Prediction of Parkinson's Disease using Multimodality based Machine Learning Methods

Kushagra Agarwal[1] and Shreeya Pahune[2]

[1,2] International Institute of Information Technology Hyderabad, India

Abstract

Parkinson's disease, a severe neurodegenerative disorder leading to reduced motor skills, does not currently have a reliable test for accurate diagnosis. Early diagnosis of the disease is critical and therapeutic intervention at this stage could significantly reduce the progression of motor and non-motor deterioration. A holistic view of the disease can be obtained by combining several data modalities. Here, we present Multimodal data based supervised machine learning models, which can accurately predict PD using Clinico-Demographic, Biospecimen and Genetic data obtained from the Parkinson Progression Marker Initiative, with a high accuracy of 90% and an AUC-ROC of 0.98. The multimodal approach is compared with single modality based predictions and the superiority of the former can be attributed to its added supplementary information dimension. The various data modalities are analyzed using different techniques to better understand important features and their correlations.

Keywords: Parkinson's Disease prediction; *Multimodal data*; *Machine Learning*;

1 Introduction

Parkinson's disease (PD) is a dopamine-receptor-based neurological condition caused due to the death of dopaminergic neurons in the substantia nigra pars compacta (SNpc). The most common symptom of Parkinson's disease is difficulty in moving and can cause a person to move very slowly as a result of it. Parkinson's disease is a progressive neurological disorder that includes both motor and non-motor symptoms. Aside from a number of common symptoms, each person has their own unique experience with the disease. People with Parkinson disease typically exhibit the following symptoms and signs: bradykinesia (slowness of movement), hypokinesia, rigidity, and rest tremor [8, 39]. Patients' initial clinical symptoms, such as tremor dominant (TD) or postural instability and gait disturbance (PIGD), have been used to categorize them and predict their long-term clinical outcomes. Patients with PIGD have been documented to have faster disease progression and have more non-motor symptoms, but the interpretation of available research is difficult due to the long-term instability of initial clinical classifications [1, 9, 54, 33, 3, 12]. Drugs, as well as less common conditions like multiple cerebral infarction and degenerative conditions like progressive supra nuclear palsy and multiple system atrophy, may also cause Parkinsonism. Although Parkinson's disease is mainly a movement disorder, it also causes other problems such as depression and dementia. Later on, autonomic disruptions and pain may occur,

and the disease may develop to cause severe impairment and handicap, as well as a lower quality of life for the individual affected.

Currently, there is no reliably accurate test that can tell the difference between Parkinson's disease and other disorders with similar manifestations. The diagnosis is largely clinical, based on the patient's medical history and physical examination. The presumed delay in the initiation of motor symptoms that leads to PD diagnosis, as well as the associated neurodegeneration that occurs during this pre-motor phase, represents a missed opportunity for early therapeutic intervention that could dramatically slow or stop the progression of PD-related deterioration [10]. The PD inpatient multidisciplinary treatment concept (Parkinson's disease Multimodal Complex Treatment (PD-MCT)), which was developed in Germany in 2008 [38], includes motor and physical activity as key treatment components. An important positive treatment impact of PD-MCT on motor function in PD patients has been identified, which can be sustained in several parameters over a six-week span and identify predictors for motor function improvement [29].

There is a growing need to provide an early and accurate diagnosis for PD using available data modalities, including but not limited to, imaging data, clinical data, biochemical markers, and genetic biomarkers [45]. It is well known that when applied to numerous data sources, machine learning-based predictive modeling has tremendous potential for clinical prediction problems such as the diagnosis of complex diseases as it can identify trends that a single clinician might miss. However, there often is an unmet need for a larger dataset and additional steps need to be taken to characterize the disease process well [18]. This need for well defined comprehensive datasets can be met by initiatives like the Parkinson Progression Marker Initiative (PPMI) [44], which is a large study aimed at identifying biomarkers for Parkinson's disease progression. The data from the PPMI analysis is a valuable resource for learning more about Parkinson's disease etiology. The PPMI dataset has different data modalities for individual patients and this allows for more generalizable and accurate predictive models [52]. Clinical assessments [43] include Movement Disorders Society-Unified Parkinson Disease Rating Scale (MDS-UPDRS) and Hoehn and Yahr scales [26]. MoCA can be used to assess global cognition [16, 48] and may be a successful screening tool for cognitive function in PD with many benefits [30, 61]. Apart from MoCA other cognitive tests include the Hopkins Verbal Learning Test-Revised (HVLT-R) for memory assessment, Benton Judgment of Line Orientation (BJLO) to assess visuospatial function, Symbol-Digit Modalities Test (SDMT) to assess processing speed-attention, Letter-Number Sequencing (LNS) and semantic (animal) fluency to assess executive abilities-working memory [6, 7, 62, 67, 25]. Neurobehavioral testing include the Geriatric Depression Scale (GDS), State-Trait Anxiety Inventory (STAI), and Questionnaire for Impulsive-Compulsive Disorders (QUIP) [69, 37, 68]. Additional assessments like Epworth Sleepiness Scale and REM sleep behavior disorder (RBD) questionnaire assess sleep behavior, Scales for Outcomes in Parkinson's Disease-Autonomic (SCOPA-AUT) is used to assess autonomic function, and the 40-item University of Pennsylvania Smell Identification Test (UPSIT) examines olfactory function [49, 63, 65, 66].

