Physica Medica 69 (2020) 233–240

Contents lists available at ScienceDirect

Physica Medica

journal homepage: www.elsevier.com/locate/ejmp

Original paper

Machine learning methods for optimal prediction of motor outcome in Parkinson's disease

Mohammad R. Salmanpour[a], Mojtaba Shamsaei[a], Abdollah Saberi[b], Ivan S. Klyuzhin[c], Jing Tang[d], Vesna Sossi[e], Arman Rahmim[e,f,g,*]

[a] Department of Energy Engineering and Physics, Amirkabir University of Technology, Tehran, Iran
[b] Department of Computer Engineering, Islamic Azad University, Tehran, Iran
[c] Department of Medicine, University of British Columbia, Vancouver, BC, Canada
[d] Department of Electrical & Computer Engineering, Oakland University, Rochester, MI, USA
[e] Department of Physics & Astronomy, University of British Columbia, Vancouver, BC, Canada
[f] Department of Radiology, University of British Columbia, Vancouver, BC, Canada
[g] Department of Radiology, Johns Hopkins University, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Parkinson's disease
Outcome prediction
Motor symptom (MDS-UPDRS-III)
Predictor and feature subset selection algorithms

ABSTRACT

Purpose: It is vital to appropriately power clinical trials towards discovery of novel disease-modifying therapies for Parkinson's disease (PD). Thus, it is critical to improve prediction of outcome in PD patients.

Methods: We systematically probed a range of robust predictor algorithms, aiming to find best combinations of features for significantly improved prediction of motor outcome (MDS-UPDRS-III) in PD. We analyzed 204 PD patients with 18 features (clinical measures; dopamine-transporter (DAT) SPECT imaging measures), performing different randomized arrangements and utilizing data from 64%/6%/30% of patients in each arrangement for training/training validation/final testing. We pursued 3 approaches: i) 10 predictor algorithms (accompanied with automated machine learning hyperparameter tuning) were first applied on 32 experimentally created combinations of 18 features, ii) we utilized Feature Subset Selector Algorithms (FSSAs) for more systematic initial feature selection, and iii) considered all possible combinations between 18 features (262,143 states) to assess contributions of individual features.

Results: A specific set (set 18) applied to the LOLIMOT (Local Linear Model Trees) predictor machine resulted in the lowest absolute error 4.32 ± 0.19, when we firstly experimentally created 32 combinations of 18 features. Subsequently, 2 FSSAs (Genetic Algorithm (GA) and Ant Colony Optimization (ACO)) selecting 5 features, combined with LOLIMOT, reached an error of 4.15 ± 0.46. Our final analysis indicated that longitudinal motor measures (MDS-UPDRS-III years 0 and 1) were highly significant predictors of motor outcome.

Conclusions: We demonstrate excellent prediction of motor outcome in PD patients by employing automated hyperparameter tuning and optimal utilization of FSSAs and predictor algorithms.

1. Introduction

Parkinson's disease (PD) is the second most prevalent neurodegenerative disorder after Alzheimer's disease, affecting over 1% of individuals over the age of 60 [1]. PD is characterized by neuronal loss in the substantia nigra with the loss of dopaminergic terminals in the basal ganglia [2–4], resulting in a series of motor and non-motor symptoms such as resting tremor, rigidity, bradykinesia, postural instability and autonomic dysfunction [5]. There are presently no proven disease-modifying therapies for PD. Substitutive treatments with levodopa [6] and dopaminergic agonists [7] remain as the most effective treatments for PD [8]. However, these are associated with adverse reactions, including motor and psychic complications. Furthermore, these only provide temporal improvement through relief from early symptoms and do not stop disease progression [9].

There are currently two main quantitative measures for PD progression, including: 1) the Hoehn and Yahr scale [10], and 2) the Unified Parkinson's Disease Rating Scale PD Rating Scale (UPDRS) [11]. The Movement Disorder Society (MDS) sponsored UPDRS consists of 65 items and has four parts: first part involves Non-Motor Experiences of Daily Living, the second part is on Motor Experiences of Daily Living, the third part is on Motor Examination, and the fourth part is on Motor

[*] Corresponding author at: Department of Radiology, University of British Columbia, Vancouver BC, Canada.
E-mail address: arman.rahmim@ubc.ca (A. Rahmim).

https://doi.org/10.1016/j.ejmp.2019.12.022
Received 12 September 2019; Received in revised form 11 December 2019; Accepted 23 December 2019

Complication. This also includes the use of sum of scores for each part in preference to a total score of all parts [12]. Parashos et al. [13] claimed that UPDRS II and III were more reliable than some other questionnaire tests as measures of disease activity at baseline, and of disease progression in early, untreated PD subjects. Thus, the PD motor symptom severity was quantified using the MDS-UPDRS Part III [14]. Noyce et al [15] showed that patients with higher UPDRS III score have a higher risk factor for declining in motor disturbances.

Towards the discovery of disease-modifying therapies, there is an important need to establish biomarkers of disease progression [16]; e.g. this is the primary aim of the Parkinson's Progressive Marker Initiative (PPMI) [17]. New biomarkers of PD would allow better designs of disease modifying trials, with greater power to ascertain efficacy [18]. PD is a heterogeneous disease; for example, in a multicenter study of PD, Hely et al. [19] showed that during 10 years of observation (excluding patients developing signs and symptoms atypical for PD in follow-up), 9 out of 126 patients progressed to confinement to bed or a wheelchair unless aided, whereas 13 patients remained without significant functional restriction. Novel biomarkers may enable more accurate stratification of PD based on the expected prognosis. There is also a significant interest in using the biomarkers for prediction of disease outcome [20–25,28], to properly adapt clinical trial studies as applied to appropriate patients.

In addition to identifying individual novel PD biomarkers, a promising area of research is to use machine learning and data mining to build predictive models of PD progression, using multiple biomarkers as inputs. Approaches based on machine learning aim to build classification or prediction algorithms automatically by capturing statistically robust patterns present in the analyzed data [26]. In our past efforts we used radiomic (texture) measures to show improved correlation with clinical measures [29] and, when combined with clinical measures, improved prediction of motor outcome (MDS_UPDRS III) [20]. Furthermore, our recent studies confirmed significant improvement of outcome prediction based on discovering of patterns in images using deep learning (see the Section 4) [30,31]. Emrani et al [32], using machine learning methods, introduced a new combination of PD biomarkers (SCNA-3UTR, total cholesterol, SBR in left and right putamen, α-Syn, GUSB, DJ-1 and UPSIT are identified as significant diagnostic biomarkers), showing significantly improved prediction of outcome (total MDS_UPDRS) in PD including discovery of the high effect of CFS in prediction.

In previous work, Ramani et al. [27] conducted a trial based on prediction of motor and total UPDRS scores from voice measures. Data mining techniques based on feature relevance analysis and classification were utilized towards identifying severity of disease. The random tree classification algorithm was used, producing excellent classification results, also showing that feature selection algorithms helped in prediction of correct class labels. Similarly, Nilashi et al. [33] used computational tools of data mining to improve prediction of early PD using voice recordings from potential patients. Their results depicted that combining clustering, PCA (Principle Component Analysis), and SVR (Support Vector Regression) can dramatically improve accuracy of PD prediction.

