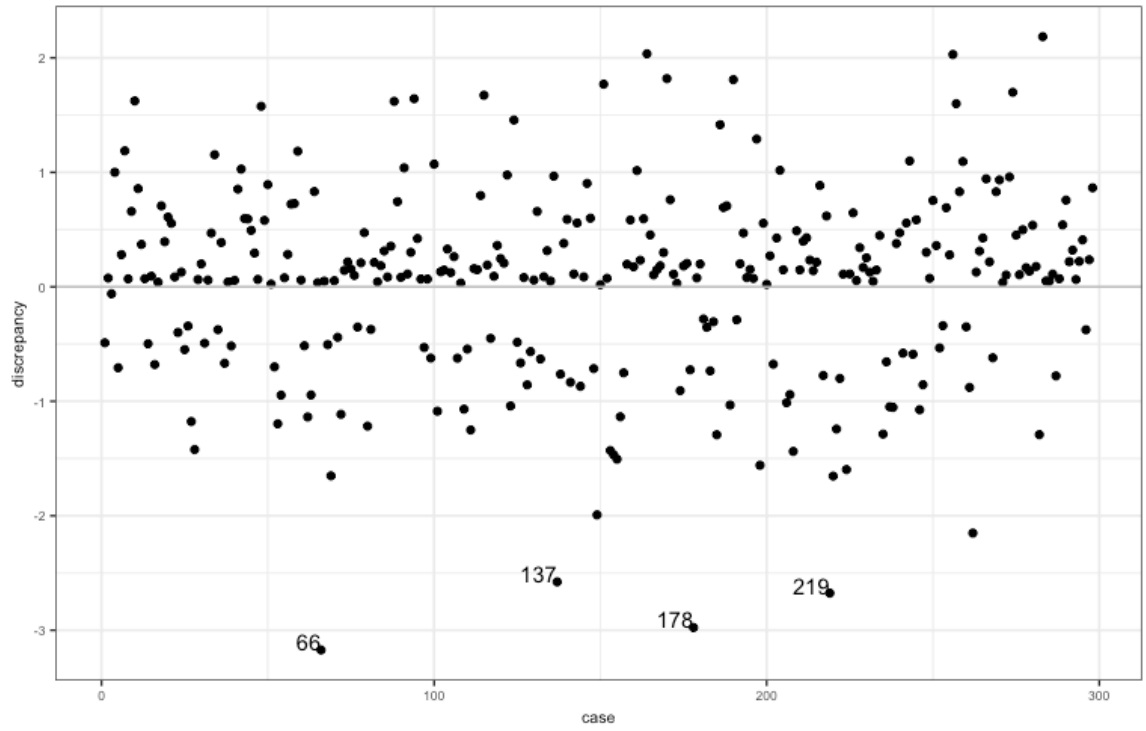


FIG S1- 5 Discrepancy: 4 of studentized residuals exceeded > ± 2.5

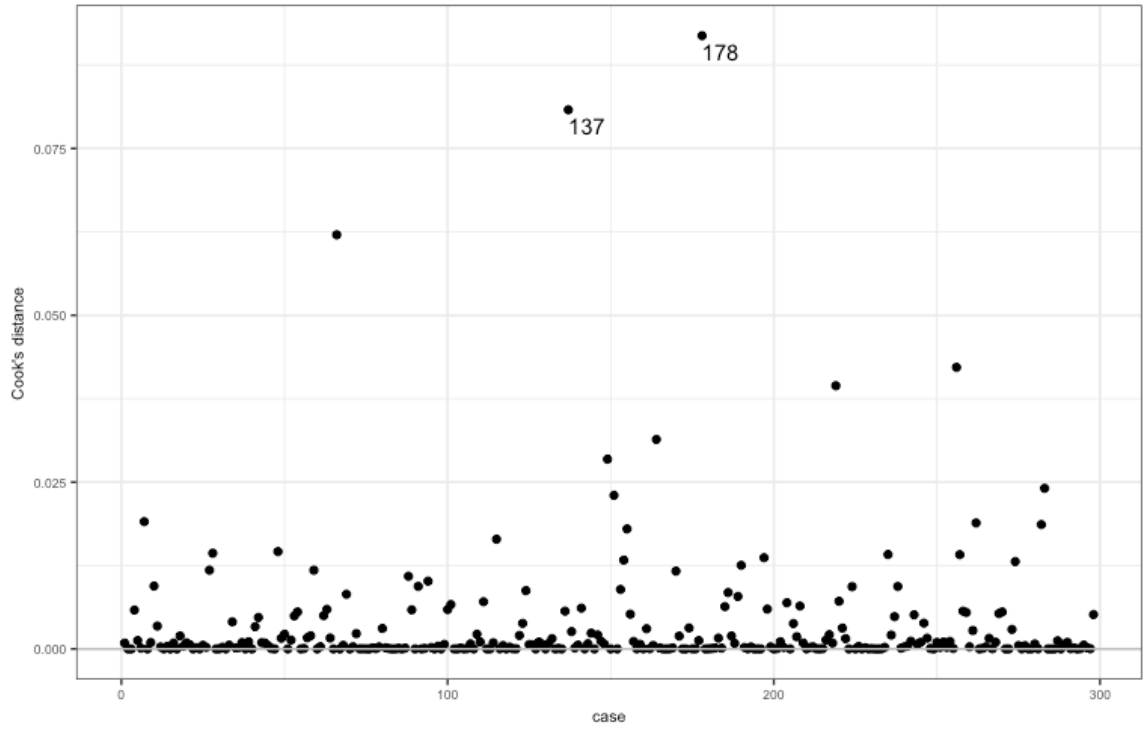


FIG S1-6 Cook's Distance: no cases approached 1

Expression 1

```
trainControl(method="repeatedcv", number=10, repeats=5,
             savePredictions = TRUE, classProbs = TRUE,
             search = "random", # reduce tree accuracy and ROC
             summaryFunction = twoClassSummary)
```

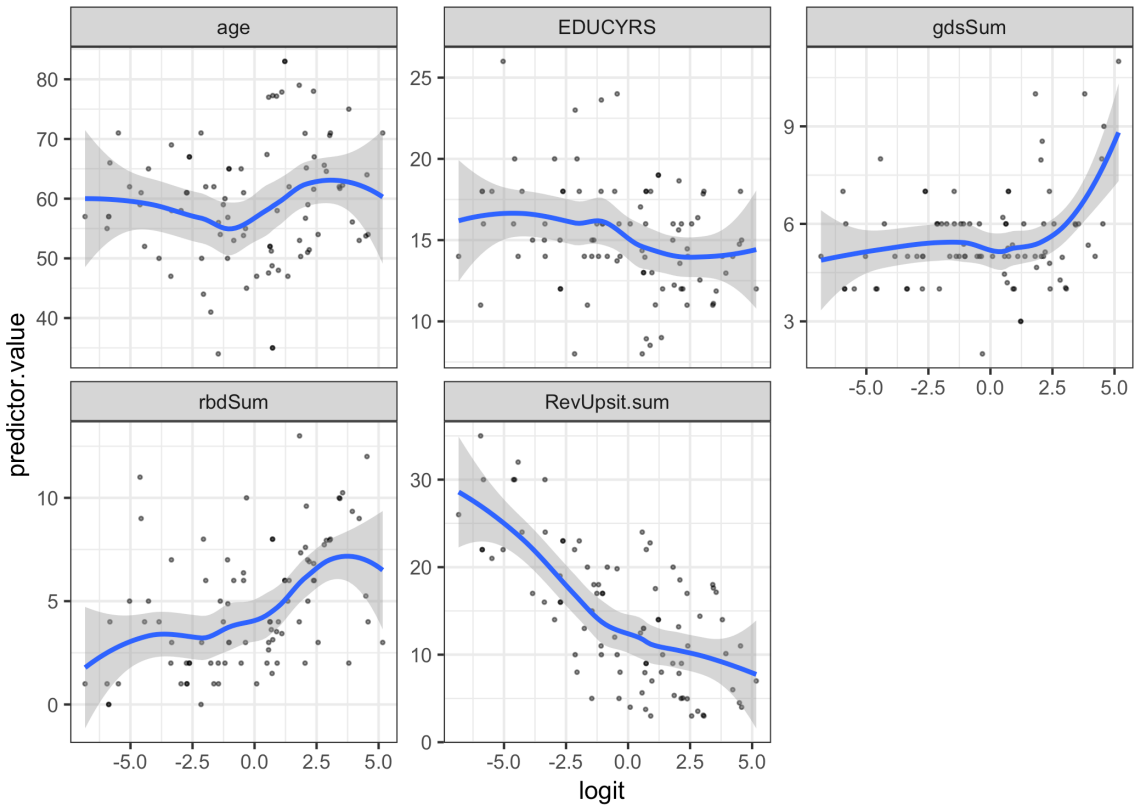


FIG S1-7 Linearity of logit scatterplots with smoothers, early PD/ SWEDD: gdsSum = Geriatric depression scale; MCATOT = Montreal cognitive assessment test; rbdSum = rapid eye movement behaviour disorder questionnaire; RevUpsit.sum = University of Pennsylvania smell test