Biospecimen markers obtained from the cerebrospinal fluid can also be used for diagnosis. There has been a lot of work in understanding the differential presence of such biomarkers in Healthy Controls (HC) and PD patients. The concentrations of α-syn, t-tau, and p-tau, but not Aβ1-42, in the cerebrospinal fluid have been found to be lower in PD than in HC, however there is a substantial overlap between the two groups [32]. Also in

theory, α-syn aggregation in the CNS of PD patients could reduce -syn release into the CSF, similar to the reduction of Aβ1-42 release into the CSF in Alzheimer's disease due to amyloid plaque formation. Basic research and neuropathology studies have indicated that α-syn and tau interact in the brain [19, 24, 28, 70]. The findings of a strong association between CSF α-syn and tau support this theory. However, since the strong association was also observed in HC, any interactions between α-syn and tau proteins are not unique to PD and do not contribute to the pathogenesis of PD [32]. A study of Non-Motor Symptoms in Early Parkinson's Disease discovered clinical and biological variables linked to both baseline burden and progression predictors. There was a connection between a lower baseline Aβ1-42 level and a greater longitudinal increase in NMS [60].

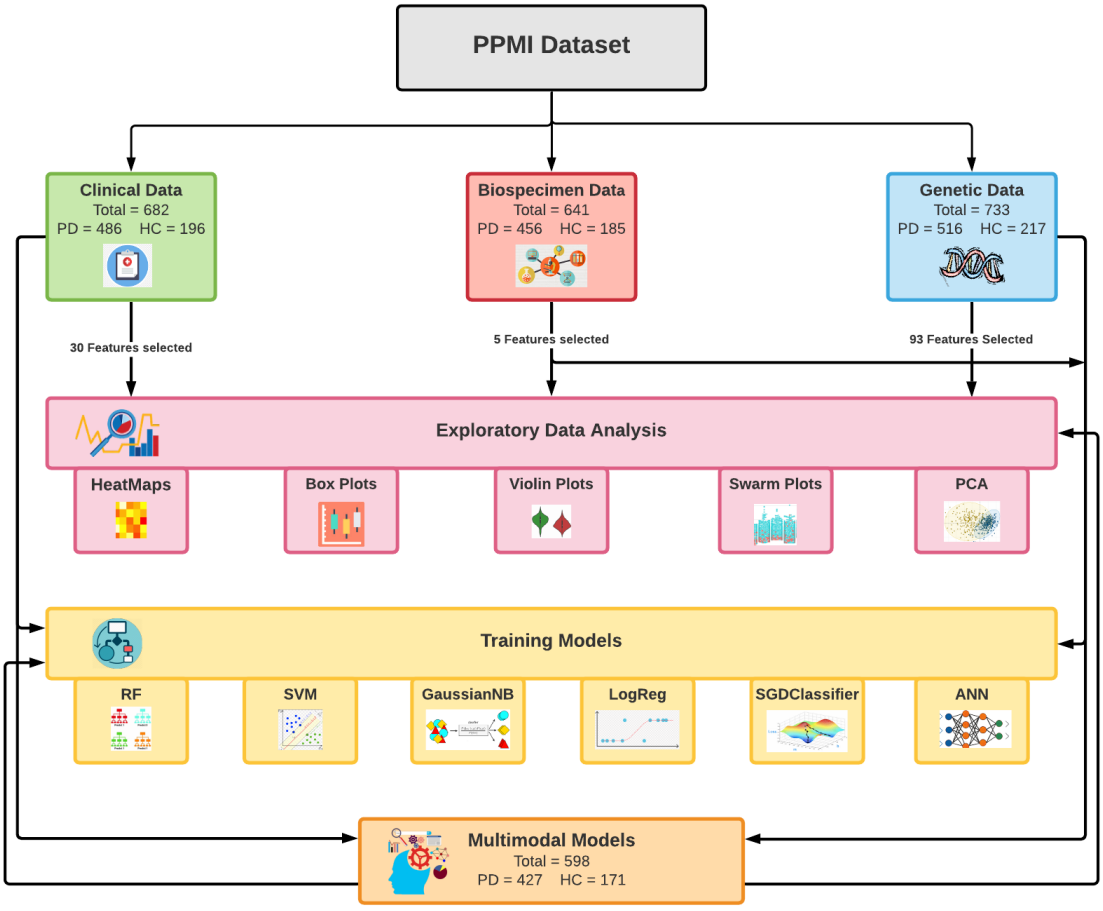

Figure 1: Workflow for the study

Over the last few years, genetic causes resulting in PD have been studied. Most of the common genetic variants associated with PD lie in the PARK loci and related gene regions. PARK loci are regions of the genome where these genes map, and they are PARK1 (SNCA), PARK2 (Parkin), PARK5 (UCHL1), PARK6 (PINK1), PARK7 (DJ-1), PARK8 (LRRK2), PARK9 (ATP13A2), PARK11 (GIGYF2) and PARK13 (OMI/HtrA2) [14, 36]. 27 of the 218 informative genotyped variants reported were associated with PD susceptibility. Of

these eight SNPs were correlated with age at the onset of PD (one SNCA SNP: rs1372520; four PINK1 SNPs: rs3738133, rs1043502, rs1043424, and rs2078073; one Omi/HtrA2 SNP: rs17010022; and two SNCB SNPs: rs1352303 and rs4868670) [14]. Another study [13] reported a different set of 27 statistically significant SNPs associated with susceptibility to PD, lying in the PARK loci and related gene regions. After correcting for multiple testing, only the MAPT SNP rs2435200 was linked to PD susceptibility. However, uncorrected p-values for 16 additional MAPT variants, seven SNCA variants, and one each of LRRK2, PARK2, and UCHL1 variants indicated that they were all relevant. 29 SNPs lying in the SNCA and associated epigenetic landscape, corresponded to high susceptibility, 10 of them to low susceptibility and 4 demonstrated variability in clinical data [59]. An analysis of SNPs lying in the SNCA and MAPT region showed that they are common risk factors for PD [20]. 