Indeed, using only the most relevant features may improve the prediction accuracy. Furthermore, most predictor algorithms are not able to work with a large number of input features, and thus it is necessary to select the optimal features to be used as inputs, as we have demonstrated in a different prediction task [34]. The process of feature selection can be performed either manually or automatically using feature subset selector algorithms (FSSAs). Our present effort includes extensive search for and focus on optimal combination of machine learning methods and FSSAs for the task of predicting motor outcome in PD patients. Using multiple clinical data measured at baseline (year 0) and year 1, we set to predict the severity of motor symptoms (MDS-UPDRS-III) at year 4 following enrollment of patients with *de novo* PD. This work includes three parts. In the first part, we create manual combinations of features, and each combination is assessed using predictor algorithms. In the second part, we employ Feature Subset Selector algorithms (FSSA) for pre-selecting more effective features, and each combination selected by FSSAs is assessed using predictor algorithms. In the final part, we create all possible combinations for the 18 features, to assess relative contributions of each feature to the prediction task.

2. Methods and materials

2.1. Machine learning methods

Two group of algorithms were employed: 1) Predictor algorithms; and 2) Feature Subset Selector algorithms (FSSAs).

2.1.1. Predictor algorithms and utilizing automated machine learning hyperparameter tuning to adjust parameters of them

A range of optimal predictor algorithms were selected amongst various families of learner and regressor algorithms. These are all listed in the Supplement (Part II, Section 1). Specifically, we selected 10 predictor algorithms: 1) LOLIMOT (Local Linear Model Trees) [35,36], 2) RBF (Radial basis Function) [37], 3) MLP-BP (Multilayer Perceptron-Back propagation) [38,39], 4) LASSOLAR (Least Absolute Shrinkage and Selection Operator – Least Angle Regression) [40,41], 5) RFA (Random Forest Algorithm) [42,43], 6) RNN (Recurrent Neural Network) [44,45], 7) BRR (Bayesian Ridge Regression) [46–48], 8) DTC (Decision Tree Classification) [49–51], 9) PAR (Passive Aggressive Regression) [52–54], 10) Thiel-Sen Regression [55–57] and 11) ANFIS (Adaptive neuro fuzzy inference system) [58,59]. In this work, we automatically adjusted intrinsic parameters such as the number of neurons and number of layers in the predictor algorithms etc. via automated machine learning hyperparameter tuning. Such hyperparameter tuning was used in various algorithms such as LOLIMOT, RBF, RNN, MLP-BP, RFA and ANFIS so that algorithm parameters were automatically optimized given one of the data arrangements prior to formal training/validation/test processes. Automated tuning, implemented with our own in-house code, performs an error minimization search scheme, seeking to optimize the hyperparameters starting with random initialization, pursuing a systematic trial-and-error search scheme for tuning the parameters. All applied algorithms except LASSOLAR, BRR, DTC, PAR and Thiel-Sen Regression were implemented in MATLAB R 2016 b platform. The remaining algorithms were implemented in the Python 3.7.2 platform.

2.1.2. Utilizing FSSAs for feature selection

6 FSSAs were employed and compared to select the most effective features (see the supplement (Part II, Section 2) for more details): 1) GA (Genetic Algorithm) [60,61], 2) ACO (Ants Colony Optimization) [62,63], 3) PSO (Particle Swarm Optimization) [64,65], 4) SA (Simulated Annealing) [66,67], 5) DE (Deferential Evolution) [68,69], and 6) NSGAII (Non-dominated Sorting Genetic Algorithm) [70,71]. All algorithms aimed to minimize the prediction error by selecting the best combination of features, while NSGAII additionally aimed to reduce number of features. These are also elaborated in the supplement. Subsequently, the selected features were tested on the predictor algorithms mentioned in Section 2.1.1. All applied optimization algorithms were implemented in MATLAB R 2016 b platform.

2.2. Longitudinal patient data

Data were extracted from the PPMI database (www.ppmi-info.org/data). As predictors, we considered the following 18 clinical features: (1–6) MDS-UPDRS, parts I, II and III in year 0 and 1, (7–8) demographics (age, sex), (9–16) dopamine transporter (DAT) SPECT images measures, namely putamen as well as caudate uptake, both left and right in years 0 and 1, and (17–18) disease duration (DD), taken with

Table 1
Various combinations between eighteen features.

Features	SET 1	SET 2	SET 3	SET 4	SET 5	SET 6	SET 7	SET 8	SET 9	SET 10	SET 11	SET 12	SET 13	SET 14	SET 15	SET 16	SET 17	SET 18	SET 19	SET 20	SET 21	SET 22	SET 23	SET 24	SET 25	SET 26	SET 27	SET 28	SET 29	SET 30	SET 31	SET 32
UPDRS I - year 0		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨		▨
UPDRS I - year 1		▨		▨		▨		▨		▨		▨						▨		▨		▨		▨		▨		▨				
UPDRS II - year 0			▨	▨			▨	▨			▨	▨			▨	▨			▨	▨			▨	▨			▨	▨			▨	▨
UPDRS II - year 1			▨	▨			▨	▨			▨	▨							▨	▨			▨	▨			▨	▨				
UPDRS III - year 0	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
UPDRS III - year 1	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨					▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
Cau-Right - year 0	▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨				
Cau-Left - year 0	▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨				
Pu-Right - year 0	▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨				
Pu-Left - year 0	▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨					▨	▨	▨	▨				
Cau-Right - year 1	▨	▨	▨	▨													▨	▨	▨	▨												
Cau-Left - year 1	▨	▨	▨	▨													▨	▨	▨	▨												
Pu-Right - year 1	▨	▨	▨	▨													▨	▨	▨	▨												
Pu-Left - year 1	▨	▨	▨	▨													▨	▨	▨	▨												
Age	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
Gender	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨
DD-Diag	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																
DD-sympt	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨																

respect to time of diagnosis (DD-diag.) as well as time of appearance of symptoms (DD-sympt.). DAT SPECT imaging is elaborated in the supplementary material. For consistency, we only included patients who were off medication (e.g. Levodopa/dopamine agonist) for > 6 h prior to testing/imaging [17]. These selection criteria yielded a group of 204 PD subjects (149 males, 55 females; average age 67.6 ± 10.0 years at time of baseline examinations; range 39–91 years), with a widely distributed year 4 outcome MDS-UPDRS-III (mean outcome value 31.4 ± 10.6 years, range 8 to 77). For better consistency, MDS-UPDRS scores were averaged within ± 6 months. In this study, we considered three approaches included:

2.2.1. Experimental combinations

As first approach, 32 combinations of the 18 features were manually selected and considered (Table 1), and each combination was assessed using 10 (of 11 above-mentioned) predictor algorithms (mentioned in Section 2.1.1) (ANFIS was excluded because it is not appropriate for combinations with relatively large number of features).