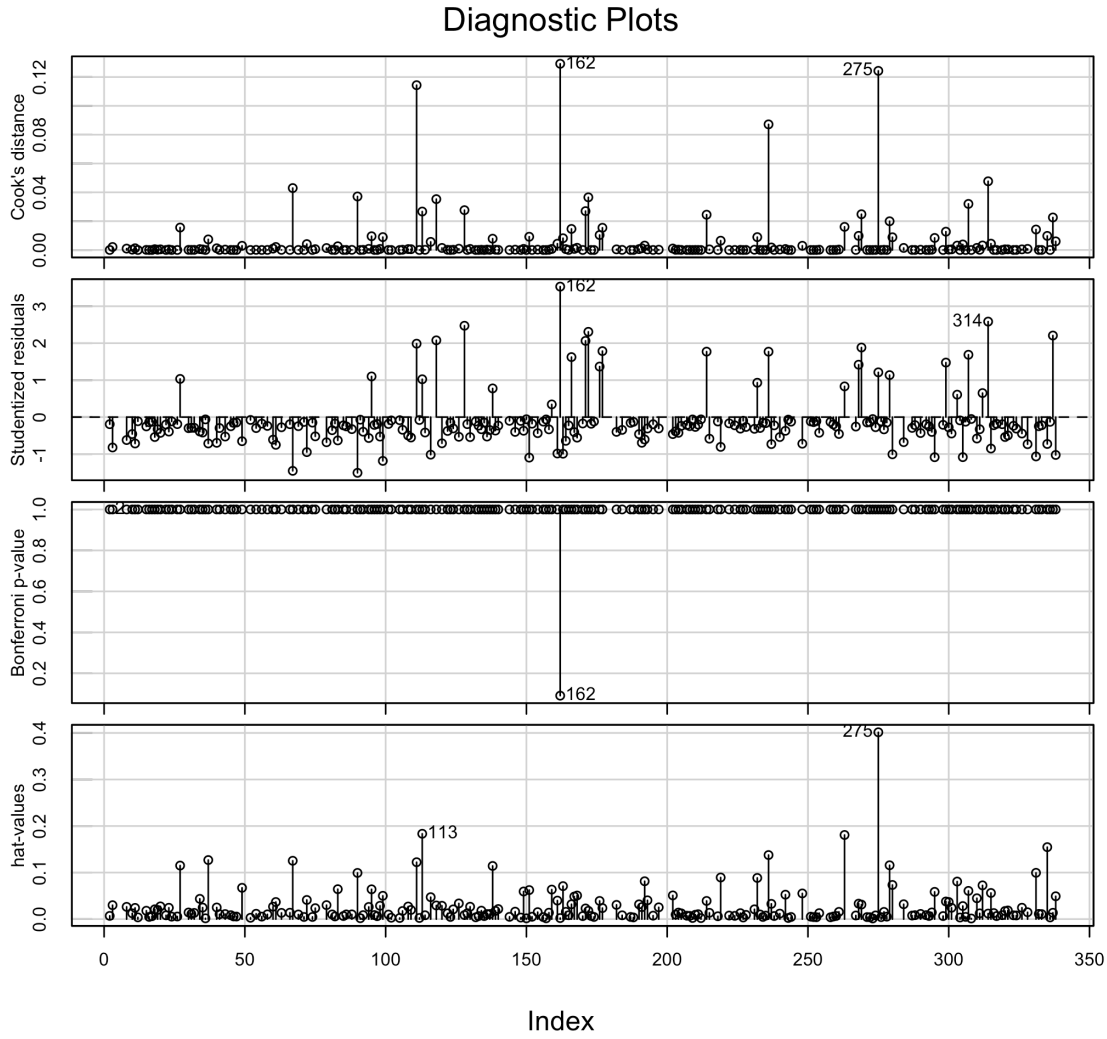


FIG S1-8: early PD/ SWEDD diagnostic plots

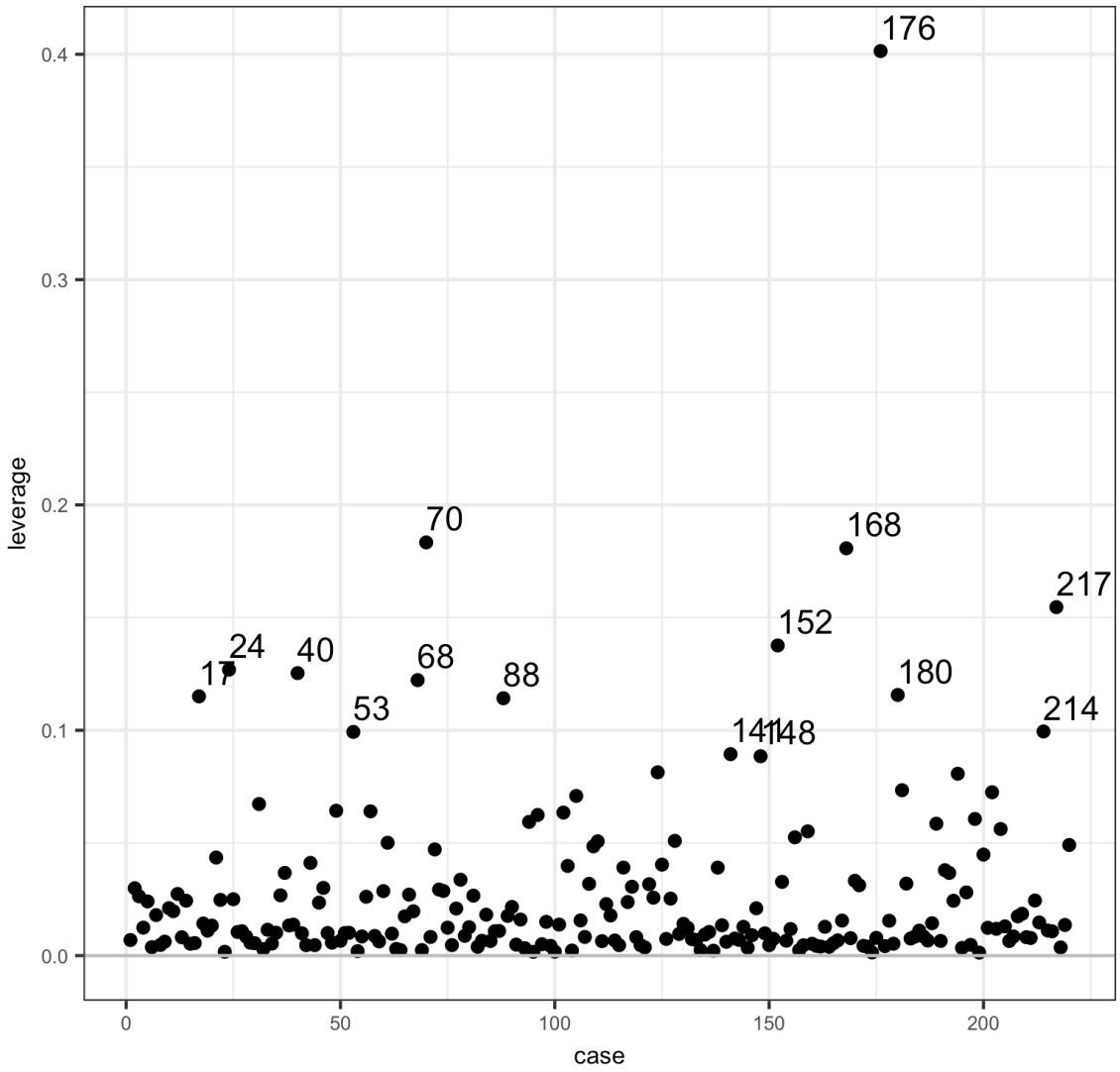


FIG S1-9: Leverage early PD/SWEDD: 7% (15/220) with > .082

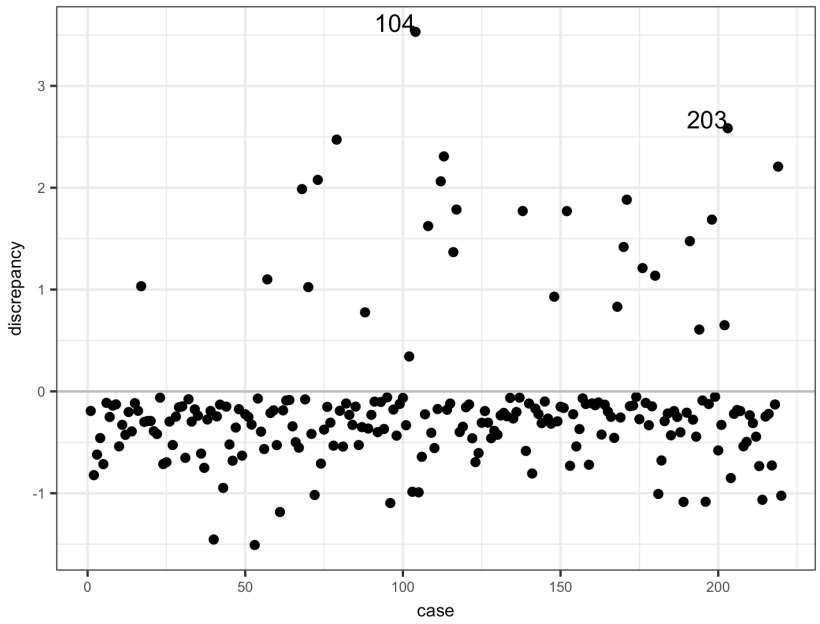


FIG S1-10: Discrepancy early PD/SWEDD

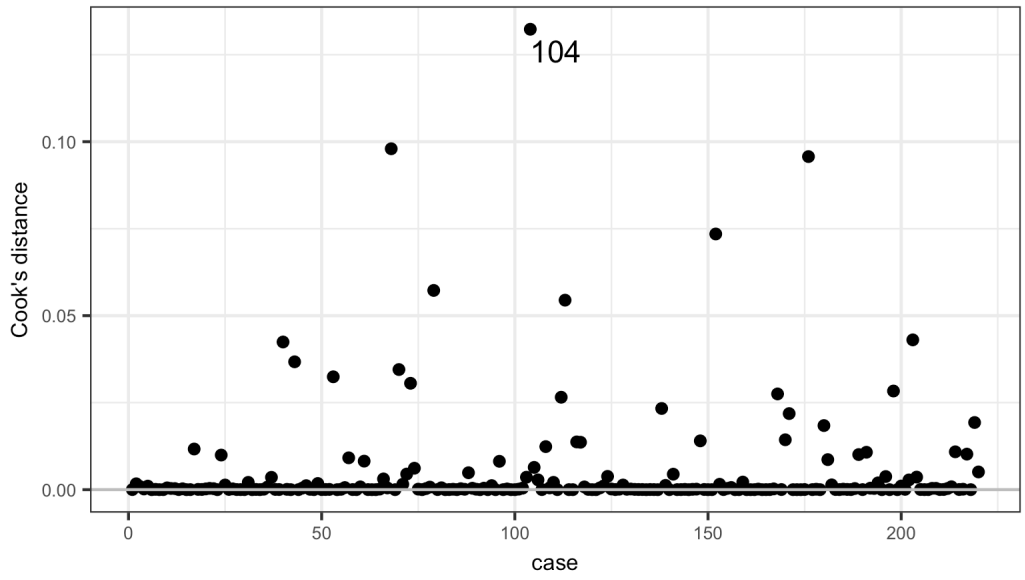


FIG S1-11: Cook's Distance early PD/SWEDD

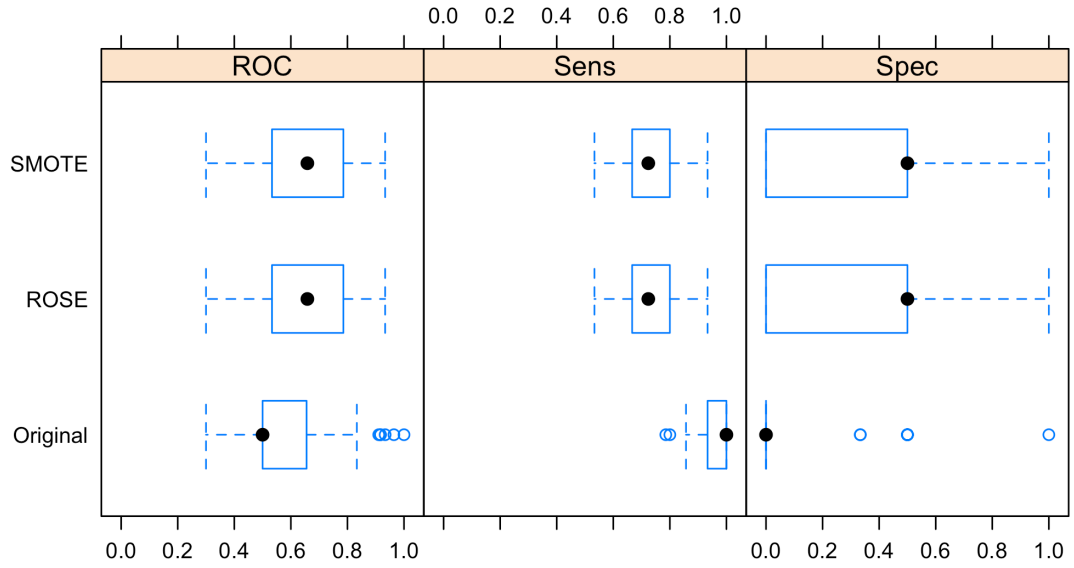


FIG S1-12: Decision tree: Early PD/SWEDD ROC AUC trained tree model original vs. subsampled data

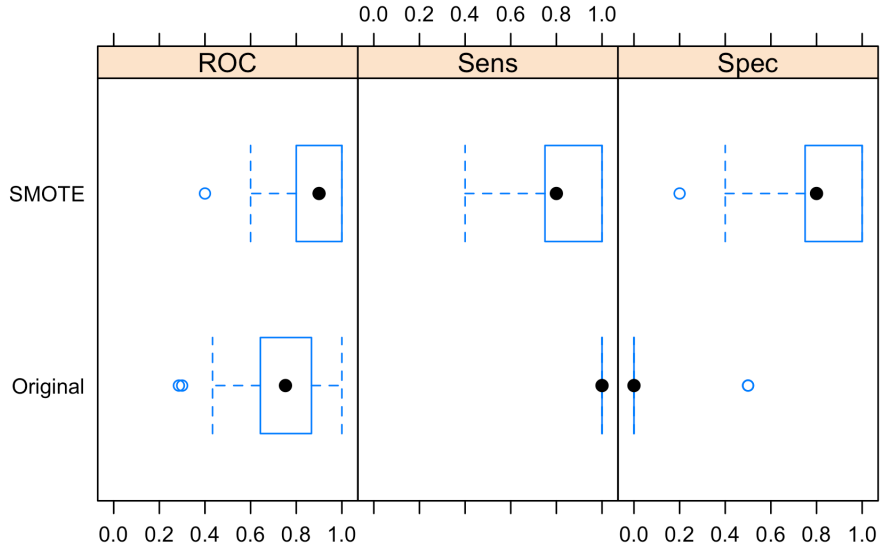


FIG S1-13 Random forest: Rndom forest models:, early PD/SWEDD., original vs. subsampled (SMOTE) data

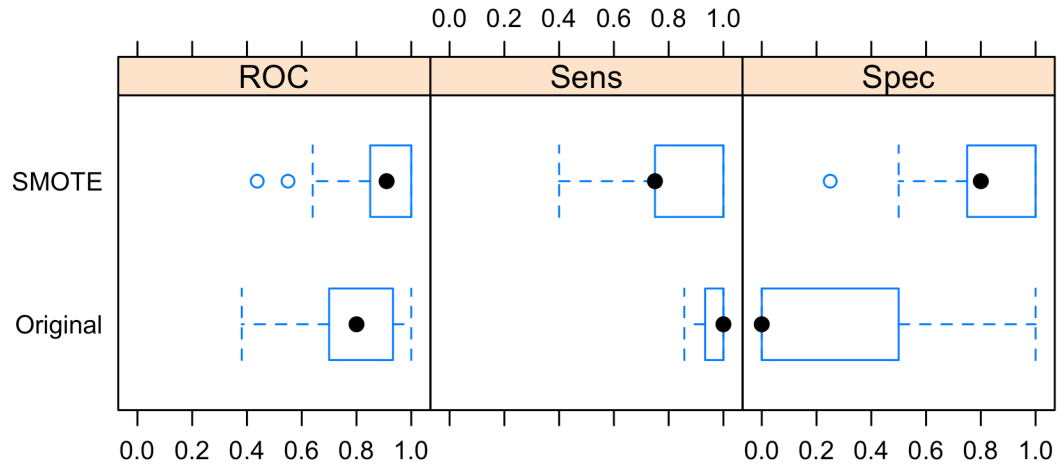


FIG S1-14 GLM: early PD/SWEDD ROC original vs. subsampled (SMOTE) data

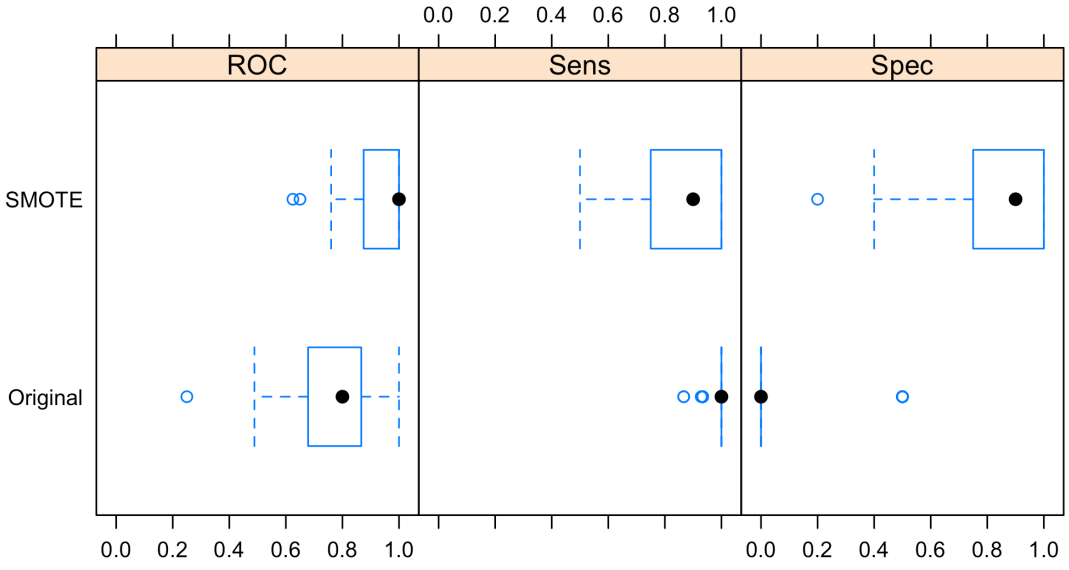


FIG S1-15: XGBoost: early PD/SWEDD ROC AUC XGBoost , original non-sampled vs. subsampled (SMOTE) data models

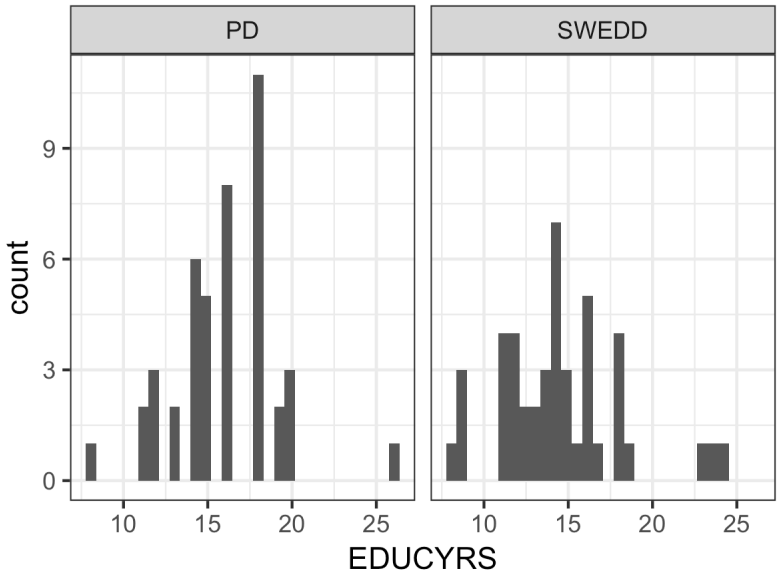


FIG S1-16: Early PD/SWEDD, years of education; EDUCYRS = years of education

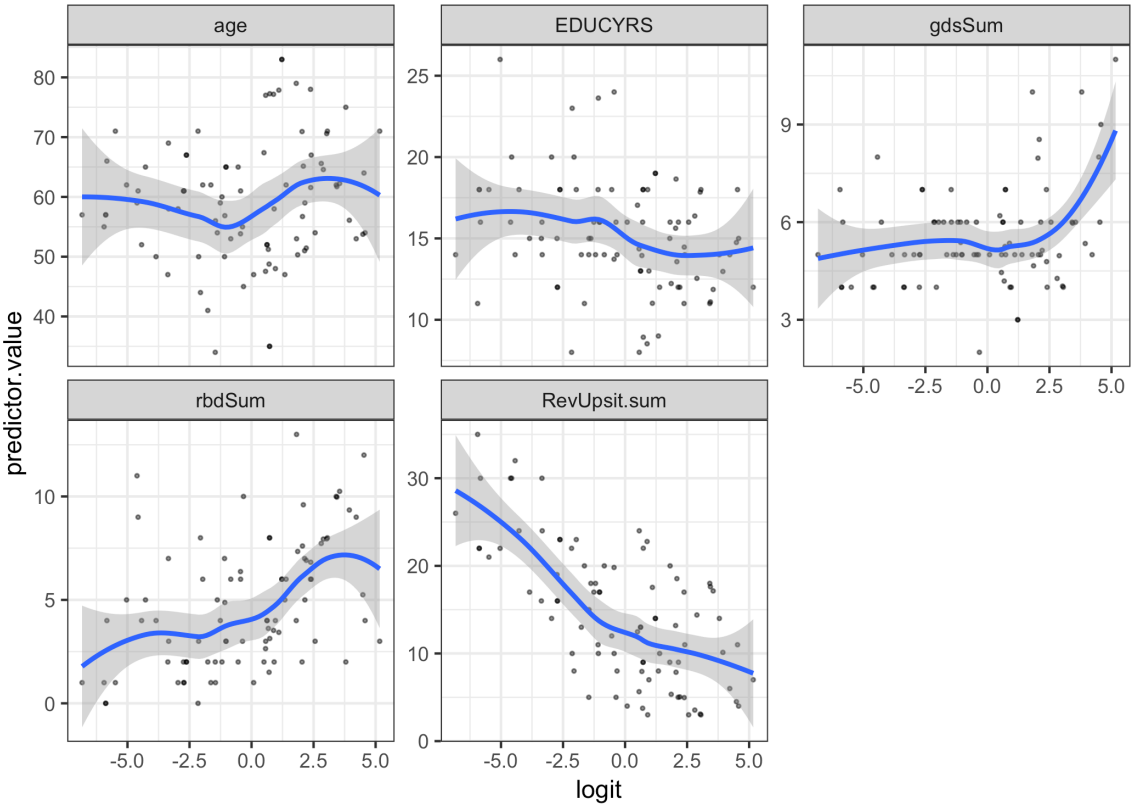


FIG S1-17 Linearity of the logit scatterplots with smoothers (early PD/SWEDD). rbdSum = rapid eye movement behaviour disorder questionnaire; RevUpsit.sum = University of Pennsylvania smell test; gdsSum = Geriatric depression scale; EDUCYRS = years of education

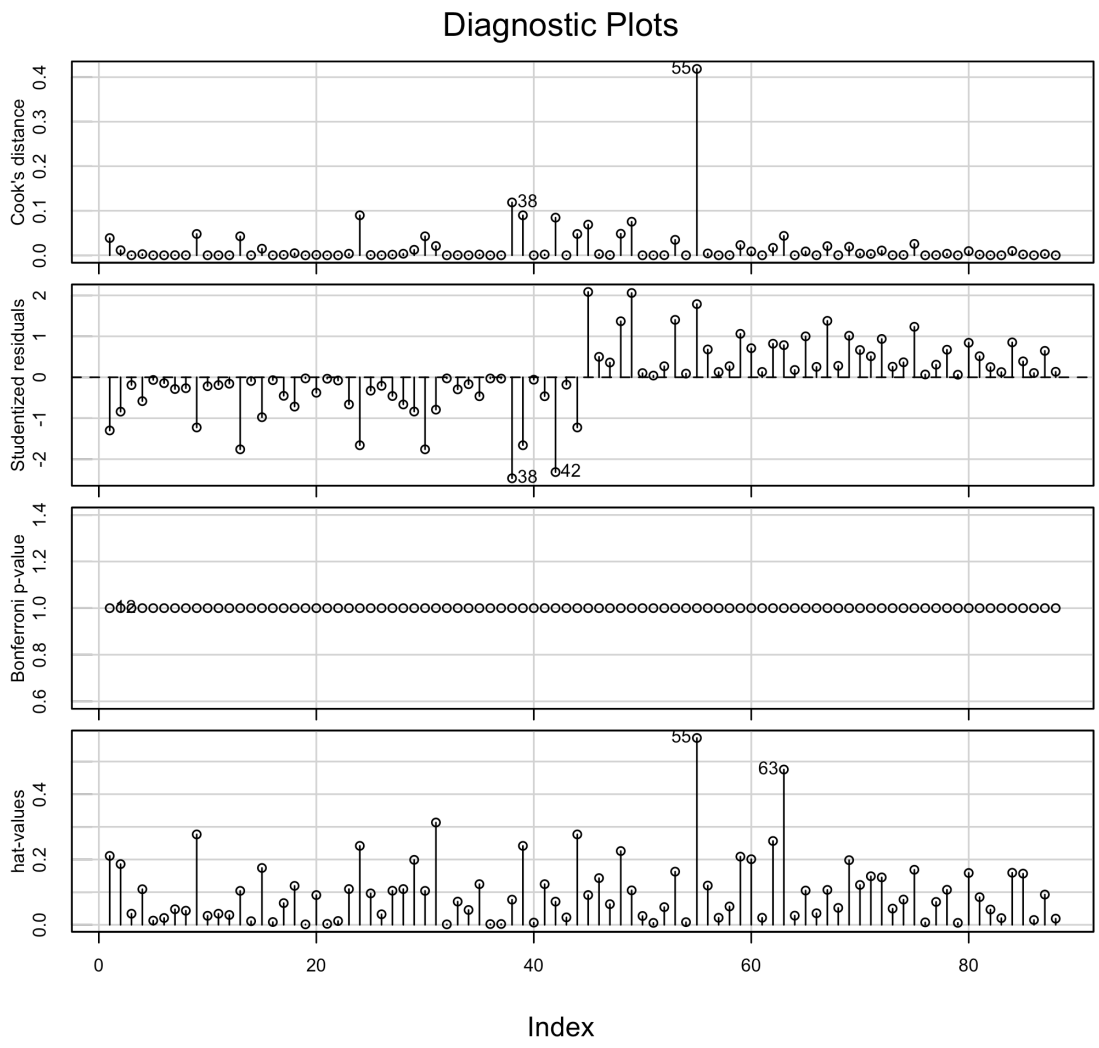


FIG S1-18: Composite: leverage (hat-values), discrepancy (studentized residuals), Cook's Distance values.

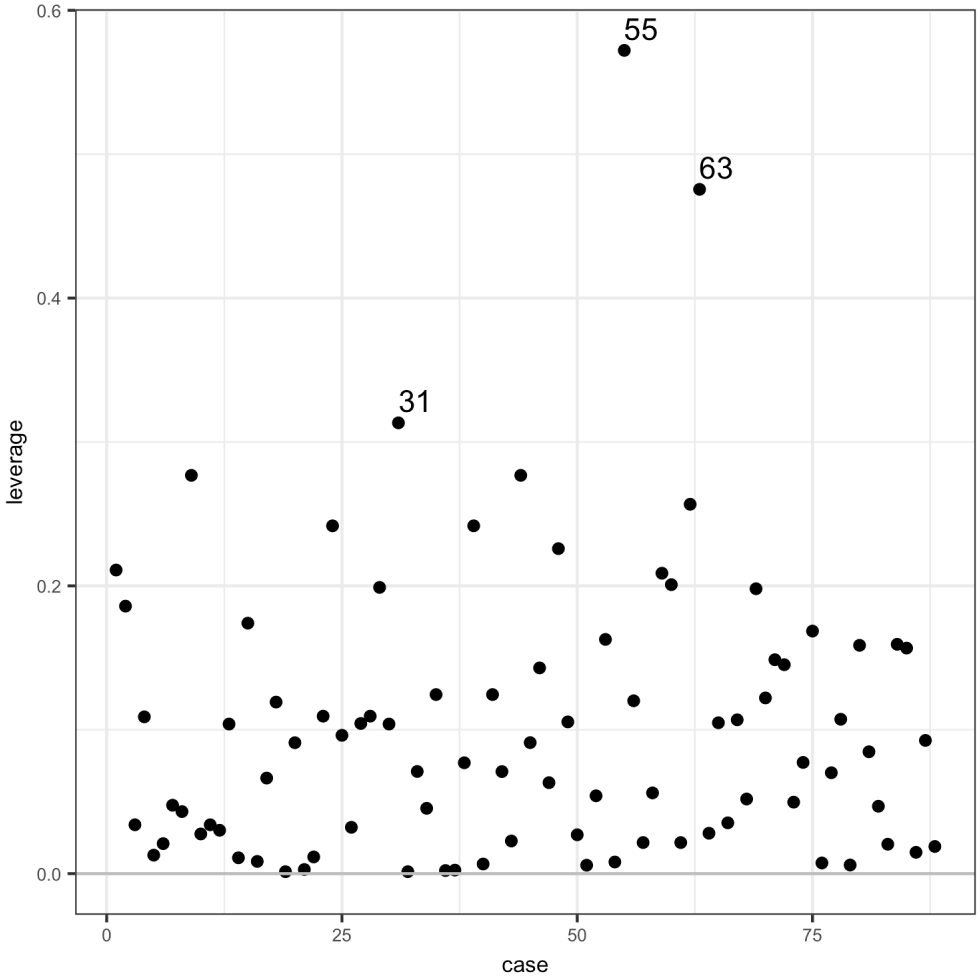


FIG S1-19: Leverage values; early PD/SWEDD

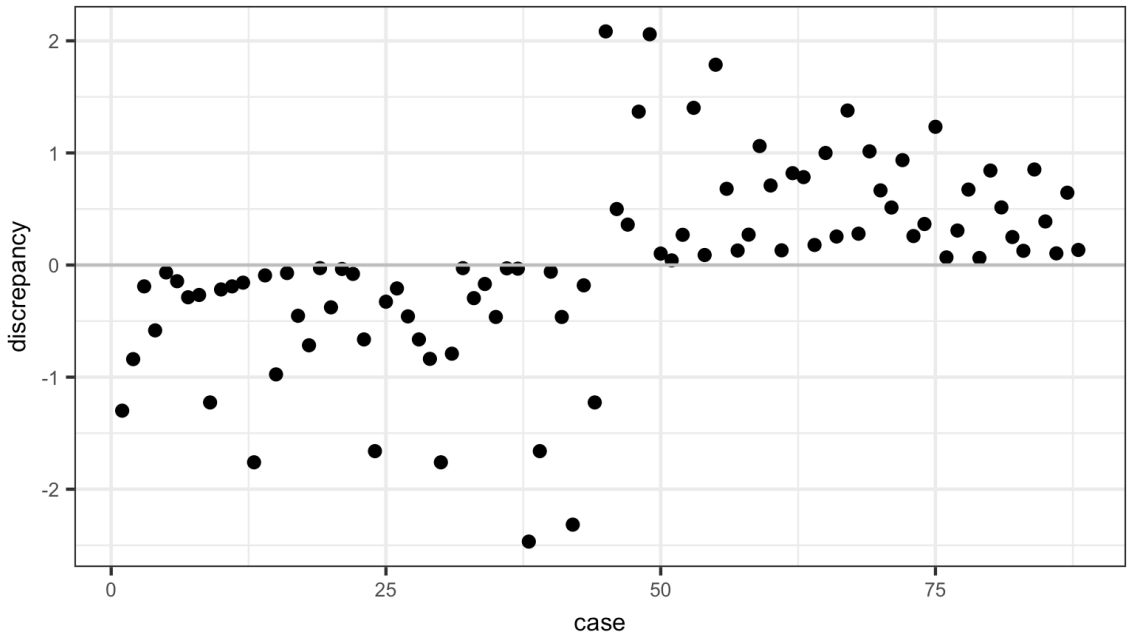


FIG S1-20 Discrepancy values; early PD/SWEDD

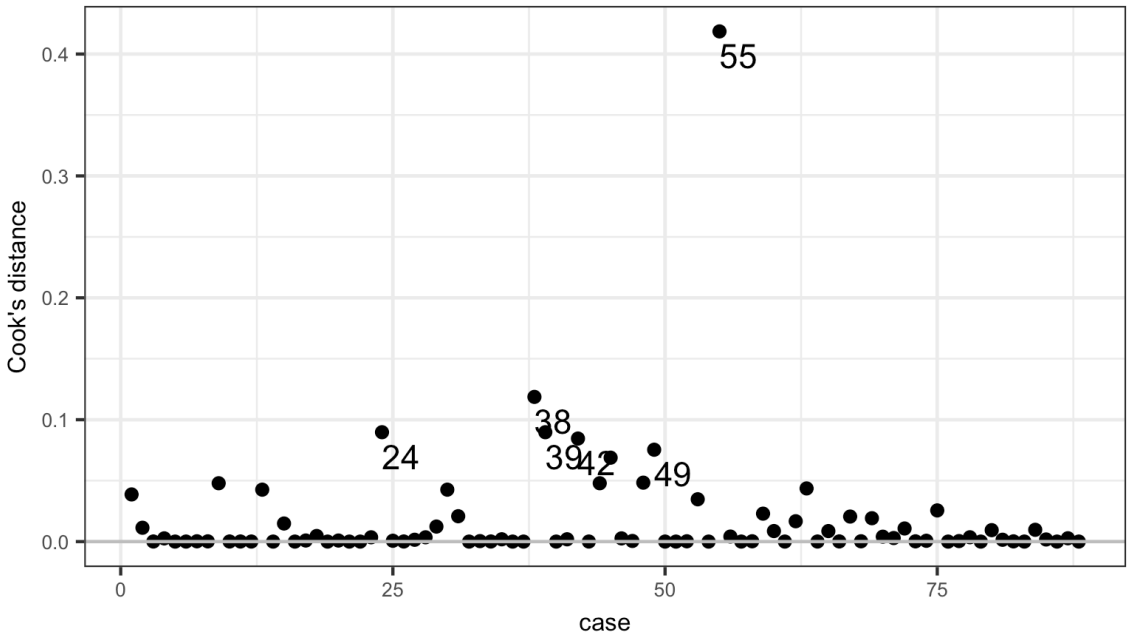


FIG S1-21 Cook's Distance; early PD/SWEDD

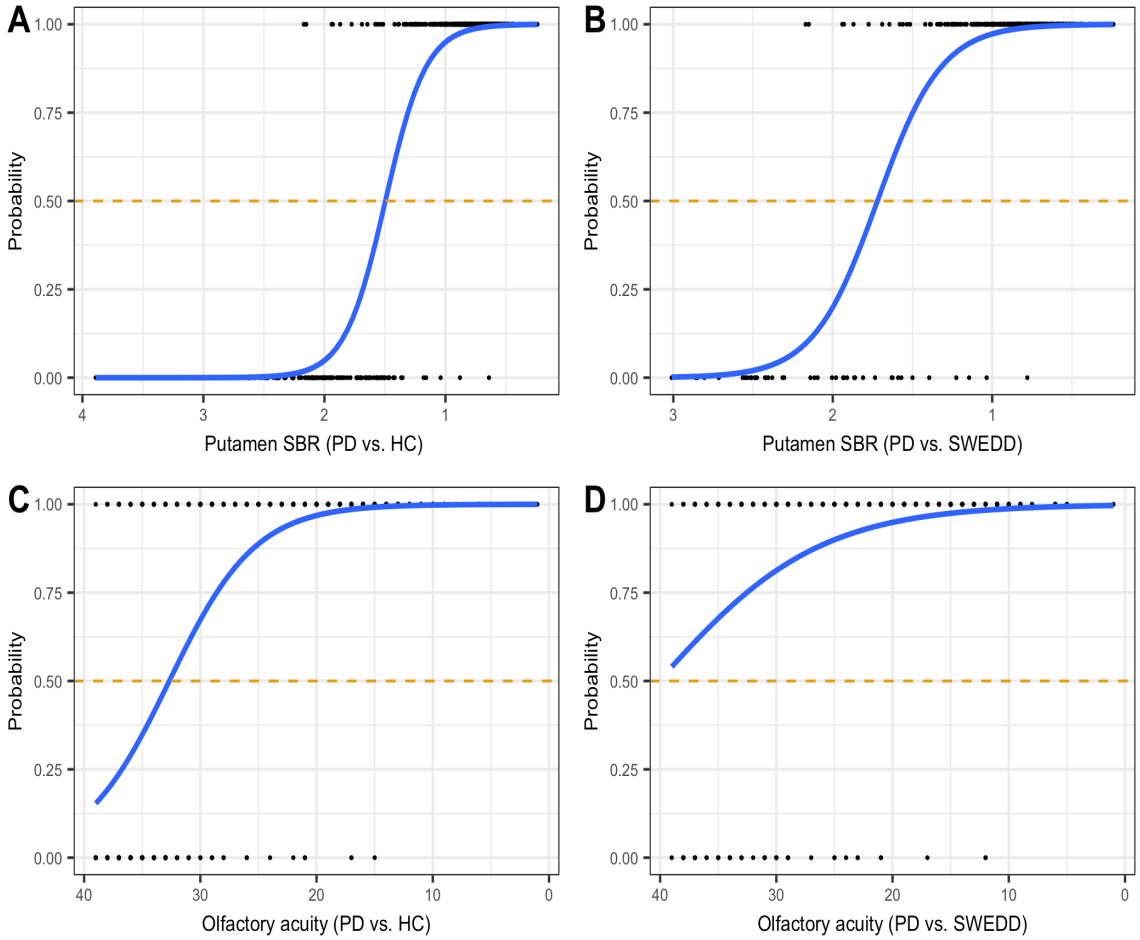


FIG S1-22 Logistic regression plots**: (A)** x-axis is DAT scan mean putamen values, y-axis proportion of cases who are early PD relative to controls; (**B)** x-axis is DAT scan mean putamen values, y-axis is proportion of cases who are early PD relative to scanswithout dopaminergic deficit (SWEDD); (C) x-axis is olfactory acuity based on reverse scaled UPSIT, y-axis is proportion of cases who are early PD relative to controls; (D) x-axis is olfactory acuity based on reverse scaled UPSIT, y-axis is proportion of cases who are early PD relative to SWEDD.

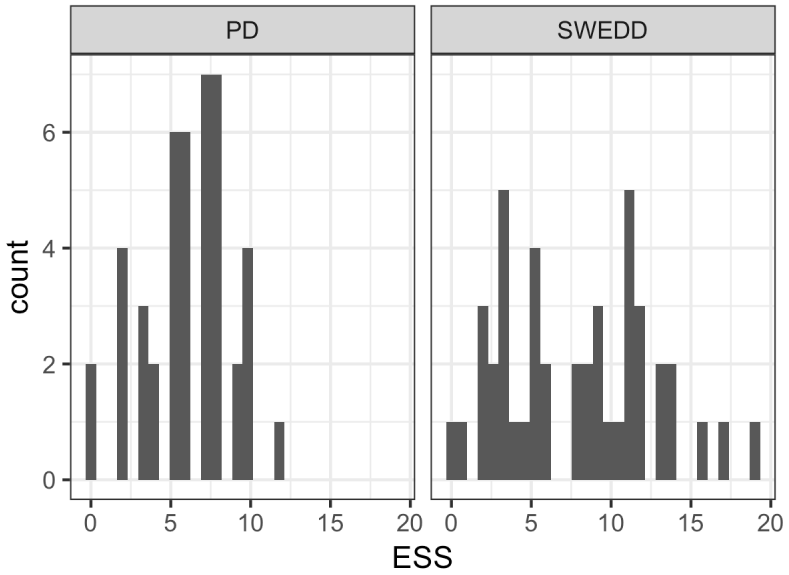


FIG S1-23 early PD/SWEDD Epworth sleepiness scale; ESS = Epworth sleepiness scale

Supporting Information II

Table S2-1: Early PD/controls logistic general additive model (GAM). Diagnostics used were gam.check and qq.gam from the mgcv package (Wood et al,. 2018).

Table S2-2: Early PD/SWEDD logistic general additive model (GAM). Diagnostics used were gam.check and qq.gam from the mgcv package (Wood et al,. 2018).

FIG S2-1: Early PD and controls qq.gam plot

FIG S2-2: Early PD and controls scatterplot, logistic GAM

Early PD/SWEDD: one data-set (internal evaluation using k-fold resampling)

Table S2-2: Early PD and SWEDD logistic general additive model (GAM). Diagnostics used were gam.check and qq.gam from the mgcv package (Wood et al,. 2018).

FIG S2-3: Early PD and SWEDD qq.gam plot

FIG S2-4: Early PD SWEDD scatterplot. Logistic GAM

TABLE S2-1

Family: binomial
Link function: logit

Formula:
gam(ENROLL_CAT~ s(age, bs="tp", k=3) + s(RevUpsit.sum, bs="tp", k=6) +

```
    s(rbdSum, bs="tp", k=4) + s(MCATOT, bs="tp", k=5) +
    s(NP1CNST, bs= "tp", k=4) + s(pTau, bs= "tp",k=3),
    data= LR1_rs1,
    method = "REML",
   family= binomial(link= 'logit'))

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    6.287      3.214   1.956   0.0504

Approximate significance of smooth terms:
                    edf   Ref.df  Chi.sq  p-value
s(age)            1.000    1.000   4.339   0.0372 *
s(RevUpsit.sum)   2.160    2.669  53.537 1.42e-11 ***
s(rbdSum)         1.000    1.000   4.703   0.0301 *
s(MCATOT)         2.739    2.945   3.577   0.2875
s(NP1CNST)        1.763    2.097   5.394   0.0749 .
s(pTau)           1.000    1.000   3.825   0.0505 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.585   Deviance explained = 55.9%
-REML = 90.787  Scale est. = 1,  n = 298
```

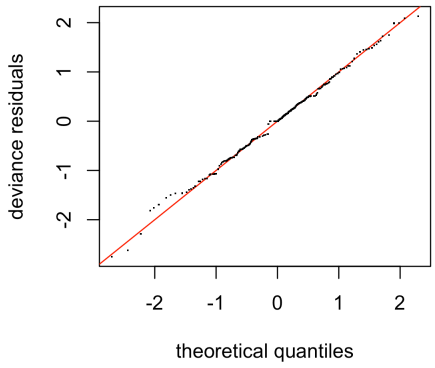


FIG S2-1: Early PD and controls, qqplot (qq.gam)

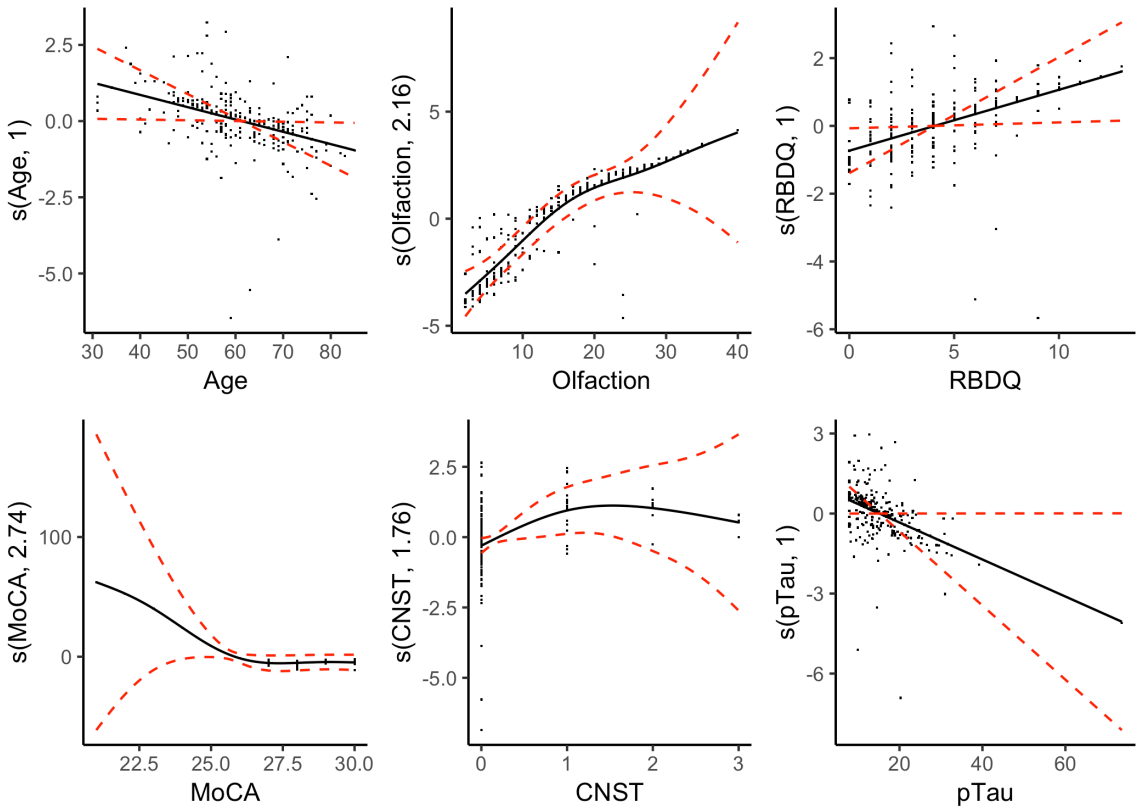


FIG S2-2 Early PD/controls logistic GAM scatterplots: The solid line is the predicted value of the dependent variable as a function of the x axis, where s(x) is fit as the outcome as a smooth of x. The dashed lines are plus-or-minus two standard errors. The y-axis is in the linear units, here logits, centered on 0 (50/50 odds), and included both positive and negative values (for graph function details see https://cran.r-project.org/web/packages/mgcViz/vignettes/mgcviz.html)

TABLE S2-2: Early PD and SWEDD, logistic general additive model (GAM)

Family: binomial
Link function: logit

Formula:
ENROLL_CAT ~ age + RevUpsit.sum + rbdSum + s(EDUCYRS, bs = "tp", k = 7) + gend + gdsSum

Parametric coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-10.20406	3.43193	-2.973	0.002946 **
age	0.13050	0.04095	3.187	0.001438 **
RevUpsit.sum	-0.40482	0.09974	-4.059	4.94e-05 ***
rbdSum	0.57431	0.16794	3.420	0.000627 ***
gend	2.63400	1.13550	2.320	0.020358 *
gdsSum	0.59875	0.31026	1.930	0.053623 .