2.2.2. Combinations selected by FFSAs

In our second approach, we utilized 6 FSSAs (mentioned in Section 2.1.2) for selecting optimal combinations from 18 features. Thus, 6 optimal combinations selected by FSSAs considered (Table 2), and each combination with 204 patients was assessed using 11 predictor algorithms (mentioned in Section 2.1.1).

2.2.3. Testing all possible combinations

In our third approach, we considered 262,143 (all) possible combinations for the 18 features. All combinations were exhaustively considered and applied to LOLIMOT which was found to be the best predictor algorithm (as later shown in Sections 3.1 and 3.2), and predictive performance of all combinations were assessed using MAE (Mean Absolute Error). This method assesses which features are independent predictors of outcome that cannot be substituted by combinations of other features. Though such an approach may not be feasible when significantly more features are considered, in this work it can help shed light on effectiveness of feature selection methods, and on the relative contributions of individual features.

2.3. Analysis procedure

Processing was performed in three stages. In the first stage, we manually selected different combinations within 204 subjects (mentioned in Table 1). We generated 4 randomized arrangements of the dataset. In each run, we allocated 64% of total patients to training (to avoid under-fitting), 6% of total patients (or about 9% of training process patients) to training validation in order to minimize over-fitting from the training step, and 30% (for reliability) to final testing (repeating this process 4 times). Subsequently, for each predictor algorithm, the MAE as well as STD (standard deviation) of the predicted vs. true MDS-UPDRS III scores in year 4 were reported. These were computed from the final test sets in the randomized arrangements for more reliable and appropriate assessment of our results.

In the second stage, we first performed systematic feature selection via FSSAs of the dataset with 204 patients and 18 features, identifying optimal combinations of features for prediction of outcome. We then created new combinations from the 204 patients (for each mentioned randomized arrangement) based on optimal combinations (Table 2). For each randomized arrangement with optimal features, we considered ~64% for training, ~6% for training validation and ~30% for final testing. The MAE and STD from the final test sets were then reported.

In the third stage, we considered all possible combinations of the 18 features for the 204 patients. They were then applied to LOLIMOT, which found as the best predictor algorithm in prior sections, for finding independent features. A single arrangement from each combination (~64% for training, ~6% for training validation and ~30% for final testing) was used and assessed via MAE.

3. Results

3.1. First stage analysis involving manually selected sets of input features

Results of first approach (mentioned in Section 2.2.1) are shown in Fig. 1. The figure demonstrates that LOLIMOT resulted in the best performance (i.e. lowest prediction MAE) over a wide range of combinations. The best results for LOLIMOT were achieved for sets 18, 21 and 22, and some algorithms also reached similar results for those sets, but were less consistent when including other range of features. Overall, in prediction of year 4 MDS-UPDRS-III (outcome range [8–77]), MAE as low as 4.32 ± 0.19 were achieved. The best results were observed when features MDS-UPDRS I (year 0 and 1), MDS-UPDRS-III (years 0 and 1), putamen as well as caudate uptake (both left and right; years 0 and 1), age and gender were in the sets. Some combinations such as sets 1, 3, 6, 7, 8, 19, 20, 23 and 24 also had good results, although they were less consistent than sets 18, 21 and 22.

Fig. 2 shows one of the results obtained by LOLIMOT where the Y axis is the predicted outcome and X axis reflects the true outcome, in addition to comparison to performance by DTC. Overall, LOLIMOT was seen as the best predictor algorithm.

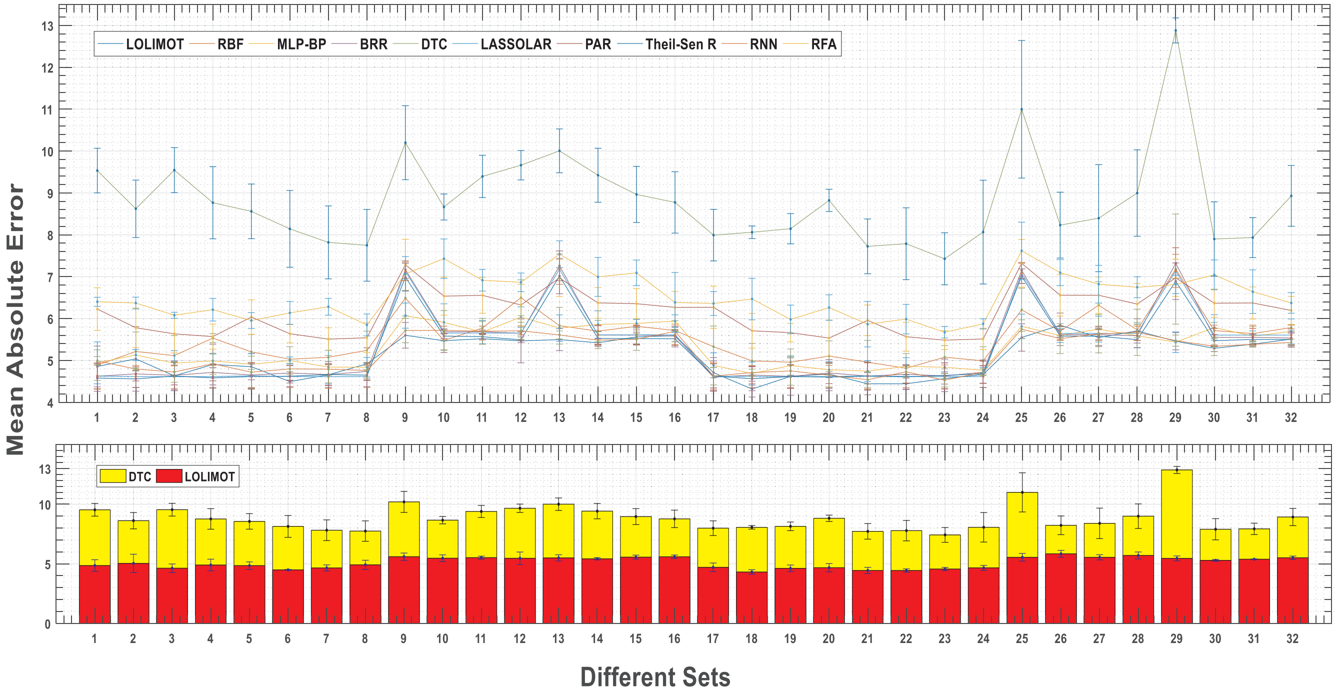

Fig. 1. (Top) MAE for outcome prediction from 10 predictor algorithms, each from 32 combinations shown in Table 1. ANFIS was excluded from results due to significantly poorer performance. (Bottom) Amongst the 10 algorithms, the best and worst performers are shown as bar plots. The error bars represent the standard deviation. In each part, X axis lists the predictor algorithms and Y axis shows their MAE.

3.2. Second stage analysis involving subsets of input features selected by FSSAs

In the second approach (mentioned in Section 2.2.2), the features selected using different FSSAs (mentioned in Section 2.1.2), are shown in Table 2.