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:

```
                edf    Ref.df   Chi.sq p-value
s(EDUCYRS) 4.119  4.912  14.39  0.0126 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.635   Deviance explained = 62.3%
-REML = 33.054  Scale est. = 1     n = 88
```

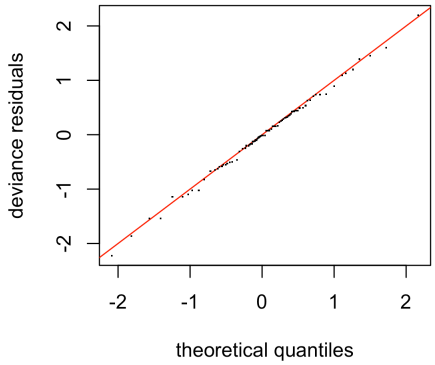


FIG S2-3 Early PD and SWEDD, qqplot (qq.gam)

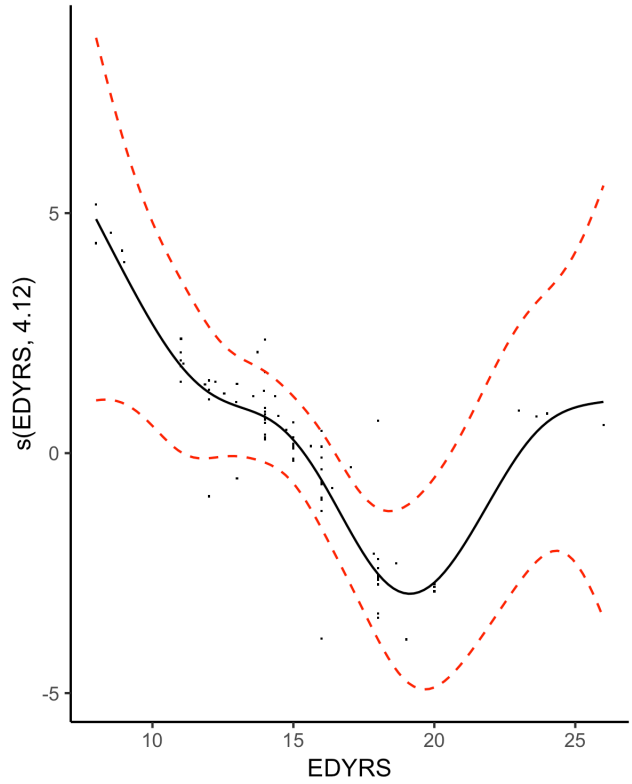


FIG S2-4 Early PD/controls logistic GAM scatterplot (years of education): EDYRS = years of education**.** The solid line is the predicted value of the dependent variable as a function of the x axis, where s(x) is fit as the outcome as a smooth of x. The dashed lines are plus-or-minus two standard errors. The y-axis is in the linear units, here logits, centered on 0 (50/50 odds), and included both positive and negative values (for graph function details see https://cran.r-project.org/web/packages/mgcViz/vignettes/mgcviz.html)

Supporting information III

Predictive classification model evaluation metrics

Error in context of predictive classification models is generally not expressed as MSE, RMSE or R^2; such measures are appropriate for linear regression (which has a quantitative response variable) but typically not appropriate for the qualitative response outcome in classification models (M Kuhn, 2013). Examples of binary (dichotomous two-class) qualitative response outcomes include: yes or no, positive or negative, class 1 or class 2, 1 or 0. In binary classification problems, error is generally the proportion or fraction of model misclassifications. This is often summarized in a confusion matrix, which specifies the number of cases (data instances) correctly classified (predicted) by the model and the number of cases incorrectly classified. In the 2 x 2 confusion matrix immediately below (SIII Table 1), the numbers on the diagonal represent the number of cases correctly classified by the model (262); the other values are errors, model misclassifications.

		Observed	
Predicted		HC	PD
	HC	*73*	18
	PD	18	*189*

HC= health controls; PD = early Parkinson's

SIII Table 1

Model accuracy, or error, is then simply the proportion of correct classifications relative to all predictions. Here this would amount to about 88% accuracy (262/298). However, as discussed below, accuracy can be a misleading indicator. The values in the above confusion matrix were taken from the early PD versus controls general additive model (GAM) used in the current research.

Perhaps confusingly, logistic regression is an example of an algorithm that has properties of both a classification model (it has a binary, dichotomous outcome) and also a quantitative and parametric model; it estimates the probability (with a 0-1 range, which is certainly not dichotomous) of a data instance being a case, the class of interest. Generally, in a binary classification model, the class or category of interest (e.g. Y=1, Y=positive, Y= yes) is

decided at the outset. A classification model then predicts the probability of being a member of the class of interest based on information from the predictors (features).

Another error or loss term used quite often is Log Loss. A well-endorsed intuitive explanation of Log Loss is provided by Stack Exchange (Exchange, 2017) for balanced data, where in binary (two-class) classification and pr_i is the predicted probability attributed to the correct class, $LogLoss = -\log(\mathrm{pr}_i)$. A $\mathrm{pr}_i = 1$ is ideal and indicates the model attributed 1 to the correct class resulting in a $LogLoss = 0$. Antithetically, a $LogLoss(\mathrm{pr}_i) = +\infty$ indicates that the actual class was incorrectly attributed a probability of 0 by the model. In general, a lower Log Loss value indicates more accurate model prediction.

A predictive model generates the probabilities that indicate the extent of predicted case membership, which, in the medical field context, could be indicative of membership either in the presence of disease (P) or absence (1-P) of disease groups. Moreover, in R, the "predict()" function facilitates comparison of predictions among different types of models. It will create an object that will reveal model predictions, and it can be used with virtually any model. The "predict()" function is typically used in creating a confusion matrix, and simplifies application of a model to new data. The classification numbers in a confusion matrix (as in the 2 x2 matrix above) depend on the classification threshold. A threshold is set, typically 0.50 (a 50% cutoff to discriminate the predicted class); statistical software normally defaults to a 0.50 threshold. Using the medical field context again, cases below this threshold are predicted as members of the disease absence group; cases above the threshold are predicted as positive, and hence members of the group with disease. The continuous probabilities (or, alternatively, discrete predicted class labels) produced by a model's binary classification allows calculation of four different fractions of a 2 x 2 confusion matrix $\begin{matrix} TN & FN \\ FP & TP \end{matrix}$. As in the SIII Table 1, the numbers on the diagonal are collectively the correct predictions and are made up of true positive (TP) and true negative (TN) values. TP is the fraction is the proportion of cases correctly predicted as with disease; TN is the fraction of cases correctly predicted without disease. The false positive fraction (FP: equivalent to Type I error or commission error) is the proportion of cases incorrectly predicted as diseased; and the false negative (FN: equivalent to Type II error or omission error) fraction is the proportion of cases with disease incorrectly predicted as healthy or without disease.

From the SIII Table 1, it can also be seen that the total actual observed number of controls is 91 (73+18) and the total actual observed number of early PD is 207 (18+ 189). If a model threshold other than 0.50 is selected it will alter confusion matrix values. As discussed shortly, the default 50% threshold is typically less than optimal.

Binomial classification model performance is often assessed with the receiver operating characteristic (ROC) area under the curve (AUC). Binary transformed probabilities from TP, TN, FP, FN fractions are used to derive performance measures including accuracy (briefly discussed above), Cohen's Kappa, sensitivity, and specificity. Cohen's Kappa (Cohen, 1960) or simply the Kappa statistic, was originally used to compare agreement between experiment raters, but it also measures the performance of a model, its accuracy, relative to random chance. It can be expressed as $K = \frac{O-E}{1-E}$, where O is observed accuracy and E expected accuracy, the latter referring to random chance. In general a higher Kappa is better: 0.41-0.60 is rated as moderate and a Kappa >.75 as excellent (Landis & Koch, 1977), though there is currently not a universally accepted scale for interpretation of the Kappa statistic. Despite wide spread usage, the Kappa statistic has earned sharp criticism from more than a few sources. Reported problems include, but are not restricted to, redundancy (due to high correlation) with the general accuracy measure, a basis on random chance probabilities that are not random, and underestimation of the probability of correct classification (Foody, Campbell, Trodd, & Wood, 1992; C. R. Liu, Frazier, & Kumar, 2007; Olofsson et al., 2014; Pontius & Millones, 2011).

Accuracy is often confused with the ROC AUC statistic (often referred to a simply the AUC), but accuracy and the AUC metrics differ. Accuracy, otherwise known as diagnostic accuracy, again, is simply the number of correct predictions a model or classifier makes expressed as the proportion of correct predictions over all predictions. To elaborate, accuracy considers the predicted true negative (TN), true positive (TP), false negative (FN) and false positive (FP) values that result from an algorithm. In the context of binary classification, accuracy (A) can be derived from the 2 x 2 confusion matrix illustrated above and can be expressed as $A = \frac{TP+TN}{\sum(TP,TN,FP,FN)}$. In short, diagnostic accuracy is the proportion of correct predictions made by the classifier, and it imposes a cutoff threshold of > 0.50 (Field, 2012). The accuracy of the (GAM) model as revealed in SIII Table 1 was as already conveyed about 88%. It warrants note, that accuracy is not reliable when classes are severely

imbalanced. Consider a fictitious example of TP = 94 (with disease), TN =2(controls), and FP = 2, FN =2. In such a circumstance, accuracy would be 96% (TP+ TN/TP+TN+FP + FN), and controls would have a nil recognition rate, meaning accuracy is largely reflecting the underlying class imbalance. Moreover, it is underlined that the values in the confusion matrix and hence accuracy, as well as Cohen's Kappa, sensitivity, and specificity, depend on the class discrimination threshold.

The AUC (see Graphic 1 below), by contrast, is regarded as invariant to class skew (Fawcett, 2006), and it is not altered by a given threshold. With respect to classification threshold, the AUC combines both sensitivity and specificity. Paired points of sensitivity, plotted on the y-axis and specificity, plotted on the x-axis, occur across all cutoff thresholds. Each paired instance of sensitivity and specificity corresponds to a ROC space single point (Fawcett, 2006). With regard to class imbalance or skew, the AUC attributes equal weight to sensitivity (the percentage of correctly identified positives) and specificity (the percentage of correctly identified negatives). Sensitivity and specificity are discussed further below.

In a model with poor accuracy the AUC may fall on the diagonal, which reflects only a 50% chance of correct classification; the ideal model will have an AUC of 1. ROC curves falling in closer proximity to the upper left hand corner reflect more accurate models: AUC values of .60 - .70 reflect relatively poor models while AUC values of .90 and up indicate a high degree of model classification separation. The general interpretation the AUC, say of .90, would be that the model has a 90% chance to discriminate between the class designated as positive (the with disease group typically) and the class designated as negative (typically the group without disease) (Hanley & McNeil, 1982).

Sensitivity (also known as recall) is the true positive rate (TPR), also referred to as the true positive fraction (TPF): the percentage of events/cases with pathology correctly identified as positive for the presence of pathology. Specificity is the false positive rate (FPR; type I error rate), also referred to as the false positive fraction (FPF); the percentage of non-events/non-pathology cases correctly identified as negative and without pathology. In a ROC curve, there is a trade-off between the sensitivity and specificity. If sensitivity is high (perhaps set by a predetermined threshold) specificity will lower; if specificity is high sensitivity will be lower. Each point in ROC space then, is a function of a given threshold or cutoff. Importantly, as already noted, the AUC

statistic is invariant to threshold change; different thresholds can change sensitivity and specificity but will not alter the AUC measure of model performance. Specifically, a different classification threshold (e.g. .40 rather than .50) alters confusion matrix $\begin{matrix} TN & FN \\ FP & TP \end{matrix}$ values but not the AUC.

As distinct metrics, sensitivity and specificity offer insight to model classification error. Again, specificity is the model's percentage of correctly identified negatives (non-events; typically healthy controls). A model with high specificity has a lower type I error (or false positive rate). Specificity (*SPE*) can be expressed as $SPE = FPR = \frac{FP}{FP+TN}$ where *FP* is the number of false positives and *TN* is the number of true negatives, and FPR is the false positive rate. Sensitivity, as just defined, is the true positive rate (TPR), the model's percentage of correctly identified positives (events, or those with disease). Sensitivity can be expressed as $SN = \frac{TP}{TP+FN} = 1 - FNR$, where *SN* is sensitivity, *TP* the number of true positives, and *FN* is the number of false negatives. Given the true positive, false negative, false positive and false negative measures from which sensitivity and specificity are derived, it is not surprisingly that error, in the context of a confusion matrix, can be couched in terms of sensitivity and specificity (M Kuhn, 2013).

Critical to the sensitivity and specificity indices is selection of the appropriate cut-off. Often the .50 classification cut-off has insufficient sensitivity to detect an event or case of pathology. Or, antithetically, specificity may be poor resulting in a low correct identification percentage of those without pathology. Of course, to eliminate unnecessary diagnostic procedures, a diagnostic test that is both highly specific (to rule out disease) and sensitive (to identify disease) is ideal. While achieving this may often not be possible, estimated threshold cutoffs based on maximal sensitivity and specificity are in aid of such predictive disease screening. One approach is to calculate a balanced threshold; a threshold that maximizes both sensitivity and specificity across all classification cut-off points to arrive at an optimal cut-off threshold. This is known as the Youden's index (Youden, 1950b). Youden's *J* index can be expressed as: $J = SN + SPE - 1$. Various software offerings provide a Youden Index option for ROC AUC analysis. Another method of finding an appropriate cut-off value is to use the base rate of a given disease or event's prevalence in the population (Habibzadeh, Habibzadeh, & Yadollahie, 2016). Should there be a need to prioritize identifying disease, a cut-off with relatively high sensitivity, say > .80

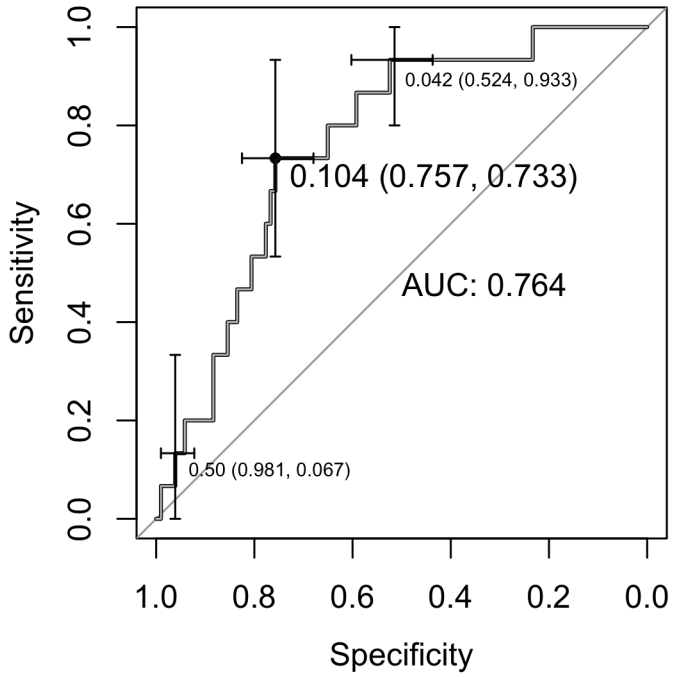
(which lowers the false negative rate or type II error) would be effective for screening disease as true positive cases are less likely to be missed. However, prioritizing sensitivity comes at the expense of specificity (Habibzadeh et al., 2016). Of note, deriving a cutoff threshold from a training or validation data-set post hoc can result in an overly optimistic model performance estimate (Ewald, 2006). Ideally, the cutoff threshold should be estimated from an evaluation data-set not used in training or testing the model (M Kuhn, 2013), though often alterative data sources for this approach are not available.

It is important to underline that accuracy – a much used performance measure- is dependent not only on the classification threshold but on the frequency of each class. Antithetically, and as already discussed, the AUC is unaffected by imbalanced data (Fawcett, 2006; Provost, 1998). Indeed, it can be seen that the AUC TP and FP rates are ratios independent of class prevalence (Fawcett, 2006). Consequently, and especially in the context of imbalanced data, the AUC statistic is regarded as the preferred indicator of performance (M Kuhn, 2013) and an index of model *global* accuracy.

Accuracy, as already noted, is sensitive to differential class distribution, particularly for large case instance discrepancy. Here is another fictitious example that further illustrates the difference between accuracy and metrics sensitivity and specificity. Sensitivity and specificity (integral to the ROC AUC), are not sensitive to class imbalance. For example, consider a circumstance with 200 positive observations (with disease) and 200 negative observations (controls) and confusion matrix values from some model of TP = 140, TN = 160, FP = 40 and FN = 60. Accuracy is given the expression $A = \frac{TP+TN}{\sum(TP,TN,FP,FN)}$, sensitivity (SN) is $SN = \frac{TP}{TP+FN}$ and specificity (SPE) is $SPE = \frac{TN}{TN+FP}$. These metrics then amount to the following: Accuracy or A =. 75 (75%.), SN = .70 and SPE = .80. If the negative samples increased such that TN and FP rates increased to 1600 and 400 respectively, accuracy is altered to .79, but sensitivity and specificity remain unchanged. However, if the threshold at which class labels are predicted is altered then accuracy as well as both sensitivity and specificity will be impacted (Freeman & Moisen, 2008; Tharwat, 2018).

Consider a AUC used to determine the predictive classification accuracy of a trained model on validation or test set data. While the AUC is threshold-invariant, changes in threshold can be readily demonstrated to, in

turn, change sensitivity (the number of positive cases correctly identified) and specificity (the number of negative cases correctly identified). Typically, a decrease in threshold increases sensitivity and returns more positives, while an increase in threshold heightens specificity and returns more negatives. In short, specificity and sensitivity are inversely proportional. This is illustrated ROC AUC Graphic 1 below.



GRAPHIC 1 | The ROC curve. Sensitivity = TP/(TP+FN); Specificity = FP /(FP and TN); TP = true positive; FN= false negative, FP = false positive; TN = true negative

The point (0,0), in the lower left corner, corresponds to a threshold at which all cases are classified as negative; point (1,1), in the upper left corner, corresponds to the threshold at which at all cases are positive. At the default 0.50 threshold, (error bars) near the lower left corner, sensitivity is only about 7% but specificity is 98%. By contrast, at the threshold of less than 5% (.042), the upper-most set of numbers, sensitivity is 93% but specificity is reduced to 52%. The best balance of sensitivity and specificity with this data occurs at the threshold of .104, which is a Youden Index (Youden, 1950b) optimized point that minimizes the mean positive and negative error rates while maximizing the sum of specificity and sensitivity. At a threshold of .104, sensitivity is 73% and specificity is 76%. The error bars in the above graph reflect variation in specificity and sensitivity.

Cross-validation

The litmus test of a trustworthy model's accuracy is performance on data "unseen" by the existing model; data on which the model was not trained – a cross-validation paradigm. Comparable results on training and evaluation data-sets indicate the model will generalize well (Tabachnick, 2007). Ideally, model accuracy should be tested on a new data-set (from the same population), though availability of such entirely new samples is quite infrequent. Alternatively, a single data-set can be partitioned (prior to pre-processing and feature selection) into separate train and test/validation set portions. In this quite common circumstance, the model is trained on a train partition subset, typically 60-80% of the data, and then the model is tested on the remaining test/validation partition "unseen" by the model during training.

One method of splitting the data into training and test or validation subsets is simple random sampling. Another approach is stratified random sampling, which in the context of a classification analysis preserves proportions within classes by sampling within classes (e.g. as achieved by the "createDataPartition" function in the caret package function) (M. Kuhn, 2019, March, 3). The training error rate (see *Evaluation metrics* above) and performance metrics (e.g. AUC, specificity, and sensitivity) are computed on the training data but of greatest interest is the performance outcome when the model is applied, in cross-validation fashion, to the unseen partition or new data. External and internal types of cross-validation can be distinguished. External cross-validation is the validation process described so far. Internal validation refers to cross-validation within the model itself using only a portion of the data allocated to train the model. The random forest, and and XGBoost model types are examples of algorithms with built-in, internal cross-validation whereby multiple models are internally tested on (hold-out) data not used to train a given model. Further, a model is often then externally cross-validated; tested on an unseen data partition or new data set. Internal cross-validation improves the models ability to generalize through a resampling method.

Resampling and internal validation

In general, resampling provides internal cross-validation by fitting the model on a subset of the data allocated to train the model. A portion of the training data is held-out for model performance testing. Such data resampling is an iterative process but how data subsets are selected differentiates resampling methods. K-fold and the bootstrap are commonly adopted resampling methods for cross-validation. In the k-fold resampling method

data is partitioned into approximately equal sized k-subsamples; one of the subsamples (the first fold) is held out for validation testing while the other k-1 subsamples are used to train or fit the model. It warrants emphasizing that class predictions are made on the validation or held-out data not on the data used to train a model. For example, if k=10, k-1 or 9 samples are used to train a model, 1 sample is held-out for validation on which predictions are computed; 90% of the data is therefore allocated to training and 10% is allocated to validation. Again, this is an iterative process, with many models being built allowing optimal hyper-parameters/tuning parameters to be selected. With each k-repetition a different segment of the observations is treated as the validation test data-set (then returned to the training data-set), which results in k-test error estimates of model predicative performance. A k-fold cross-validated (CV) estimate is the average of the k-test error estimates (Gareth, 2013).

The bootstrap differs. The model is built on random samples taken from the original data-set with replacement (Gareth, 2013), meaning that the same cases can be repeatedly used to build the model which can result in bias; greater that 63% of the same cases has been reported (M Kuhn, 2013). In addition some observations may not be allocated to the validation test data-set at all. By contrast, in the k-fold approach every observation will be used for both training and validation and each observation is allocated for validation not more than once. Moreover, though it is a point of debate (Bengio & Grandvalet, 2004), it has been contended that the k-fold technique (10-20 folds) reduces variance between (training) models relative to a method such as leave one out cross-validation (Y. Zhang & Yang, 2015). Leave-one-out CV is a particular application of k-fold CV where k is replaced by n (n= number of samples) and the model is fit on n-1 training data observations. With sufficient k-fold repetitions (e.g. k=10), similar variance and hence results have been reported for both k-fold and leave-one out methods (Molinaro, Lostritto, & van der Laan, 2010). While a single CV method will not be ideal for all data-sets, for smaller sample sizes acceptable bias and variance outcomes can be achieved using the 10-fold CV approach (M Kuhn, 2013).

Sampling methods

Sampling (also referred to as subsampling) techniques include down-sampling, up-sampling as well as hybrid approaches. In a very cursory description of subsampling methods, down-sampling data randomly samples classes and matches class proportions to the class with the lowest frequency of observations, though this could lead to loss of important majority class information. Up-sampling, in a largely antithetical procedure, randomly samples the minority or rare event class with replacement such that the minority class is the same

frequency as the class with the greater frequency of observations. Random over sampling examples (ROSE) (Menardi, 2014), and synthetic minority over-sampling (SMOTE) (Chawla, Bowyer, Hall, & Kegelmeyer, 2002) are hybrid techniques that simulate new minority class data points while also down-sampling the majority class. Both are over-sampling methods that create artificial samples but they have differing sampling methods. ROSE generates artificial samples from the predictor (feature) space of the minority class using a bootstrapping approach. SMOTE, in a k-nearest neighbours approach, synthesizes new minority instance samples somewhere along a line between an existing randomly selected minority instance and that instance's nearest (existing) minority instance neighbour. A number of studies have reported such subsampling methods allay data imbalance related issues (Batista, 2004; Burez, 2009; Jeatrakul, 2010; Van Hulse, 2007), though mixed results have been reported elsewhere (M Kuhn, 2013). Combined use of the both the ROC AUC performance metric and subsampling have been recommended to counter issues stemming from imbalanced data (Brownlee, 2015, August 19).

Collinearity and feature elimination

Models are impacted by the correlation structure among predictors and with the response variable. Models built using predictors with high collinearity (e.g. > *r* .75 between predictors) can obscure interpretation of results. Two highly correlated predictors may each have similar important relationships with the response variable. In regression (linear or logistic) the coefficients of highly correlated predictors have sizable overlap in contribution to the response variable; these coefficients are therefore not unique and difficult to interpret. Moreover, in models based on algorithms without collinearity issues (e.g. tree-based models), the importance of one predictor can be masked by another predictor with which it is highly correlated, which can lead to an inaccurate conclusion that one of these predictors is unimportant (Exchange, 2018, August 3; Field, 2012; M Kuhn, 2013).

Under the rubric of feature (predictor) elimination however, high correlation among predictors is not a concern; the objective is not interpretation of the data but simply to reduce irrelevant or redundant predictors. Feature elimination can remove predictors that similarly impact the response variable. Eliminating such redundant parameters mitigates overfitting (Everitt, 2010). Feature elimination is often undertaken to arrive at the most potent group of predictors. For logistic regression, a stepwise approach automates sequential addition of a

candidate predictor to a model, then evaluates it for elimination by refitting the model (Harrell, 2013; Hosmer, 2013). In this commonly used method, the Akaike information criterion (AIC) is increasingly used as the criterion for predictor elimination. AIC is a measure of model fit that imposes a penalty for each added predictor; a lower AIC value is better. A preferred variant of stepwise regression is backwards stepwise regression (Field, 2012), where all predictors are initially incorporated in the model, then iteratively removed or retained based on the change in model AIC. This is a recursive feature selection technique. A stepwise approach is useful in narrowing a set of variables predetermined by research as relevant, but without prior screening to include only subject-relevant predictors at the outset, it is possible that even noise can be selected by this method as a significant predictor (Flack & Chang, 1987; Freedman, 1983).

Many statistical algorithms have built-in mechanisms to reveal predictors of greatest import to a model. Regression models automatically indicate the importance of predictors: for example the logistic regression predictor coefficient z-scores and odds ratios convey predictor contribution to a model. Decision tree models (L. Breiman, Freidman, J., Olshen, R., & Stone, C., 1984) have a summary measure, Goodness of Split, indicating contribution of each predictor to splits in a tree resulting in classification. In random forest (L. Breiman, 2001) a measure called Mean Decrease in Gini also measures the importance of a predictor, but as random forest is an ensemble of trees, this measure is across all random forest trees. In XGBoost (T. Chen, Guestrin, C., 2016) a measure called Gain is typically used to determine feature importance, and it too calculates the contribution of each predictor across all trees in the model. For more information see the section *The models: logisitic regression, decision tree, random forest and XGBoost.*

Alternatively, generic or unifying predictor of importance estimators are available that can be applied to several model types. One such offering is the varImp function [114)] for ranking the importance of predictors used in a model. If a modelling algorithm has a built-in calculation of predictor importance the varImp function calls the built-in importance function (but see (Overflow, 2016, October 5)). The varImp function can scale predictor contribution to a model between 0 and 100, with the predictor of lowest import is given a value of zero. How the varImp function works depends on the model type. The documentation, including online sources

(https://topepo.github.io/, 2019, March, 3; M. Kuhn, 2019, March, 3) stipulates scaling may be able to incorporate model-based inter-predictor correlations (M Kuhn, 2013).

A review of classification-related literature determined 64% of studies adopted inappropriate feature selection validation, which resulted in overfitted models that had pessimistic results on test data (Castaldi, Dahabreh, & Ioannidis, 2011). This emphasizes the importance of removing non-informative variables. Validating the feature set has been recommended as a solution, whereby the training data-set is used for feature selection, whereby feature selection is conducted using internal cross-validation (e.g. 10-fold cross-validation) involving resampling of hyper-parameters (M Kuhn, 2013). This allows feature selection to be applied to the held-out samples, which approximates testing on an independent data-set.

Modeling and the caret package

Model building typically involves a repetitive process. Relevant variables are used often in several trials in pursuit of the optimal inferential or predictive variable combination. The classification and regression training (caret) package (M. Kuhn, 2019, March, 3) expedites the iterative model development process. Resampling (bootstrapping, leave-one-out cross-validation or k-fold cross-validation) methods are used to arrive at estimates of model performance. An objective of caret resampling is to find optimal model tuning parameters automatically while also greatly facilitating comparisons of multiple model types. The resampling "injects" variation into the modeling process to aid in generalization on future samples. The caret package automates model tuning currently for 237 model types (e.g. CART, fuzzy rules, glmnet, knn, random forest, XGBoost, etc.).

Tuning parameters of model types relevant to the current work are the decision tree complexity parameter (determines the "price" of misclassification and tree depth) the random forest mtry (the number of randomly sampled variables used to spit data) and in XGBoost there are multiple tuning parameters (for details see *The models: logistic regression, decision tree, random forest, and XGBoost)*. Caret resampling can expedite the process of finding optimal hyper-parameter settings, and offers the latitude to use a measure other than just misclassification error to select the optimal tuning parameter settings. Specifically, Kappa or the ROC AUC can be adopted as criteria to find optimal tuning parameter settings. Borrowing from the online caret reference (https://topepo.github.io/caret/model-training-and-tuning.html), the model hyper parameter tuning process can be conveyed in a simple algorithm, here incorporating 10-fold resampling and the AUC performance measure:

```
"for a given tuning parameter do
    for each resampling iteration do
        Hold-out specific samples (10% of the samples)
        Optionally pre-process the data
        Fit the model on the remainder (90% of the samples)
        Make predictions on the hold-out samples using ROC AUC
    end
    Calculate average performance across hold-out predictions
end
Determine the optimal tuning parameter(s)
Fit the final model to all of the training data using the optimal tuning parameter(s)"
```

The caret train module schematic above runs 10 times, using a range of tuning parameter settings, and with each iteration the module trains a model on 90% of the data but tests a given (tuning) parameter on 10% of hold-out data to which the training module was not exposed. The testing or evaluation criterion (the ROC AUC here) results across predictions on hold-out models are averaged, the optimal tuning parameter (the one with the highest ROC AUC) is found, and the optimal parameter(s) is then used to fit a final model to all of the training data. The final model, benefiting from variation lent by the resampling process, will have improved generalization relative to a model without such resampling. The measures (ROC AUC, and associated sensitivity and specificity or metrics Kappa and accuracy) used to evaluate tuning parameters also provide an estimate of model future performance. It is noteworthy, that by repeating 10-fold cross-validation to some extent (e.g. 10 fold repeated 3 times or 10 fold repeated 5 times) variance across held-out data samples is further reduced (M Kuhn, 2013).

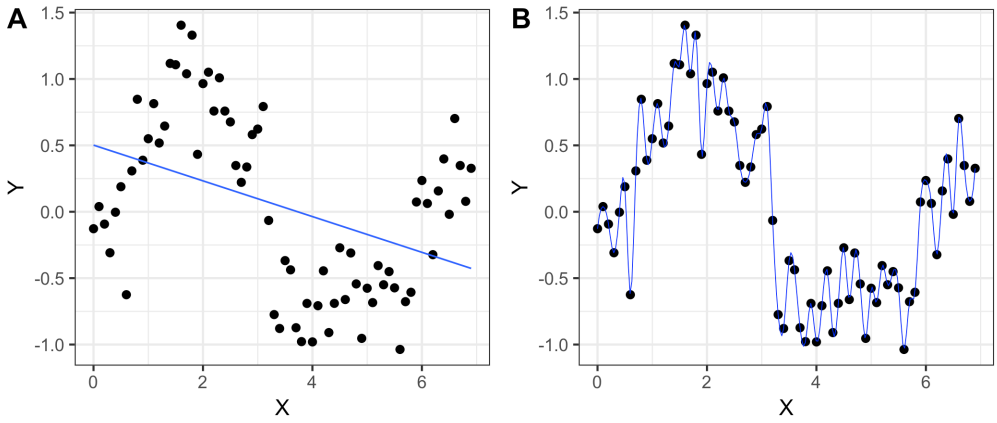
Caret can work in conjunction with built-in resampling methods. Using random forest out-of-bag (OOB) sampling as an example, the OOB sampling (a bootstrap with replacement) rate is calculated while the model is being built. By contrast, caret (internal) cross-validation, such as k-fold, makes predictions after the random forest model OOB rate has been computed, and the caret predictions are made on hold-out samples (without replacement).

It is important to note, that, the general linear model (glm) does not have tuning parameters, and caret will not tune logistic regression coefficients. Consequently, a logistic regression model's coefficients are unaltered by caret resampling. Moreover, it warrants adding that the caret package also includes functions for random stratified data splitting and pre-processing. Pre-processing, not discussed in the introduction, often takes the form of centering and scaling. Centering subtracts an average predictor value from that predictor's full set of values (e.g. x – mean(x)). Scaling simply refers to conversion of predictor values to z-scores. Centering and scaling can therefore remove the influence of original scale values that may detract from a model's ability to find relationships in the data. When

centering is applied to all predictors, all will have a common mean of zero, though predictors do not necessarily need to be centered on their mean. It is important to note, that if zero constitutes a meaningful value in predictors, centering is not generally recommended (Cohen et al., 2003).

Factors impacting predictive modelling

In a cursory outline of the bias-variance trade-off, bias refers to an over simplified analysis of a complex problem. A linear regression model assumes a linear Y and X_1, X_2, … X_n relationship but the relationship may not be linear (see Graph 2 A) in the population and important patterns in the data may be unaccounted for by the model. In general, parametric models, including linear and logistic regression as well as discriminant model types, have higher bias relative to more flexible models including non-linear tree-based algorithms; the latter tend to have higher variance. A model with high variance closely adheres to the pattern of the observations and is not constrained by any assumption of linear data relationships (as in Graphic 2 B).



GRAPHIC 2 | Bias (A) vs. variance (B)

Accordingly, a high bias but low variance model will fit a straight line to the data as in Graphic 2 A, while a high variance but low bias model will typically fit a curve passing through all points, as in Graphic 2 B. But the high variance model can be overly sensitive; so sensitive that it can even learn noise in the training data, which will likely result it having poor accuracy when applied to another data sample from the same population, given the new sample with not have identical noise to the training data on which the model was built. Ideally, the aim of a model is to have both low bias and low variance, and models pursue a bias-variance trade-off that both efficiently learns the patterns in a training data sample while also generalizing well. The bias-variance trade-off is often

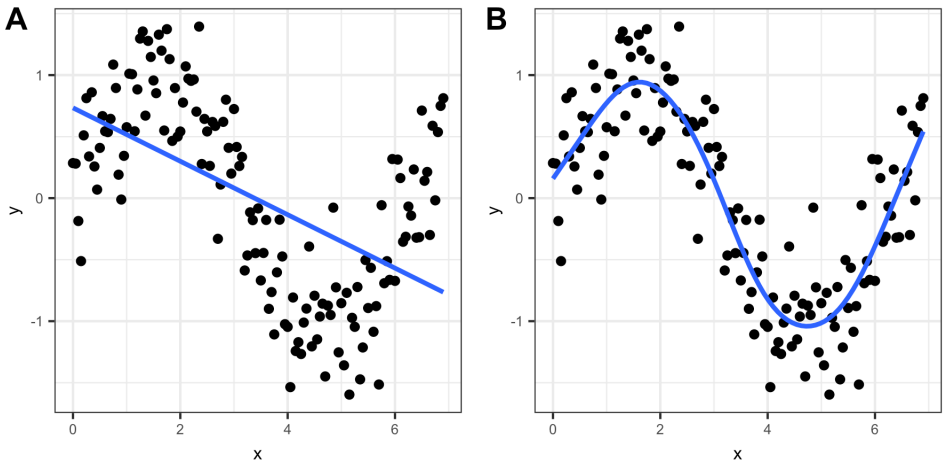
referred in terms of an underfitted-overfitted trade-off, and again the challenge is to find the balance between over and under fitting that captures key data patterns in the training data while also generalizing with high accuracy (Gareth, 2013).