These featured combinations selected via the selector algorithms were then applied to the 11 predictor algorithms as mentioned in Section 2.1.1, with the results plotted in Fig. 3.

As it was shown, most subset feature selector algorithms selected combinations leading to prediction errors less than 5. The performances were best for features selected by GA and ACO (MDS – UPDRS I – year 1, MDS-UPDRS III – year 1, Left putamen uptake – year 1, Age, Gender). In the case of LOLIMOT predictor algorithm, lowest MAE of 4.15 ± 0.46 was reached, though other algorithms such as LASSOLAR, Bayesian Ridge, Theil Sen R and RNN also performed well. Plot of the results of GA and ACO as applied to LOLITMOT are shown in Fig. 4. The NSGAII selected 4 among 18 features, and subsequently, prediction via LOLIMOT reached a MAE ~4.48 ± 0.22, while other algorithms such as RBF, MLP-BP, LASSOLARs, BRR, Thiel-Sen R and RNN also resulted in approximately similar results compared to LOLIMOT. The DE and PSO selected features similar to GA, so that when those applied to predictor algorithms, Thiel-Sen R resulted in ~4.61 ± 0.26, although some other algorithms such as LOLIMOT, LASSOLAR, RNN and BRR performed nearly similarly. The SA selected number feature similar to the GA, so while those selected features applied to all predictor algorithms, the LASSOLAR reached MAE ~4.51 ± 0.26, although other algorithms such as LOLIMOT, MLP-BP, BRR, Thiel-Sen R and RNN similarly reached an appropriate result. Results for GA, ACO and NSGAII were comparable ($p > 0.05$; paired *t*-test & Friedman test) relative to the lowest errors in the original set, while some errors for DE, PSO and SA were statistically worse (see Supplementary materials, section III, Supplemental Tables 1&2, bolded p-values). Thus, as also shown in our previous work on prediction of cognitive decline (Montreal Cognitive Assessment; MoCA score) [34], FSSAs were also able to effectively select optimal combinations from large datasets.

Between MDS_UPDRS III – year 0 and outcome, Pearson and Spearman correlation coefficients of 0.63 and 0.67 were reached,

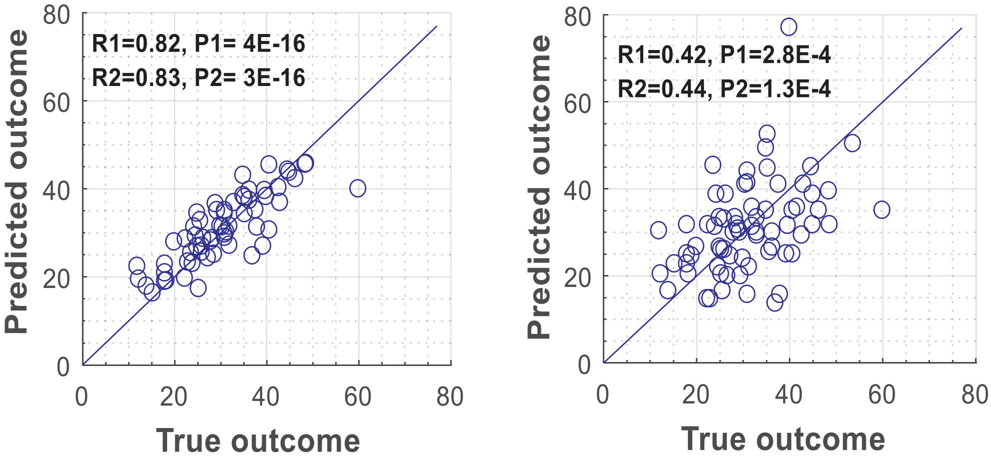

Fig. 2. Plots of outcome prediction from (Left) LOLIMOT and (Right) DTC for 61 patients, making use of features in set 18. X axis shows the actual scores and Y axis depicts the predicted MDS_UPDRS III scores. R1 is the Pearson product-moment correlation coefficient and R2 is the Spearman's rank correlation coefficient between true vs. predicted outcomes. P1 and P2 are p-values corresponding to R1 and R2, respectively ($p < 0.05$ is significant).

Table 2
Feature subsets selected by 6 Feature Subset Selector Algorithms (FSSAs).

Selectors	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
GA	UPDRS I – year 1	UPDRS III – year 1	Left putamen uptake – year 1	Age	Gender
ACO	UPDRSI – year 1	UPDRSIII – year 1	Left putamen uptake – year 1	Age	Gender
PSO	UPDRSII – year 0	UPDRSIII – year 1	Left putamen uptake – year 1	DD-diag	Gender
SA	UPDRSII – year 0	UPDRSIII – year 1	Left putamen uptake – year 0	DD-diag	Gender
DE	UPDRSII – year 0	UPDRSIII – year 1	Left putamen uptake – year 1	DD-diag	Gender
NSGAII	UPDRS II – year 0	UPDRS III – year 1	Age	Gender	–

respectively, while, between MDS_UPDRS III – year 1 and outcome, these became 0.76 and 0.72 (see Supplementary materials, section III, Supplemental Tables 3–6, Bolded). Moreover, due to high correlations between MDS_UPDRS III – year 0 and 1 (Pearson: 0.81; Spearman: 0.83; p-values < 0.001), the FSSAs selected one of both scores (MDS_UPDRS III – year 1 instead of year 0). In addition, using predictor algorithms LOLIMOT, RBF and RNN, we performed a prediction task using solely the MDS_UPDRS III score at year 0 or year 1 and also using both years. The comparison of those results against results by GA and ACO (5 important features selected) are shown in Table 3.

Overall, by switching from sole usage of MDS_UPDRS III in year 0 (left column) to integrating the important five predictive features (right column), significantly lower errors were reached (all p-values < 0.02 using paired *t*-test; and < 0.03 using nonparametric Friedman test). In addition, for LOLIMOT and RNN, there were also significant improvements compared to sole usage of MDS_UPDRS III in year 1, while improvements for RNN were also significant compared to usage of both years 0 and 1.

3.3. Third stage analysis studying all possible combinations of 18 features

In the last approach (in Section 2.2.3), we studied all existing combinations, as applied to LOLIMOT. Fig. 5 shows how much each feature contributed to prediction of outcome, considering all possible

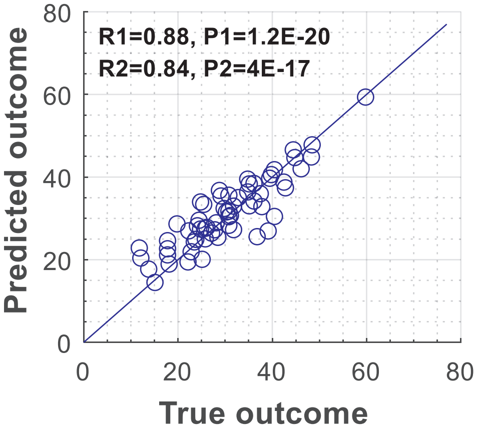

Fig. 4. A typical predictive performance of LOLIMOT (following GA and ACO pre-selection of features) for 61 patients. X axis shows the actual scores and Y axis depicts the predicted MDS_UPDRS III scores. R1 is the Pearson product-moment correlation coefficient and R2 is the Spearman's rank correlation coefficient between true vs. predicted outcomes. P1 and P2 are p-values ($p < 0.05$ is significant).