In general, the objective of a predictive model is to find an estimate $\hat{f}$ that informs about $\hat{Y}$ utilizing systematic X information. Parametric models make the assumption the variables are linearly related. In logistic regression (a linear model), the specific assumption of linearity is that there is linear relationship between the predictors and the logit of outcome. A generic linear model taking the general form

$$f(X) \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

estimates coefficients (beta values) or parameters $\beta_0 + \beta_1 + \beta_2 + \cdots \beta_k$ from the data, where $_k$represents the $_{kth}$ predictor (Gareth, 2013).

Nonparametric tree-based models, as already noted, do not make assumptions about f (such as the relationships among variables are linear). Tree-based models are not reliant on linear combinations of predictors(M Kuhn, 2013). The general additive model (GAM) is a non-linear model type but one that includes both non-linear and generalized linear model link structures (see *The general additive model*). The lines in Graphic 3 plot the same simulated data as in Graphic 2, but the GAM spline curve in Graph 3 (B) reflects a much improved bias-variance trade-off: the model conforms to the general shape of the data while not adhering to every single data point nuance.



GRAPHIC 3 | A = linear method line; B= general additive model (GAM) spline curve

Parameters, estimated from the data, define a quantifiable contribution to the model. For example, logistic regression coefficients are parameters (as are support vector machine support vectors and neural network weights). Hyper-parameters, also referred to as tuning parameters, are not derived from the data by means of some analytical formula. Typically they can be entered manually and can be improved or tuned by changing their values manually or programmatically. For a given analysis, the optimal tuning parameter value is unknown (M Kuhn, 2013) and should be determined. The decision tree complexity-parameter (a hyper-parameter) the Random forest mtry hyper-parameter and the XGBoost (T. Chen, Guestrin, C., 2016) eta hyper-parameter are examples. The complexity parameter (a hyper-parameter) in a decision tree (L. Breiman, Freidman, J., Olshen, R., & Stone, C., 1984) determines how many times the data splits, in the process of predicting class labels, into branches, which controls the size of the tree. The complexity parameter value is available in a table that is created when a tree model is executed. It is common practice to select the complexity parameter value associated with the lowest cross-validation error corresponding to the lowest number of splits. The XGBoost eta hyper-parameter controls the rate at which new trees correct errors from a prior sequence of trees. The random forest (L. Breiman, 2001) mtry parameter, as previously mentioned, determines how many predictors are used to split the data at given point. The default is the square root (rounded) of the number of predictors, but this can be manually changed. Further details on hyper-parameters are outlined under *The models: Logistic regression, decision tree, random forest, and XGBoost*. Tuning parameter values that generalize optimally can be determined by validating a model on data separate from that on which the model was trained; specifically by applying the trained model on out-of-sample data. This can be efficiently accomplished with resampling (see *Resampling*).

The models: Logistic regression, general additive, decision tree, random forest and XGBoost

The model probabilities are summarized as discrete predictions in Chapter Four and Five confusion matrices. The related model performance metrics are also provided in Chapters 4 and 5 and summarized in Chapter Six. This section focuses on an overview of model algorithms. Tree and GAM model settings are also reviewed. It should again be noted that while this section reviews model-specific measures, the current study used the AUC as the main performance metric; model sensitivity, specificity, Kappa and accuracy measures were also

reported (see the above mentioned Chapters). Analyses for all models can be replicated using the R code in the Mods.R file (current version is ModsApril19.R)

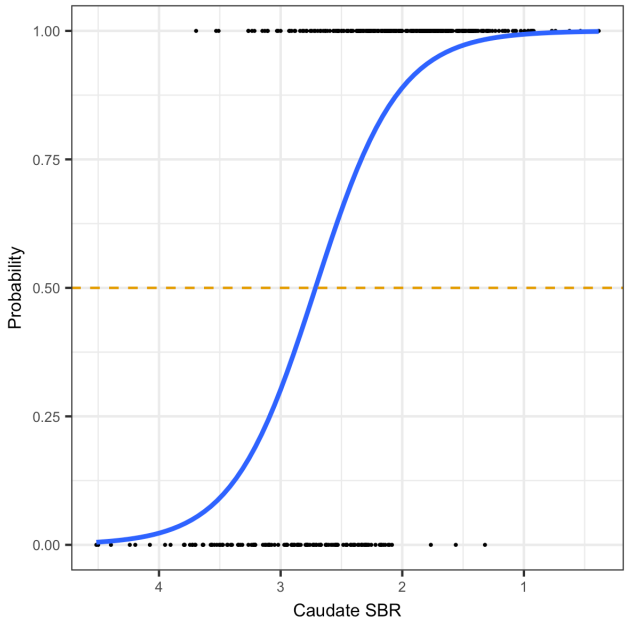
Logistic regression

In binary logistic regression, as with other binary classification models, the response variable takes a binary form 0 or 1. In this dummy coding, 0 = negative, which is synonymous with non-event, non-case; 1 = positive, which is synonymous with event occurrence or case (e.g. a case of pathology). As already outlined (see *Predictive classification model evaluation metrics*), with a binary response variable the predictive model output is continuous probabilities of presence/positive (P) or absence/negative of an event, such as disease. The probabilities indicate the extent of predicted case membership in either the presence of disease (P) group or absence (1-P) of disease group. Probability expresses the chance, quantitatively, that an event will happen; the number of times some event occurred divided by the total possible number of times the event could have occurred. Logistic regression is a departure from model types such as tree-based models in that it provides quantification of individual predictor and outcome relationships in the form of coefficients, the Wald statistic, and the odds ratios.

Logistic regression uses maximum likelihood (ML) to fit a model, where, very generally, ML selects coefficient estimates that make the observed values most probable. The ML process of deriving coefficients is a trial and error iterative procedure; it is not analytic as in linear regression which uses the least squares criterion. ML considers the likelihoods of subjects or cases having a particular Y-outcome (class label) based on predictor values and the estimated coefficients. ML, as its name suggests, provides the maximum likelihood coefficient estimates that make the sample observations as likely as is possible. The maximum likelihood of a given sample is often symbolized as L [141].

A logistic regression model predicts the probability of case group membership, the probability predicted from the known observation values of a predictor(s). This is depicted on y-axis in Graphic 4, which is logistic regression probability plot. The s-shaped pattern of the Graphic 3 conveys a non-linear and dichotomous relationship between predictor X and response variable coded 0 for controls and 1 for early PD. The probability of early PD relative to controls is on the y-axis (ranging from 0 to 1) and this is a function of (X) the caudate striatal binding ratio (SBR) on the x-axis. Quite simply, the predicted probabilities - the quantitative expression defining the chance that early PD will occur - are graphed against observed caudate SBR values. Caudate SBR refers to the average caudate DAT values (left + right caudate/2). Specifically,

the predicted probability $\widehat{pr}_i$ of being a case (of early PD) for the *i*th individual increases as (*X*) caudate SBR DAT value diminishes on the x-axis. Moving left to right along the x-axis the .50 or 50% cutoff occurs at a little < 3, after which the $\widehat{pr}_i$ (of being a case) is > .50. As caudate SBR values get progressively smaller (moving left to right) the $\widehat{pr}_i$ of being a case of early PD increases sharply as caudate SBR approaches a value of 2.



GRAPHIC 4 | Early PD vs. control caudate SBR probability plot
SBR = SPECT striatal binding ratio: left + right caudate/2

Logistic regression is often expressed as the natural logarithm. In the fraction on the left side of the expression below the numerator is the predicted probability $\hat{Y}$ =1 ($\widehat{pr}_i$), and the denominator amounts to the predicted probability $\hat{Y}$ = 0.

$$ln\left(\frac{\widehat{pr}_i}{1-\widehat{pr}_i}\right) = \beta_0 + \beta_1 X_{1i},$$

The *ln* on the left side of the expression uses the base of the natural logarithm (~2.72); it transforms the left side of the above equation to a logit, more commonly referred to as the log-odds or the log of the odds. This transformation expresses the distinctly non-linear relationship (a binary or categorical response variable is non linear) being modelled in a linear form (Field, 2012). The linear, right portion of the model is now expressed in probability units.

Again, the right side of the above equation indicates the logit of *Y* follows a linear form; β_0 is the intercept and β_1 is the coefficient for predictor X_1. The intercept β_0 is also referred to as the constant, which in the Graphic 3 example would be the predicted log-odds of caudate SBR for controls. Interpretatively, for every unit increment in predictor X_1 there will be a β_1 increment in the log-odds of outcome (a β_1 increment in the predicted probability of a case = 1, here early PD). With

multiple predictors, the logistic regression predicted probability of being a case for *i*th participant can be represented in the expression $\widehat{pr}_i = \frac{1}{1+e^{(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots\beta_k X_{ki})}}$, where e is also the natural logarithm and the bracketed part of the denominator represents a multiple linear equation. Each regression coefficient is, as in multiple linear regression, a partial regression coefficient: the effect of β_1 is interpreted controlling for other coefficients ($\beta_{1,}\beta_{2,}$ ... β_k) which are held constant (Cohen et al., 2003).

A predictor coefficient's interpretation is enhanced when it is converted to an odds ratio. In most software packages this is easily accomplished with an exponential function (exp). Predictor odds ratios can be expressed as $exp(\beta_{0,}\beta_1 X_{1i} + \beta_2 X_{2i} + \ldots \beta_k X_{ki} + \beta_0)$, where exp is the exponential function; $e^{(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\cdots\beta_k X_{ki})} = \frac{\widehat{pr}_i}{1-\widehat{pr}_i}$, where e is the natural logarithm (2.718282) and the odds ratio is the predicted probability $\widehat{pr}_i$ (ranging from 0 to 1) of having case status to the predicted probability of not having $(1 - \widehat{pr}_i)$ case status. Interpretation of the odds ratio (OR) is straightforward: an OR = 1 indicates the absence of a predictor relationship with *Y* and all participants have equal odds of case group membership; an OR < 1 indicates that increments in a predictor reduce the odds of the case group membership; OR > 1 indicates that as the predictor increases so do the odds of case group membership. Further, a unit increase in a predictor is associated with the multiplicative amount by which the odds of case group membership (Y=1) is changed. For example, if predictor X_1 had an OR of 2.58, this would indicate that a unit increment in X_1 multiplies the odds of being a case by 2.58.

Measures of the how well a model fits the data include the deviance statistic and the related likelihood-ratio, the Hosmer-Lemeshow goodness of fit test (Hosmer, 2013), the Akaike information criterion (AIC) (Akaike, 1974), and pseudo R^2 (McFadden, 1974). The logistic regression example (model 1) output in Table (A) serves as reference facilitating the explanation of the model goodness of fit (GOF) measures. Table (A) also assists in differentiating logistic regression given it provides quantification of individual predictor and outcome relationships (coefficients, the Wald statistic, and the odds ratios) not typically available in the tree-based models. Sections 4.2 and 5.2 provide substantive detail of the logistic regression specific outcome measures (e.g. coefficients, deviance, etc) for the PPMI early PD vs controls, and early PD vs. SWEDD analysed respectively. To avoid redundancy, the Table (A) data reflects an example logistic regression model built using an alternative data set: the Pima Indians data included with the mlbench package (Newman, 1998). The binary outcomes for this data set are diabetes positive (Y= 1) and negative, without diabetes (Y= 0).

The deviance statistic is directly related to log-likelihood. Finding the ML typically involves calculating the log-likelihood. The latter can be expressed as

$$\sum_{i=1}^{N} [Y_i ln\left(\widehat{pr}(\hat{Y}_i)\right) + (1 - Y_i) ln(1 - \widehat{pr}(\hat{Y}_i))],$$

where $\widehat{pr}$ is the predicted probability and the expression sums the probabilities associated with both the actual and predicted outcomes (Tabachnick, 2007) and is grossly analogous to the SS residual (residual sum of squares) to the extent that it provides the amount of unexplained variation remaining subsequent to the fitting of the model; a poorly fitted logistic model will have a larger log-likelihood (Cohen et al., 2003; Field, 2012). The deviance statistic equals negative 2 times the log-likelihood (*LL*): -2 × *LL* = – *2LL*. It is often used in place of the *LL*, largely because it has a chi-square distribution permitting simpler significance value calculation and comparison of logistic models, or comparison of the baseline model that represents only the constant without any predictors to a model including various predictors. New model deviance subtracted from baseline model deviance results in a difference called the likelihood ratio (Field, 2012). The likelihood ratio (which actually is subtraction not a ratio) has a chi-square distribution, where the degrees of freedom equal the number of parameters (i.e. the number of predictors plus 1 for the constant) in the new model minus the number of parameters in the baseline model: χ^2= (-*2LL*(new)) – (-*2LL*(baseline) (Field, 2012). If the baseline model is the null model (with only the constant and no predictors) the constant is the only parameter, meaning it has 1 degree of freedom.

TABLE A: model 1, logistic regression

Variable	$\widehat{\boldsymbol{\beta}}$ (***SE***)	*z*	*p*	*95% CI*	*Odds ratio*	*95% CI (Odds) ratio)*
Intercept	-0.77 (.62)	-0.49				
BMI	0.08 (.013)	5.47	< .0001***	(05, .11)	1.08	(1.05, 1.11)
Pedigree	0.83 (.29)	2.90	= .00412**	(.27, 1.40)	2.29	(1.30, 4.04)
Glucose	0.03 (003)	10.53	< .0001***	(.03, .04)	1.03	(1.03, 1.04)
Age	.03 (.007)	3.94	< .0001***	(.02, .05)	1.03	(1.02, 1.05)
	Null deviance:	993.48	on 767 df			
	Residual deviance:	747.23	on 763 df			
	AIC:	757.23				

Note: BMI = body mass index; Pedigree = diabetes insulin function

Comparing the fit of the logistic regression model 1 (see output Table A above) to the null model (with only the constant) a likelihood ratio test (model 1 deviance – null model deviance) indicated the model with predictors had a significantly better fit, $\chi^2(4) = 246.25$, $p < .0001$. The Hosmer-Lemeshow GOF test (Hosmer, 2013), which determines if predicted classification probabilities are similar to observed proportions, indicated the model was a

good fit to the data, $\chi^2(2) = 2.65, p = .27$. Note, that in the Hosmer-Lemeshow GOF test, a small chi-square value and large p-value suggest a good model fit to the data. The AIC (Akaike, 1974) statistic, another measure of fit when comparing two models, can be expressed as AIC = $-2LL + 2k$, where k is the number of predictors included and $-2LL$ is the deviance statistic. A smaller AIC is better reflecting that AIC (via $+ 2k$ in the expression) penalizes a model with more predictors. The smallest AIC occurs in models that combine the best fit and parsimonious use of predictors (Cohen et al., 2003). AIC is often used as the predictor selection (or feature elimination) criterion in stepwise regression and application of AIC in a stepwise procedure has been well documented (Harrell, 2013; Hosmer, 2013). For example, adding the additional predictor triceps (a skin fold thickness measure) from Pima Indians data to the predictors already included in model 1 to create a new model (model 2), we find that AIC becomes a little higher (758.24). In a stepwise regression using AIC (Venables, 2002) (MASS package) for predictor selection, triceps was removed from the regression model. However, a different predictor may have both improved model fit while also maintaining a low AIC, in which case the stepwise procedure would have retained the newly added variable in a five variable model. The final model GOF measure considered is the McFadden pseudo R^2 (McFadden, 1974) (see Cohen et al. 2003 for a discussion of linear regression variance vs. pseudo variance of logistic regression). Pseudo R^2 (pR^2) can be expressed as $pR^2 = 1 - \left[\frac{ln(LLM)}{ln(LL0)}\right]$, where $ln(LLM)$ is the fitted model log likelihood and $ln(LL0)$ is the null model (just the constant) log likelihood. McFadden pseudo R^2 ranges between 0 and 1; a value > .2 is considered satisfactory while a value < .2 is inadequate to explain the target outcome. The original four predictor model in Table A has a $pR^2 = .248$, while the five predictor model (not shown), where the triceps measure was added, has a $pR^2 = .249$. Where parsimonious models are preferred, the four-predictor model would likely be used given the additional predictor in five-predictor model only marginally increased the pseudo R^2 value. With respect to the coefficients, and interpreting just the predictor pedigree (a diabetes insulin function), a unit difference in pedigree, holding other predictors constant, is associated with .83 increase in the log-odds of diabetes; an effect that significantly differs from zero, z = 2.90, p = .004. A unit difference in pedigree multiplies the odds of diabetes by 2.29 (95% CI 1.30, 4.04). It

warrants note that the z-score, often referred to as the Wald (W) statistic is $W = \frac{\hat{\beta}}{\widehat{se}(\hat{\beta})}$, where the coefficient $\hat{\beta}$ is divided by the coefficient standard error.

Also, and with respect to coefficients, is optional application of coefficient standardization and scaling. Continuous or semi-continuous (integer rating scales with response scored over a range, e.g. 0 to 10) predictors can be converted to z-scores and used in a logistic model (Pampel, 2000). While this will not alter model goodness of fit measures (e.g. deviance), these predictors are all now measured commonly in standard deviations; i.e. a one standard deviation change in a predictor corresponds to the change in number of standard deviations in outcome. However, a binary (0, 1) predictor (such as female and male) converted to a z-score understandably loses its original interpretive clarity. In regards to scaling, semi-continuous integer rating scales coving a large range (e.g. 0 to 40) have smaller coefficient estimates relative to binary range variables like gender (0, 1) or a predictor with a 1-5 range. On a 0 to 40 scale 1-unit change covers just .025 of the scale, while a 1-unit change in 1-5 range predictor covers a considerably greater extent of the scale: 0.2 or $1/5^{th}$ of the scale. Consequently both coefficients and odds ratios of the predictor with the small-scale range will be more impressive. To facilitate more appropriate interpretation, semi-continuous rating scale predictors can be rescaled, whereby a given predictor is rescaled to have a binary (0 to 1) range by dividing each score by the range maximum. All semi-continuous scaled predictors will then have a 0 to 1 range with coefficients and odds ratios based on the same scale (Cohen et al., 2003). In the Table A model 1, there were no semi-continuous rating scales nor were standardized versions of the predictors provided.