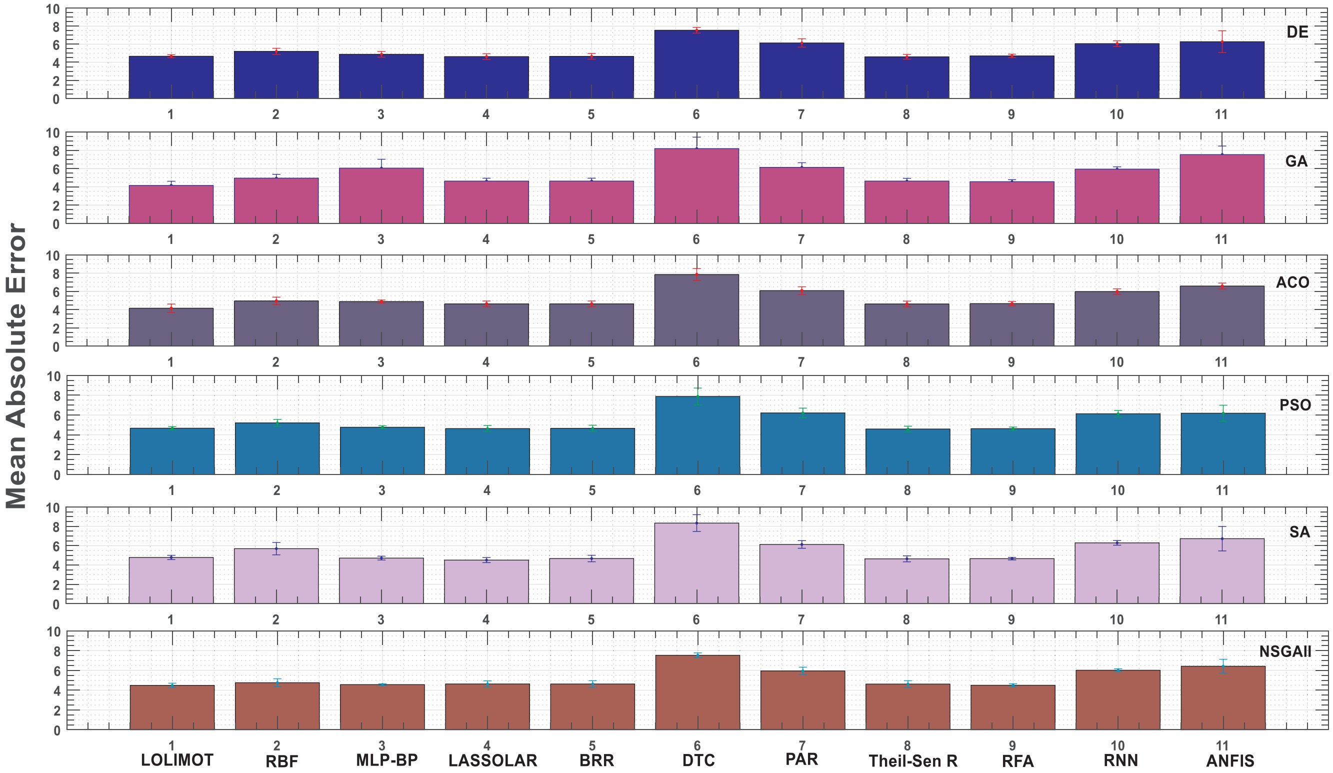

Fig. 3. Performance plots for application of the 6 FSSAs followed by 11 predictor algorithms. The error bars represent the standard deviation. In each part, X axis lists the predictor algorithms and Y axis shows their Mean absolute errors.

Table 3
Performance results for prediction task when using MDS_UPDRS III in year 0 alone vs. year 1 alone vs. both years vs. using all five important features.

	Mean Absolute Error (only MDS_UPDRS III – year 0)	Mean Absolute Error (only MDS_UPDRS III – year 1)	Mean Absolute Error (both MDS_UPDRS III – year 0 and 1)	Mean Absolute Error (all five important features)
LOLIMOT	5.39 ± 0.20	4.60 ± 0.08	4.54 ± 0.17	4.15 ± 0.46
RBF	6.10 ± 0.44	5.25 ± 0.47	5.03 ± 0.33	4.95 ± 0.42
RNN	5.65 ± 0.10	5.08 ± 0.14	5.08 ± 0.11	4.56 ± 0.22

combinations of features excluding vs. including each of the specific 18 features. All results were statistically significant (Paired *t* test, even after Bonferroni multiple testing correction) except for feature 16 (gender), while using non-parametric Friedman test, all those except for feature 4 (MDS_UPDRS II – year 1) were statistically significant; this statistical significance is because of the very large number of comparisons (131071) in each set (see Supplementary materials, section III, Supplemental Table 7). Nonetheless, only features 5 and 6 (UPDRS III – years 0 and 1) which reduce errors (negative sign) by a magnitude of ~1 are likely clinically significant, given that magnitude of overall prediction errors reached in our work are ~4.

4. Discussion

Prediction of PD progression is an important and challenging problem [72]. Accurate prediction of outcome in PD has a number of benefits [73], including: 1) helping to make social and occupational decision for patients recently diagnosed with PD and to improve their likely future physical functioning, 2) potentially better understanding processes and pathways associated with disease progression, 3) reinforcing design and interpretation of clinical trials of neuroprotective and symptomatic therapy. In this work we explored a range of predictor algorithms and aimed to find the best combinations of features to result in significant improvement in prediction of outcome in PD. In the first part of our study, 32 combinations of 18 conventional features were experimentally selected for various arrangements of 204 PD subjects), and various predictor algorithms were applied to the combinations. In the second part of study, subset selector algorithms were used for selecting the best combinations between all features [34].

In the case of manual feature combinations (Section 3.1), LOLIMOT resulted in best performance. Overall, in prediction of year 4 MDS-UPDRS-III (outcome range [8–77]), predictions errors (MAE) as low as 4.32 ± 0.19 were achieved. Such prediction performance far exceeds results in other prior works where absolute errors of the order of 9 were obtained when using similar features [20]. Best results were observed when features MDS-UPDRS I (year 0 and 1), MDS-UPDRS-III (years 0 and 1), putamen as well as caudate uptake (both left and right; years 0 and 1), age and gender were in the sets. Conventional DAT SPECT images were seen to have some (but not large) effect. Some combinations such as sets 1, 3, 6, 7, 8, 19, 20, 23 and 24 also had good results, although they were less consistent than sets 18, 21 and 22.

In the second part of our efforts, we employed 6 FSSAs for pre-selecting more effect features. The features selected through these different algorithms are shown in Table 2. As shown in Fig. 3, all FSSAs selected acceptable combinations, so that the MAE decreased to 4.15. Most FSSAs selected combinations leading to prediction errors less than 5. GA and ACO selected 5 features (MDS – UPDRS I – year 1, MDS-UPDRS III – year 1, Left putamen uptake – year 1, Age, Gender), and upon application of the selected features, LOLIMOT reached the least MAE ~4.15. DE and PSO selected features similar to GA, so that when those were applied to predictor algorithms, Thiel-Sen R reached the least MAE ~4.61. SA selected features similar to GA, and for the selected features, when applied to the predictor algorithms, LASSOLAR reached the smallest MAE ~4.51. NSGAII selected 4 among 18 features, so that when applied to predictor algorithms, LOLIMOT reached the least MAE ~4.48. Some predictor algorithms in each approach resulted in values similar to the smallest errors. Most errors thus obtained using GA, ACO and NSGAII were comparable (not statistically different) relative to the lowest error obtained by the first approach (4.32 ± 0.19), while others were significantly different. In other words, using FSSAs accelerated selection of optimal combinations, which we have observed also in different prediction tasks [34]. Our results confirm existing some new combinations that have not yet been investigated.

Due to presence of high correlations between MDS_UPDRS III scores in year 0 vs. year 1, as well as higher correlation between MDS_UPDRS III – year 1 vs. outcome relative to MDS_UPDRS III – year 0 vs. outcome, all FSSAs selected MDS_UPDRS III – year 1 instead of MDS_UPDRS III year 0 (Table 2). As shown and discussed in the context of Table 3 for LOLIMOT, RBF and RNN, we reached errors statistically significantly lower when switching from employing MDS-UPDRS III – year 0 to integrating all 5 important features. There were also improvements relative to usage of MDS-UPDRS – year 1 score only (significant for both LOLIMOT and RNN), or usage of both year 0 and year 1 (significant for RNN). Overall, poorer performances resulted from usage of MDS-UPDRS III – year 0 or 1 scores only, or their combined usage, compared to using the most important five features as selected by GA and ACO.

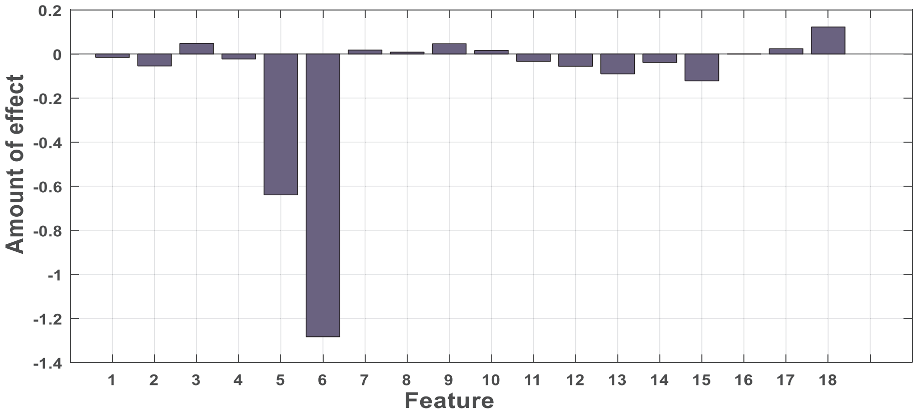

Fig. 5. Evaluation of the effect of each feature in all combinations of features including vs. excluding it. Mean difference between absence and presence of each feature. Negative sign means that with adding specific feature, the errors decrease and positive sign is inverse. The 18 features are listed in Table 1.

Importantly, we saw that not using imaging information (features selected by NSGAII) can lower performance to around 4.48 (in prediction error) which was not statistically significantly than our 4.32 performance (see Supplementary materials, section III, Supplemental Table 1&2). In our other investigations we also found that conventional imaging measures do not correlate well with clinical measures (correlation coefficient = −0.008, p-value = 0.94) [29] nor improve prediction (9 ± 0.88) [20]. However, radiomics analysis of DAT SPECT images, going beyond conventional imaging measures, was seen to provide significant improvements (4.12 ± 0.43, $p < 0.001$) in both tasks. Further, in our recent ongoing efforts involving deep learning based prediction of outcome [30], significant improvements (3.22 ± 2.71, p-value < 0.05) were observed, involving implicit discovery of patterns in images, although fluctuation of results (STD = 2.71) were high. Similarly, another study of ours [31] only used MDS_UPDRS III and DAT SPECT images in year 0, showing significant improvement (Accuracy: 70.6%+/-7.7%, p-value < 0.001) of outcome prediction based on discovering of patterns in images using deep learning. In other words, our present research indicates that there is a need to move beyond conventional imaging metrics for improved prediction of outcome.

In our final investigation in the present work, we created all existing states (existing combinations). Fig. 5 showed an individual figure how much each feature effected prediction of outcome, considering all possible combinations of features excluding vs. including each feature. It is particularly appropriate to assess which features are independent predictors of outcome that cannot be substituted by combinations of other features. It is clearly seen that features 5 and 6 (MDS-UPDRS-III years 0 and 1) were the most significant predictors of outcome with largest effects. Other features, though reducing errors statistically significantly (except for gender based on Paired *t* test or except for MDS_UPDRS II year 1 based on Friedman test), did not make actually notable reduction in errors that would foresee a clinical advantage. Inclusion of feature 18 (disease duration from symptoms) actually degraded prediction performance; this may be related to the very subjective nature of this measure, suggesting that measurement of disease duration from diagnosis may be a more valuable measure.

We wish to also importantly note that the training validation step after each (epoch of) training (prior to testing) aims to tackle potential over-fitting. At the same time, even with this, as more and more machines are compared and tested, it becomes increasingly plausible that over-fitting may occur. To address this, we note that in our study, LOLIMOT persisted in producing good results even when we reduced data dimension via feature selection, followed by predictive modeling (see the second section of our results); i.e. even with a significantly reduced feature set, we obtained similar results. Overall, LOLIMOT was seen as the best predictor algorithm, GA and ACO as the best feature subset selector algorithms, and also MDS-UPDRS III – years 0 and 1, MDS-UPDRS II – year 1, MDS-UPDRS I – years 0 and 1, Age, and DAT SPECT putamen as well as caudate uptake – year 1 as the best predicting features of outcome. Comparison of current work and prior related works includes the following: our predictive performances are dramatically better than previous efforts when no images are used, important combinations were discovered for the multiple features, and effect of each feature on outcome predictions was separately investigated.

The limited size of a dataset is a limiting factor in outcome prediction; as such, to maximize our numbers, we had to select a set of 204 patients for which imaging data was available for all patients. We could not find more patients within the PPMI dataset even after optimal feature selection for additional testing. In our work, we used features subset selection algorithms to reduce the number of features (for size reduction) to avoid over-fitting, although it was possible to utilize extraction algorithms such as PCA which we hope to explore in future work. At the same time, we believe feature selection might be more clinically informative than feature extraction, providing insights as to which features are most important, whereas in feature extraction, features are combined and transformed into new dimensions and thus may not be easily interpretable.