As already noted and demonstrated, unlike the tree-based models (decision tree, random forest, and XGBoost), logistic regression parameters (coefficients) have readily available Wald (z-scores) and p-values and odds ratios that quantify the relation between each predictor and outcome. Such per predictor quantification can be of great value but comes at the price of a few assumptions that must be met for results to be valid. These assumptions, again in aid of brevity, are not tested for the example model using the Pima Indians data. However, the general assumption and diagnostic guidelines for logistic regression warrant brief outline.

The observations must be independent, such that there must be a lack of autocorrelation and hence independence of errors (residuals are not correlated); the same participants can not be measured across two time points. This can be assessed with the Durbin-Watson test (Durbin & Watson, 1951) adapted for logistic regression (Fox, 2011; Group, 2013, November 14). There is also a requirement that allows for only a small amount or no multicollinearity. Highly correlated (e.g. $r >= .75$, or perhaps less conservatively, $r >= .85$) predictors should normally be removed at the outset (M Kuhn, 2013). The extent of multicollinearity can also be assessed with a logistic regression variation of the variable inflation factor (VIF)(Fox, 2011). A VIF of > 5 associated with any predictor indicates problematic amount of multicollinearity. Logistic regression is also sensitive to outliers and influential cases.

Recommended logistic regression case-wise diagnostics to examine the effect of outliers include assessment of discrepancy, leverage, and influence measures. Specifically, case discrepancy is determined using the (externally) studentized residual, which measures the extent of discrepancy between observed and predicted $\hat{Y}$ values. A commonly adopted threshold cut-off value is ± 2 (other cut-off points have also been recommended; see Cohen et al., 2003 (Cohen et al., 2003). For Cook's distance, which is a measure of given case's global influence on the model, the widely used threshold value of 1 is typically adopted. With respect to leverage, a measure of how far case *i* is from the mean of the predictor(s), a recommended cut-off value can be calculated with "$3M_h = 3(k + 1)/n$" (Belsely et al., 1980), where M_h is mean leverage for the predictor, *k* is the number of predictors and *n* is the number of cases.

Finally, with logistic regression there is an assumption of linearity of the logit, which means there must be a linear relationship between the logit of outcome and continuous predictors. This can be assessed graphically (Zhongheng Zhang, 2016) but has traditionally been tested using the Box-Tidwell transformation (Box & Tidwell, 1962). The Box-Tidwell test amounts to incorporating in the model predictor interaction terms that are the cross-product of a given predictor and its natural logarithm (e.g. X_1* log(X_1)). Any predictors that show significant ($p < .05$) interaction with their natural log are regarded by the Box-Tidwell test as violating linearity of the logit. In the event a predictor violates this assumption, it is often converted from a continuous to percentile-based categorical variable such as a quintiles or quantiles. This is a practice common in epidemiological research that also allows

convenient framing of the relation between binary outcome and low-medium-high variable levels (Thiebaut et al., 2007; Tu, Austin, & Chan, 2001; Vickers et al., 2007; Vickers et al., 2009). However, the choice of categorizations is arbitrary, may miss portions and hence characteristics in the data, and may fail to produce sufficiently sensitive models (Bennette & Vickers, 2012; Greenland, 1995; Royston, 2000).

The general additive model (GAM)

The general additive model (GAM) is, like tree-based models, a non-parametric alterative to logistic regression that does not make any assumption of linearity between a predictor and the response variable(Jones & Wrigley, 1995). Uniquely, while a GAM is nonparametric it combines both linear and non-linear generalized linear model link structures (lm, binomial, poisson etc.). As such, GAM models could be characterized as a bridge between parametric and non-parametric models (Wegman & Wright, 1983). This is achieved by capturing the impact of a predictor variable with non-linear properties via a nonparametric smoothing function. The smoothing function can take the form of several basis functions – known transformation functions such splines (e.g. cubic splines, smoother splines, thin plate splines, etc.). The GAM fit to data is generally assessed with deviance, as is the cased in logistic regression. As such comparison between GAM and logistic regression using deviance is simplified (the model with lower deviance is a more efficient fit to the data).

In regression, a predictor is multiplied by its regression coefficient. This also occurs in a GAM, but any predictor can alternatively be multiplied by a smoothing function, such as a spline; the smoother spline replaces the predictor coefficient. A GAM could be represented as $g(E(Y)) = \beta_0 + s_1\left(X_{i_1}\right) + \cdots s_p\left(X_{ip}\right) + \varepsilon_i$. The variable *g* represents a link function (e.g. family = binomial (link= 'logit), connecting a value that is expected *E(Y)* to the predictor (s); *Y* is the response variable; *p* refers to a predictor (Hastie & Tibshirani, 1995). The β_0 is the intercept, ε_i is the error, but $s_1\left(X_{i_1}\right) + \cdots s_p\left(X_{ip}\right)$ are nonparametric smoothing functions that lend GAMs their distinctive capacity. An assumption of a GAM is that a smoother function can be estimated in a scatter-plot smoother (Jones & Wrigley, 1995).

Consider a logistic regression classification GAM to predict (the probability, *pr*) of pathology, *pr* (*y*=1|*X*) vs. *pr* (*y*=0|*X*), which includes predictors X_{i_1} and X_{i_2}. If X_{i_1} violated the assumption of linearity of the logit it

could be wrapped in a smoothing function. And, if X_{i_2} did not violate this assumption it could be run in the same model but without the need to have a nonparametric smoothing function applied to it. Such a model could be expressed as $\log\left(\frac{pr(\mathrm{X})}{1-pr(X)}\right) = \beta_0 + s_1\left(X_{i_1}\right) + \beta_2(X_{i_2}) + \varepsilon_i$, where X_{i_1} is wrapped in a nonparametric smoothing function, such as a form of spline, but X_{i_2} is not and remains parametric (Gareth, 2013).

A smoothing function typically retains more information than percentile-based variables such as quantiles. While this does likely increase sensitivity of the model it can also lead to a tendency of overfitting (Weinberg, 1995; Simon N. Wood, 2008). The more closely a smoother function follows the pattern of data the more degrees of freedom are used up (more degrees of freedom are added to the model). Figuratively, this can be conveyed by a smoother in the form of a wiggly line passing through points in a scatterplot, as in Graphic 3 (B). In a GAM, the segments forming the curve in such a scatterplot are the predictors separately fitted to polynomial functions, and knots delimit these predictor/segments. In the case of a cubic spline for example, *k* knots uses 4 + *k* degrees of freedom. In a more recent GAM package (mixed GAM computational vehicle [mgcv]), briefly outlined below, (S. N. Wood, 2019, March 21) the degrees of freedom is governed by the type of penalization used (e.g. AIC, REML, etc.). In general, the more points in a scatterplot the wiggly line hits, the higher the degrees of freedom and the greater flexibility the model has, but such flexibility incurs a propensity to overfit. The lines in Graphic 3 plot the same simulated data, but the GAM spline curve in Graph 3 B conforms much more closely to the shape of the data than the linear method in Graphic 3 A, and the gam spline curve in B hits many more data points. A curve passing through virtually all points (as in Graphic 2 B) is certain to capture more variation in the training data but too close adherence to training data variation may result in a model that generalizes poorly. In short, overfitting occurs when a model learns trends specific to the training data set and will not generalize well to new, unseen samples. While this can occur in almost any model, here with respect to the smoothing function, to counter this propensity the smoothness level of a smoothing function can be adjusted by a smoothing parameter $s_\lambda\,(X)$, where s_λ represents a smoothing function and λ (lambda) denotes the smoothing parameter that exerts control over the extent of smoothing and hence the degrees of freedom. Conveniently, the mgcv package provides automatic selection of the smoothing parameter.

There is, as might be expected, a trade-off in this adjustment between "wiggliness"and smoothness of a curve: when $\lambda = 0$ "wiggliness" of the curve is maximized as is the fit and as λ is increased the curve becomes smoother (and degrees of freedom decrease as the smoothing parameter increases). In short, a smoothing function's coefficients are penalized to ideally limit the extent of curve wiggle in order to satisfy the dual GAM objective of minimizing the degrees of freedom while also maximizing model fit to the data.

In the GAM mgcv package(S. N. Wood, 2019, March 21), smoothing parameters and coefficients can be estimated from the data by generalized cross-validation (GCV), maximum likelihood (ML), or restricted maximum likelihood (REML). For details see Wood (S. N. Wood, 2019, March 21). In the current work, smoothing parameter selection (and coefficient estimation) used REML, which involves treating wiggly spline parts as random effects terms within the likelihood framework of random effects (Simon N. Wood, 2008). The REML fitting method inhibits overfitting, at least compared to the CGV method (S. N. Wood, 2011). Smaller values of REML indicate better fitting models.

A GAM typically has distinct inner and outer loop iterative operations. While the outer loop functions to maximize the overall model fit selecting parameters making the data most probable (as in maximum likelihood), the inner loop is involved in smoothing of individual predictors using a smoothing function $s_\lambda\,(X)$ that optimizes the fit for a given smoothed predictor relative to the partial residuals (i.e. the fit to the data of X_{i_1} controlling for other predictors).

Procedurally, recommended steps to diagnose a GAM include initially fitting the model, extracting the deviance residuals, and then checking that a smoothed term's k-value does not closely approach the term's estimated degrees of freedom (S. N. Wood, 2004). In the mgcv package (S. N. Wood, 2019, March 21) the k indicates the number of basis functions for a smoothing term, such as a thin plate spline (the default in the mgvc package). For a thin plate smoother, the k-value defaults to -1, which ultimately amounts to 10 basis functions and hence a maximum of k-10 = 9. Also, k is automatically governed by choice of smoother penalization. The estimated degrees of freedom should be inspected. Where k represents the maximum upper limit on a smoother term's estimated degrees of freedom, k should not approach the smoothed term's estimated degrees too closely (S. N. Wood, 2019a). However, a value of k that is too low may not capture significant statistical outcome. Again,

while the k-value is automatically provided, it should be assessed for optimal results. One approach is to use the "gam.check()" function, which will tabulate the gam model smoothing parameter k and estimated degrees of freedom (edf).

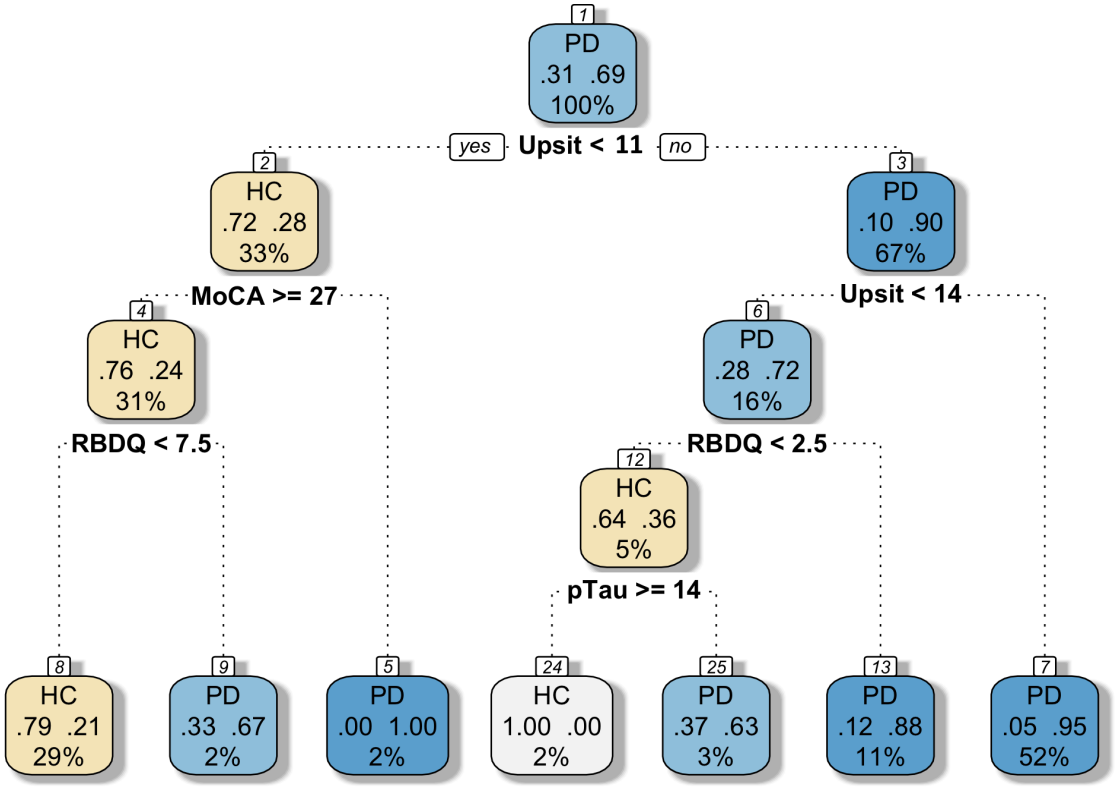
The logistic regression GAM model produces a term referred to as deviance explained, which is a pseudo R^2 goodness of fit measure. This is the same as McFadden pseudo R^2 (McFadden, 1974). In addition, as noted previously, logistic regression and GAM deviance can be directly compared (using anova with a Chi-squared test). Further, AIC values of each model type can also be compared. Of course, performance metrics ROC AUC, accuracy, sensitivity and specificity can also be assessed.

The increased sensitivity of GAM should be weighed against heightened interpretation difficultly. A smoothed variable in a GAM, while provided with a p-value, does not retain the regression parameters- the coefficients. This complicates interpretation. Also, the implementation of a GAM is more complex than logistic regression. Moreover, where the outcome of a GAM constitutes only a small improvement relative to a simpler model logistic regression model, the latter is preferable (Hastie & Tibshirani, 1995). Finally, assessment of variables as percentile-based and as smoothed functions can be useful (Bennette & Vickers, 2012). The former provides a preliminary perspective and simpler interpretation of the data while the latter provides a more comprehensive examination of the data. It warrants mention that the caret package (Kuhn,et al., 2019), while providing a general wrapper for the mgcv GAM, did not, a the time of this writing, allow specification of GAM formula (all predictors are automatically smoothed).

Decision tree

Decision tree regression and classification types exist. Accordingly, decision trees are also called classification and regression trees (CART). The open source implementation of CART in *R* is provided in the "*rpart*" package. While the current work included a decision tree classification model rather than a regression tree model, the XGBoost classifier, also included in the modelling, does employ an ensemble of regression trees rather than an ensemble of decision trees. As such, a brief explanation of a regression tree is warranted. Graphic 5 can be used to visualize a tree: the rectangular shapes are nodes (or final leaves at the bottom of the graphic), which are split, during the classification process into branches leading to other nodes further down the tree.

The *rpart*(L. Breiman, Freidman, J., Olshen, R., & Stone, C., 1984) function will create a decision classification tree if the outcome is a factor but will create regression tree if the outcome is numeric. Both decision trees and regression trees have a top-down recursive, binary splitting approach; top-down as data splitting from the outset uses the predictor with the strongest association with the dependent variable to decide on the best data split, but the split at a given step in the partitioning of data does not consider how a current split might improve a future split. Initially, all observations are in the same region, the apex or top of the tree. The apex is called the root, and is actually the start point of an inverted tree. Using Graphic 5 as reference, tree parts consist of rectangles called nodes, branches and nodes at the bottom call leaf nodes. The single *best* predictor, and in the case of a continuous predictor the latter's appropriate cut point (e.g. some score on a 0 to 10 scale), splits all the data to partition it at the outset into two separate branches. Each subsequent partitioning follows in similar fashion: the data split is based on the variable that makes the largest contribution to reducing heterogeneity of a node. New child nodes that result from a split are in turn split, repeating the procedure and this process continues successively with the final outcome shown in the leaf nodes at the bottom of the tree. The leaf nodes are used to arrive at predictions. Note, as depicted in Graphic 5, the partition or split decision uses a single predictor at each decision node. Following Graphic 5, the top of each node displays the classification, the mid-node values reflect the probability of the class at that node, and the lower node value is the percentage of observations at the node. In the first split, those with (reverse-scaled) UPSIT greater than 11 were classified as early PD, else they were classified as controls, etc.



GRAPIC 5 | Decision tree plot: Classification tree (two-class or binomial classification). PD= early PD, HC = controls; MoCA= Montreal Cogntive Assessment; pTau= phosphorylated tau (tau protein maintains microtubules in axons; a variant of tau is phosphorylated and contributes to neurofibrillary tangles); RBDQ= rapid eye movement behaviour disorder questionnaire; Upsit = University of Pennsylvania Smell test, reverse scaled. The model was constructed in *rpart* using current study early PD vs. control (HC) classification data. Optimal hyper-parameter values selected were a complexity parameter (*cp*) of 0.02197, a minsplit = 9, minbucket = 3. The root node (rectangle) at the top of the tree represents the entire data set sample or population. The top of each node displays the classification, the mid-node shows the probability of the class at that node, and the lower node value is the percentage of observations at the node.

The summary function in R provides a much more detailed output of the decision tree spits, specifying the splitting rules at each and every data partition of the tree.

As might be expected, what defines the *best* split is quantitative in a regression tree but qualitative in the classification tree. The best split(s) of the data in a regression tree is that involving the predictor, and that predictor's cut point value, that divides the data into the most homogenous classes. The predictor with the lowest residual sum of squared error (RSS/SSE) is chosen to split the data, which leads to subset (branches of the tree), lower regions of the tree of potentially more homogenous groups. The process is continued recursively. Finding the RSS can be expressed as "$RSS = \sum_{i \epsilon Sa}(y_i - \bar{y}_1)^2 + \sum_{i \epsilon Sb} y_i + \bar{y}_2)^2$", where $\bar{y}_1$ and $\bar{y}_2$ reflect average outcomes for class *Sa* and class *Sb* groups, with outcomes derived from the training data (M Kuhn, 2013).