5. Conclusion

This work explored a range of predictor algorithms and aimed to find the best combinations of features to result in improvement in prediction of outcome in PD. In the first part of our study involving 204 patients, 32 combinations of 18 conventional features were experimentally selected, and various predictor algorithms were applied to the combinations. Mean Absolute Errors as low as 4.32 ± 0.19 (in prediction of UPDRS-III motor performance in year 4) were reached via LOLIMOT, while some other algorithms in some combinations also had good results with errors less than 5 (for range of predicted values [8,77]). In the second part of study, subset selector algorithms were used for selecting the best combinations between all features, and GA and ACO selector algorithm selected the best combinations further lowering error when combined with LOLIMOT for prediction. Selected features by GA and ACO (UPDRS I – year 1, UPDRS III – year 1, left putamen Uptake – year 1, Age, Gender) had positive effect on prediction of outcome (MAE ~4.15 ± 0.46), although other FSSAs also reached acceptable combinations so that MAE in some predictor algorithms were below 4.7. This is in comparison to previous works utilizing similar features that attained errors of around 9. Third part of our work investigated the independent effect of different features in prediction of outcome. It was seen that MDS-UPDRS III years 0 and 1 were especially important predictors of outcome.

Data and code availability

All code (included predictor algorithms and Feature Sunset Selector Algorithms) and all datasets are publicly shared at: https://github.com/salmanpoor/Machine-Learning-Algorithms_M.R.Salmanpour-A.Saberimanesh-A.Rahmim.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The project was supported by the Michael J. Fox Foundation, including use of data available from the PPMI—a public-private partnership—funded by The Michael J. Fox Foundation for Parkinson's Research and funding partners (listed at www.ppmi-info.org/funding-partners). This work was also supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ejmp.2019.12.022.

References

[1] Shi C, Zheng Z, et al. Exploring the effects of genetic variants on clinical profiles of Parkinson’s disease assessed by the unified Parkinson’s disease rating scale and the Hoehn-Yahr Stage. PLoS One 2016:1–11.

[2] Brooks D, Ibanez V, et al. Differing patterns of striatal F-18 Dopa uptake in Parkinsons-disease, multiple system atrophy, and progressive supranuclear palsy. Ann Neurol 1990;28:547–55.

[3] Garnett E, Lang A, et al. A rostrocaudal gradient for aromatic acid decarboxylase in the human striatum. Can J Neurol Sci 1987;14:444–7.

[4] Stoessl A, Martin W, et al. Advances in imaging in Parkinson’s disease. Lancet Neurol 2011;10:987–1001.

[5] Jankovic J. Parkinson's disease: clinical features and diagnosis. J Neurol Neurosurg Psychiatry 2008;79(4):368–76.
[6] Antonini A, Isaias I, et al. Duodenal levodopa infusion for advanced Parkinson's disease: 12-month treatment outcome. Int Parkinson Movem Disorder Soc 2007;22(8):1145–9.
[7] Tippmann-Peikert M, Park G, et al. Pathologic gambling in patients with restless legs syndrome treated with dopaminergic agonists legs syndrome treated with dopaminergic agonists. Neurology 2007;68:301–3.
[8] Lang A, Lozano A. Parkinson's disease. Second of two parts. New Engl J Med 1998:1130–43.
[9] Savitt M, Dawson V, et al. Diagnosis and treatment of Parkinson disease: molecules to medicine. JCI 2006;7:1744–54.
[10] Goetz C, Poewe P, et al. Movement disorder society task force report on the Hoehn and Yahr staging scale: status and recommendations the movement disorder society task force on rating scales for Parkinson's disease. Mov Disord 2004;19(9):1020–8.
[11] Mischley L, Lau R, et al. Use of a self-rating scale of the nature and severity of symptoms in Parkinson's Disease (PRO-PD): correlation with quality of life and existing scales of disease severity. NPJ Parkinson's Disease 2017;3:1–20.
[12] Goetz C, Tilley B, et al. Movement disorder society-sponsored revision of the unified Parkinson's disease rating scale (MDS-UPDRS): scale presentation and clinimetric testing results. Mov Disord 2008;23:2129–70.
[13] Parashos S, Luo S, et al. Measuring disease progression in early parkinson disease: the national institutes of health exploratory trials in Parkinson Disease (NET-PD) experience. JAMA Neurol 2014;6:710–6.
[14] Movement Disorder Society Task Force on Rating Scales for Parkinson's Disease. The unified Parkinson's disease rating scale (UPDRS): status and recommendations. Movement 2003;18(7):738–50.
[15] Noyce A, Schrag A, et al. Subtle motor disturbances in PREDICT-PD participants. J Neurol Neurosurg Psychiatry 2017;88:212–7.
[16] Marek K, Jennings D, et al. Biomarkers for Parkison's disease: tools to assess Parkinson's disease onset and progression. Ann Neurol 2008;64:S111–21.
[17] Parkinson Progression Marker Initiative. The Parkinson progression marker initiative (PPMI). Prog Neurobiol 2011;95:629–35.
[18] Post B, Speelman J, et al. Clinical heterogeneity in newly diagnosed Parkinson's disease. J Neurol 2008;5(255):716–22.
[19] Hely M, Morris J, et al. The sydny multicentre study of Parkinson's disease: progression and mortality at 10 years. J Neurol Neurosurg Psychiarty 1999;67:300–3007.
[20] Rahmim A, Huang P, et al. Improved prediction of outcome in Parkinson's disease using radiomics analysis of longitudinal DAT SPECT images. NeuroImage: Clinical 2017;16:539–44.
[21] Fyfe I, Ian. Prediction of cognitive decline in PD. Nat Rev Neurol 2018;14:213–317.
[22] Arnaldi D, De Carli F, et al. Prediction of cognitive worsening in de novo Parkinson's disease: clinical use of biomarkers. Mov Disord 2017;32:1738–47.
[23] Gao C, Sun H, et al. Model-based and model-free machine learning techniques for diagnostic prediction and classification of clinical outcomes in Parkinson's disease. Sci Rep 2018:1–21.
[24] Nieuwboer A, Weerdt WD, et al. Prediction of outcome of physiotherapy in advanced Parkinson's disease. SAGE J 2002;16(8):886–93.
[25] Grill S, Weuve J, et al. Predicting outcomes in Parkinson's disease: comparison of simple motor performance measures and the unified Parkinson's disease rating scale-III. J Parkinson's Dis 2011;1:287–98.
[26] Singh Y, Bhatia P, et al. A review of studies on machine learning techniques. Int J Comp Sci Security 2007;1(1):70–84.
[27] Ramani G, Sivagami G, et al. Feature relevance analysis and classification of Parkinson's disease telemonitoring data through data mining. Int J Adv Res Comp Sci Software Eng 2012;2(3):298–304.
[28] Salmanpour M, Shamsaei M et al., Machine Learning Methods for Optimal Prediction of Outcome in Parkinson's Disease. In: IEEE Nucl. Sci. Symp. Conf. Record, Sydney; 2018.
[29] Rahmim A, Salimpour Y, et al. Application of texture analysis to DAT SPECT imaging: relationship to clinical assessments. NeuroImage: Clinical 2016;12:e1–9.
[30] Leung K, Salmanpour M et al., Using deep-learning to predict outcome of patients with Parkinson's disease. In: IEEE Nucl. Sci. Symp. Conf. Record, Sydney; 2018.
[31] Adams M, Yang B et al. Prediction of outcome in Parkinson's disease patients from DAT SPECT images using a convolutional neural network. In: IEEE Nucl. Sci. Symp. Conf. Record, Sydeny; 2018.
[32] Emrani S, McGuirk A, Xiao W.. Prognosis and diagnosis of Parkinson's disease using multi-task learning, KDD; 2017, p. 1457–66.
[33] Nilashi M, Ibrahim O, Ahani A. Accuracy improvement for predicting parkinson's disease progression. Sci Rep 2016;6(1). https://doi.org/10.1038/srep34181.
[34] Salmanpour M, Shamsaei M, et al. Optimized machine learning methods for prediction of cognitive outcome in Parkinson's disease. Comput Biol Med 2019;111:1–8.
[35] Nelles O, Fink A, Isermann R. Local linear model trees (LOLIMOT) toolbox for nonlinear system identification. Sci Direct (IFAC System Identification) 2000;33(15):845–50.
[36] Martinez-Morales J, Palacios E. Modeling of internal combustion engine emissions by. SciVerse Science Direct 2012;3:251–8.
[37] Arora Y, Singhal A, Bansal A. A study of applications of RBF network. Int J Comp Appl 2014;94(2):17–20.
[38] Alsmadi S, Khalil M, et al. Back propagation algorithm: the best algorithm. IJCSNS Int J Comp Sci Network Sec 2009;9(4):378–83.
[39] Rumelhart D, Geoffrey E, et al. Leaner representations by back-propagating errors. Nature 1986;323(9):533–6.
[40] Fonti V. Feature Selection using LASSO. VU Amsterdam, Amsterdam; 2017.
[41] Efron B, Hastie T, et al. Least angle regression. Annals Statistics 2004;32:407–99.
[42] Breiman L. Random forests. Machine Learn 2001;45:5–32.
[43] Jehad A, Khan R, Ahmad N. Random forests and decision trees. IJCSI Int J Comp Sci Issues 2012;9(5):272–8.
[44] Townley A, Ilchmann M, et al. Existence and learning of oscillations in recurrent neural networks. IEEE Trans Neural Networks 2000;11(1):205–14.
[45] Maknickiene N, Rutkauskas V, et al. Investigation of financial market prediction by recurrent neural network. Innov Infotechnol Sci, Bus Ed 2011;11(2):3–8.
[46] Efendi A. A simulation study on Bayesian Ridge regression models for several collinearity levels. In: AIP Conference Proceedings; 2017.
[47] Bishop CM. Pattern Recognition and Machine Learning, 1th ed., P. J. K. B. S. Michael Jordan, Ed., New York: Springer Science + Business Media, LLC, 233 Spring Street, New York, NY 10013, USA, 2006.
[48] Karabatsos G. Fast Marginal Likelihood Estimation of the Ridge Parameter(s) in Ridge Regression and Generalized Ridge Regression for Big Data, Statistics; 2015, p. 1–44.
[49] RodneyOD M, Goodman P. Decision tree design using information theory. Knowledge Acqu 1990;2:1–19.
[50] Chourasia S. Survey paper on improved methods of ID3 decision tree. Int J Sci Res Publ 2013;3(12):1–4.
[51] Denison D, Holmes C, et al. Bayesian methods for nonlinear classification and regression. New York: John Wiley and Sons; 2002.
[52] Crammer K, Dekel O, et al. On-line passive-aggressive algorithms. J Machine Learn Res 2006;7:551–85.
[53] Lu J, Zhao P, Steven CH. Online Passive Aggressive Active Learning and its. JMLR: Workshop and Conference Proceedings; 2014, vol. 39, p. 266–82.
[54] Blondel M, Kubo Y, Ueda N. Online passive-aggressive algorithms for non-negative matrix factorization and completion. In: Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics; 2014, PMLR, vol. 33, p. 96–104.
[55] Shah SH, Rashid A, et al. A comparative study of ordinary least squares regression and Theil-Sen regression through simulation in the presence of outliers. Lasbela U. J. Sci. Techl 2016;V:137–42.
[56] Theil. A rank-invariant method of linear and polynomial regression analysis. I, II, III, Nederl. Akad. Wetensch; 1950, vol. 53, pp. 386–392, 521–525, 1397–1412.
[57] Sen P. Estimates of the regression coefficient based on Kendall's Tau. J Am Stat Assoc 1968;63(324):1379–89.
[58] Kar S, Das S, Ghosh P. Applications of neuro fuzzy systems: a brief review and future outline. Appl Soft Comput 2014;15:243–59.
[59] Walia N, Navneet S, Harsukhpreet S. ANFIS: adaptive neuro-fuzzy inference system – a survey. Int J Comp Appl 2015;123:32–8.
[60] McCall J. Genetic algorithms for modelling and optimisation. J Comput Appl Math 2004;184:205–22.
[61] Mitchell M. Genetic algorithms: an overview. Complexity 1995;1(1):31–9.
[62] Blum C. Ant colony optimization: introduction and recent trends. Phys Life Rev 2005;2:353–73.
[63] Sivakumar P, Elakia K. A survey of ant colony optimization. Int J Adv Res Comp Sci Software Eng 2016;6(3):574–8.
[64] Bai Q. Analysis of particle swarm optimization algorithm. Comp Inf Sci 2010;3:180–4.
[65] Singh S. A review on particle swarm optimization algorithm. Int J Sci Eng Res 2014;5(4):551–3.
[66] Dolan W, Cummings P, LeVan M. Process optimization via simulated. AIChE J 1989;35:725–36.
[67] Kirkpatrick S, Gelatt C, Vecchi M. Optimization by simulated annealing. Science, New Series 1983;220:671–80.
[68] Karaboga D, Okdem S. A simple and global optimization algorithm for. Turk J Elec Engin 2004;12:53–60.
[69] Musrrat A, Pant M, Abraham A. Simplex differential evolution. Acta Polytechnica Hungarica 2009;6:95–115.
[70] Kalyanmoy D, Associate A, et al. A fast and elitist multiobjective genetic algorithm. IEEE Trans Evol Comput 2002;6:182–97.
[71] Yusoff Y, Ngadiman M, Mohd Zain A. Overview of NSGA-II for optimizing machining process parameters. Procedia Eng 2011;15:3978–83.
[72] Tiwari A. Machine learning based approaches for. Machine Learn Appl: Int J 2016;3:33–9.
[73] Marras C, Rochom P, Anthony E. Predicting motor decline and disability in Parkinson disease. Arch Neurol 2002;59:1724–8.