A classification decision tree's *best* split, at a given node, also aims for homogeneous differentiation of classes but the default *rpart* method, in a classification (decision) tree seeks a data split that results in a subpartition's optimal

class purity- ideally a node and ultimately leaf node that is a single, homogenous class. Such a node has an impurity equal to zero without any misclassification errors. By default, *rpart* classification tree node purity is based on the Gini index.

The Gini index has a range of 0 to .5 in a binary model. Given a binary (two-class) model, the objective of the Gini index is to optimize purity or homogeneity of a node, with the value zero indicative of an entirely pure, single class node. The Gini index can be expressed can be expressed as $pr_1\,(1 - pr_1) + pr_2\,(1 - pr_2)$, where pr_1 is probability of class 1 and pr_2 is the probability of class 2. If a node Gini index is $pr_1 = pr_2$ the node has maximal impurity. But if the Gini index is expressed as $2pr_1\,pr_2$, it is apparent that a node's Gini index moves towards heightened purity if either or both of the class probabilities approaches zero (M Kuhn, 2013). Moreover, the rpart model will select the data-splitting variable that has the lowest Gini index. Further, there is an optional priors augment Gini index rule that assists in selection of low risk splits (Therneau, 2018). The default priors are weights proportional to the number of data instances in a given class. These priors can be altered in *rpart* (but were not altered in the current work).

A decision tree initially produces a fully-grown tree. Such a tree can include too many splits and has tendency to overfit the training data. However, pruning back this tree to obtain a smaller tree that produces equivalent or better results without the same extent of overfitting, is typically the next step. Tree pruning is controlled through specific hyper-parameters monitored by *rpart* that can be tuned or adjusted, such as the minimum number of observations required to attempt a split at a given node (minsplit = 20, by default), the minimum number of observations required in a terminal leaf node (minbucket = $1/3^{rd}$ of minisplit by default), and the complexity parameter (*cp*). Of particular usefulness, the *cp* is a cost function; it determines the "price" of misclassification and provides an indication of error. The initial fully-grown tree provides a table (cptable) that includes cross-validated errors (xerror) and associated *cp* values. The xerrors summarize the Gini index across the nodes, and the xerror measure undergoes cross-validation (k-fold, actually 10-fold).

In *rpart* the *cp* parameter is the essential tree pruning control and it has a default of *cp* = .01. The *cp* is literally the amount of improvement in a node's relative error gained by splitting the node. Importantly, the *cp* parameter penalizes the tree for each additional split. Typically, in practice, after an initial model is run, the *cp* coefficient with the lowest xerror should be selected and entered into a new model. This will likely prune off a given split that does not improve the model and so reduce overfitting. A generally agreed optimal *cp*-value can be

arrived at using the earliest split with the lowest xerror. The appropriate *cp* value can be extracted programmatically, or simply selected from a table of model associated *cp* values, or selected automatically by the *caret* package (M Kuhn, 2013). The settings for the final model depicted in Graphic 5 were as follows: model<-rpart(Class~ Age + Upsit + RBDQ + MoCA + AB + pTau, data = datnames3, control = rpart.control(minsplit= 9,minbucket = 3, cp=0.02197802,maxdepth = 5), method = "class"). The optimal features in the latter model were determined by 10-fold resampling (10 folds x 5 repeats) in the caret package. As with all models in this work, the model and features producing highest AUC constituted the final model.

The importance of a predictor to decision tree class discrimination is measured by *rpart's* built-in goodness of split metric. The goodness of split is the sum of a given predictor's usage as the primary basis of a split in all nodes, which is also the sum of the predictor's contribution to decrease in impurity. A predictor's usage includes its use as a surrogate (i.e. when it is substituted for another variable's missing data-point). This sum is transformed in to a percentage score (100% being the maximum). Details of the *rpart* algorithm have been well documented and provide added insight to this model type (L. Breiman, Freidman, J., Olshen, R., & Stone, C., 1984; Therneau, 2018).

Trees are overly sensitive to minor alterations in the training data. This can be addressed by ensemble methods that allow splitting decisions to be averaged over multiple tree constructions. The prediction accuracy of a decision tree model can be enhanced by ensemble methods that utilize trees as building blocks (e.g. bagging, random forest, XGBoost). These tree enhancement methods are outlined next.

Random forest

A decision tree model typically has higher variance (and hence higher error on unseen data) but lower bias relative to a logistic regression model. Bias in a logistic regression model, as already discussed, stems from its assumptions, which are that there is a linear relationship between continuous predictors and the logit outcome, independence of errors holds, and predictor correlation is not high. A decision tree makes no such assumptions and has consequently low bias. However, again, decision trees are prone to high variation (and a tendency to overit), which is effectively countered in ensemble algorithms. There are ensemble algorithms that combine

predictions from multiple models to reduce variance and improve predictive accuracy. Ensemble methods include bootstrap aggregation (bagging), random forest and XGBoost.

The bagging algorithm (L. Breiman, 1996) (bagging is an abbreviation for bootstrap aggregation) takes multiple *B* bootstrap samples from the training data with replacement, trains the algorithm on each *bth* bootstrapped sample, and then arrives at an aggregated average prediction (Gareth, 2013; M Kuhn, 2013). In classification, this averaging typically involves obtaining the "majority vote" across all bootstrapped trees in an ensemble for a given test observation (Gareth, 2013). The averaged model predictions (total number of votes divided by the ensemble's number of models) are aggregated with each model having the same weight in determining into which group or class the sample is classified. This averaging process reduces variance across individual model predictions and typically improves accuracy (M Kuhn, 2013). In addition, in the process of bootstrap sampling, and with each model built in the collective or ensemble, some samples (approximately $1/3^{rd}$) are left out and are referred to as out-of-bag samples. Models for each bagged tree are fit to approximately two-thirds of observations. Because the out-of-bag samples are not used to fit a given model, they are used to test an individual model's predictive accuracy (or test error), and an average out-of-bag (OOB) based metric (the out-of-bag estimate) provides a measure of performance accuracy for the entire aggregated ensemble that is typically similar to that derived from a cross-validation method. The OOB error then, is a bootstrap estimate of the aggregated model prediction error on a given sample x_i only utilizing trees without x_i in their bootstrapped sample (Gareth, 2013). It is a form of internal cross-validation to mitigate overfitting.

While bagging may improve prediction accuracy, interpretation of bagged trees is, understandably, complicated relative to a single decision tree. Variables of importance in a solitary decision tree are evident in numeric and graphic output but such is not the case with bagging a large number of trees (e.g. 50 iterations or more). But given all predictors are assessed for consideration at all splits of every single tree, it follows that the structure of bagged trees can be similar. As a consequence high tree correlation can result that inhibits variance reduction of predictions (Gareth, 2013; M Kuhn, 2013). By contrast, in random forest there is a random selection of predictors that avoids the bias that a decision tree or simple bagging could introduce.

The original the random forest algorithm was unveiled at a conference in 1995 (Ho, 1995), further developed subsequently (Amit & Geman, 1997; Ho, 1998), with the appearance of the most widely known version in 2001 (L. Breiman, 2001). Random forest adds an element of randomness to bagging. Each spit in a random forest (decision) tree results from a single predictor randomly selected from a random subset sample of predictors. This introduces randomness that mitigates tree correlation and is in contrast to bagging where all predictors are considered for all splits of every tree.

To briefly elaborate, in addition to the ensemble bagging procedure, the random forest tree building procedure selects, from the set of all predictors, a subset of randomly sampled predictors *p* for each tree split. The number of *p* predictors in each newly, randomly selected predictor subset *m* is determined by the forest tuning parameter m_{try}. But of the *m* predictors, the actual tree split is based on only a single, solitary predictor. The default m_{try} setting for a classification analysis approximates the square root of the number of predictors (i.e. $m_{try} \approx \sqrt{p}$) (L. Breiman, 2001).

A random forest does not avoid overfitting by pruning; decision trees, as already outlined, do adopt pruning via selection of sub-trees with the lowest cross-validated error. Random forest has been demonstrated as protected from overfitting (L. Breiman, 2001), and in application this is achieved by optimizing the tuning parameters, notably the m_{try} parameter. Typically, random forest has reduced OOB prediction and test error relative to bagging (L. Breiman, 2001; M Kuhn, 2013). Optional parameter alterations have been recommended to optimize performance: assessing five different m_{try} evenly spaced values ranging from 2 to the number of predictors *p*; experimenting with the number of trees (ntree) grown by starting at 1000 and increasing the ntree number until performance no longer improves (M Kuhn, 2013). The default random forest hyper-parameter settings include, ntree = 500, nodesize = 1, and, as already specified, $m_{try} \approx \sqrt{p}$. It should also be mentioned that random forest can overcome missing values in the training data set (L. Breiman, 2001).

A random forest model has heighted complexity of interpretation, though this is largely overcome by a built-in function that estimates variable importance to the model. Two measures of feature importance in random forest are Mean Decreased Accuracy and Mean Decreased Gini values (L. Breiman, 2001; Liaw, 2018, March

25). Mean Decreased Accuracy of a variable refers to the proportion of data instances, observations, incorrectly classified when that variable is not included in the model. The Mean Decreased Gini impurity refers to the mean purity gained from the splitting of a particular variable. Larger values associated with a predictor's Mean Decreased Accuracy measure and a predictors Mean Decreased Gini indicate greater predictor import to model classification. The two measures of variable importance are highly correlated and it has been demonstrated that both measures have a bias favouring categorical predictors with more categories (C. Strobl et al., 2007) . It warrants mention, that the estimated importance of a variable in random forest may be unreliable with multicollinearity (M Kuhn, 2013) (see also https://stats.stackexchange.com/questions/59124/random-forest-assumptions). Overall, random forest is one of the highest performing analysis methods (Collins, Griffioen, Newell, & Mellor, 2018; Pal, 2005; Rodriguez-Galiano, Ghimire, Rogan, Chica-Olmo, & Rigol-Sanchez, 2012), and an undeniable appeal of this algorithm is that it often performs well "out of the box" requiring little in the way of model tuning expertise.

XGBoost

The extreme gradient boosting (XGBoost) package (T. Chen, Guestrin, C., 2016) is a variant of gradient tree boosting (Friedman, 2001). Gradient tree boosting includes a few key components: gradient descent; a loss function measuring how well model coefficients fit training data, or more generally, the cost to a model of an incorrect classification; a weak learner, which refers to a poorly performing tree or sub-model in a modelling analysis (one with accuracy just above 50% or chance); and a loss function that is minimized by adding weak learners.

First, unlike random forest that uses decision trees, XGBoost uses regression trees. Considering the broadly familiar regression setting, a gradient descent boosting model's overall objective is to go from a high loss/cost (e.g. a least squared high SSE in linear regression or high Log Loss in logistic regression or XGBoost) to the lowest cost possible. The reduction in cost is done gradually, and the residuals (actual observations minus predicted), calculated at each point in the descent from high to lower cost, constitute the gradient of descent. A tree (weak learner) is fit on residuals (the gradient) from an initial regression tree model. Then, in an additive procedure, another tree is added to the model, and the model is fit on the current set of residuals (not on the

original Y) that further reduces the residual loss. As such, the model includes tree 1 + tree 2. The process continues based on a user-defined amount of iterations (Friedman, 2001; Gareth, 2013; M Kuhn, 2013). XGBoost is distinguished from other tree boosters by two properties in particular: it uses second-order (partial) derivatives (gradients) of a given loss function to arrive at more precise information regarding gradient direction, which further minimizes the loss function. It also offers some of the most advanced regularization for controlling overfitting.

Gradient boosting combines and averages multiple weak individual tree sub-models in an ensemble to produce an improved learner. Unlike Random Forest, models are not made from entirely random subsets of features and data. Rather, and as may be evident from the last paragraph, model building is conducted sequentially (adding trees to a model), and incorporates gradient descent (including second-order gradients in XGBoost) to decrease the cost/loss function as quickly as possible, iteratively. Initially, model prediction errors for every observation are determined. A model is then fit on errors (e.g. residuals as noted above) and this process is reiterated adding such error finding models to an ensemble to reduce misclassification rate. It is also worth distinguishing that while random forest mitigates overfitting largely by training models on randomly sampled data and predictors, XGBoost, as noted at the end of the preceding paragraph, includes elaborate regularization (penalization): e.g. Ridge Regression (L2) and least absolute shrinkage and selection operator (Lasso) (L1). Briefly, in a trade-off between fitting and overfitting a model, ridge regression shrinks parameters, coefficients, by adjusting a scalar λ, which can have a broad range (e.g. .01 to 10^{10}). The λ can add to a loss function, as expressed here, $loss = \sum(\hat{y}_i - y_i)^2 + \lambda \sum \beta^2{}_n$, where λ multiplies the sum of the coefficients. This effectively penalizes model coefficients that have high values. Lasso penalizes high coefficient values, as in ridge regression, but also sets irrelevant coefficients to zero(Gareth, 2013).

As already noted, XGBoost uses regression trees. As such, MSE or RMSE are appropriate cost functions; log loss or AUC are appropriate error metrics for classification. In binary classification, a logistic transform is used (specified in model settings as “binary: logistic”). Model training incorporates both loss (e.g. Log Loss, simple classification error or AUC) and regularization functions, and learns in an additive fashion (with a weak

learner tree fit on residuals). A model can be expressed as "$\mathcal{L}^t = \sum_{i=1}^{n} l\,(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \sum_k \Omega(f_t)$", where $\hat{y}_i^{(t)}$ is *i*th instance, *t*th iteration prediction (T. Chen, Guestrin, C., 2016). The function f_t is a "greedy" function (continues until it can improve the model no further). The l in $l\,(y_i, \hat{y}_i)$ refers to a convex loss function (gradient descent) calculating the difference between y_i and $\hat{y}$. Within the $\sum_k \Omega(f_t)$ term, the Ω is a regularization or penalty (e.g. L1 or L2) term. Finally $\sum_k \Omega(f_t) = \gamma T + \frac{1}{2}\lambda||\omega||^2$, ω is a vector of leaf scores, T is the number of leaves in a tree, and w_i weights of each leaf have an associated penalty λ term. The gamma in γT is some number that can range from zero to infinitely and will depend on the loss function used (e.g. log loss, AUC, etc.) and the number of leaves (number of leaves in turn depends on the maximum depth of a tree). The second-order gradient, used to better minimize loss has been expressed as "$\mathcal{L}^t \cong \sum_{i=1}^{n}[l\,(y_i, \hat{y}_i^{(t-1)}) + g_i\, f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t)$", where the gradient statistic for the first order loss function is "$g_i = \partial_{\hat{y}^{(t-1)}}\, l(y_i, \hat{y}_i^{(t-1)})$" and for the second order loss function is "$h_i = \partial^2_{\hat{y}^{(t-1)}}\, l(y_i, \hat{y}_i^{(t-1)})$" (T. Chen, Guestrin, C., 2016).

Cover, frequency, and gain are available to measure feature importance. As with all model types, certain features improve model accuracy while others have an antithetical effect and simply add to the error. Cover refers to the number of observations that arrive at a leaf node based on a given feature. A feature's frequency is a count of the number of times a given feature occurs in all trees built. Gain is a feature's relative contribution to the model that takes in to consideration every features'contribution building all model trees. Gain is generally regarded as the most appropriate measure of feature importance (Exchange, 2019b). A higher feature associated Gain value indicates greater feature importance.

There are multiple hyper-parameter settings in XGBoost. Detailed information on all hyper-parameters (including their default settings) is readily available (T. Chen & Guestrin, 2020)(https://xgboost.readthedocs.io/en/latest/parameter.html#parameters-for-tree-booster). In the current work, seven tree booster hyper-parameters were tuned beyond their defaults to find optimal values using k-fold (10-fold, 5 repeats) internal cross-validation (internal as it was within the training data) resampling in the caret package (M. Kuhn, 2019, March, 3). Theses hyper-parameters will be briefly reviewed and are eta, colsample_bytree, gamma,

nrounds, max-depth, min_child_weight, and subsample. For the final model hyper-parameter settings used in the current work, see the XGBoost results in Chapters 4 and 5. Note the L1 and L2 regularization terms briefly outlined earlier in this section were left at their defaults: lambda (L2) default =1. The L1 term in XGBoost (alpha) has a default =0.

Hyper-parameters, among those just listed, that are particularly useful to control overfitting are those that add randomness during training and those that penalize model complexity. Colsample_bytree and subsample add randomness. Colsample_bytree does so by randomly sampling a fraction of column data for each tree; subsample randomly samples a fraction of observations for each tree. The colsample_bytree default =1; the subsample default is = 1. With respect to model complexity control, gamma, max_depth, and min_child_weight are useful. Gamma has a default = 0, which means gamma regularization is not applied. Gamma regularizes across trees (rather than within trees), preventing trees from adding nodes unless improvement exceeds or is equal to gamma(Exchange, 2019a)(https://stats.stackexchange.com/questions/418687/gamma-parameter-in-xgboost). Max_depth (default= 6) and min_child_weight (default =1) are referred to as within tree controls. The former limits tree depth and shallow trees are less likely to overfit the data. Min_child_weight (default = 1) inhibits tree splitting if a node weight (based on second order derivatives) is less than the min_child_weight.

Finally, nrounds (number of iterations, default= 100) is similar to the number of trees in random forest, and the eta hyper-parameter (also known as a shrinkage parameter; default = 0.3) controls learning rate. Setting the eta value higher slows down the learning rate of the model. However, higher values of eta result in fewer corrections per every tree added to the model and so reduce likelihood of overfitting, but may also result in underfitting. Finding the appropriate eta value may require assessing a range of values. Eta, like most XGBoost parameters, can be optimized with resampling (e.g. using the caret package caret (M. Kuhn, 2019, March, 3)).

Data for XGBoost must be in numeric format. Prior to model implementation, categorical variables (predictors/features) need to be transformed to binary (0 1) form; this is often done using one-hot encoding. Further, data should conform to a matrix format, and the dependent or outcome variable is simply a vector of labels. While XGBoost is algorithm of high complexity, it is renowned for its speed, scalability and accuracy. Garnering a plethora of accolades in recent years, winning solutions employing XGBoost have proven dominant

in machine learning challenges such as Kaggle, the KDD Cup, and in the Higgs machine learning challenge (ATLAS, 2014; Kaggle, 2018a; KDnuggetts, 2017).