3 studies were taken into consideration and in the MIHG study, SNCA had the strongest association. SNCA (rs2736990, p-value = 6.7x10-8; genome-wide adjusted p = 0.0109) and MAPT (rs11012, p-value = 5.6x10-8; genome-wide adjusted p = 0.0079) SNPs were found to be genome-wide important. 19 SNPs that were statistically significant (p-value ¡ 1×106) across all 3 studies (MIHG, CIDR and NINDS GWAS data) were reported from the joint analysis [20]. There is a significant association between PD and SNPs in CAST gene (for example, SNP rs1559085, p-value=0.0167) [2] and the sole rs1136666 SNP in GAPDH gene, which is a high risk indicator of PD [53]. From a summarised list of 39 statistically important SNPs in the SNCA region from selected studies, 25 are intronic variants that indicate association to PD. The SNP rs2736990 was reported to be the most frequent and a consistent risk factor for PD. Of all the studies considered, twelve of them found the SNP rs356219 to have a strong association as a susceptibility marker [17]. These studies motivate the importance of genetic biomarkers modality, which could further augment the accurate diagnosis of PD.

The majority of current research has focused on classification or regression results based solely on a baseline time point. The future time point is often overlooked, but the vertical span of time can produce better results [31]. Furthermore, existing approaches for PD diagnosis primarily rely on a single modality function. Different modalities, on the other hand, will have the power to supplement knowledge from various angles [40]. A huge amount of effort has been put into the task of machine-learning based predictive modeling of PD related diagnosis and severity [8, 1, 10, 31, 27, 23, 22, 5, 64, 47, 21, 42, 41, 51, 50, 57, 58]. Deep learning has become a common technique for efficiently analyzing unstructured data such as speech and audio signals, and has thus also found use in PD diagnosis and severity prediction using voice data [27, 23, 22, 5, 64]. A model was developed that allows for the data-driven diagnosis of Parkinson's disease and the differentiation of PD-mimic conditions, such as patients with Parkinsonism but no evidence of dopaminergic dysfunction [47]. The research was then expanded to include the definition of natural disease subtypes as well as an attempt to predict these subtypes along with their progression rates at baseline [21]. A study showed how incorporating multiple data modalities into modeling efforts can boost prediction efficiency. A big advantage of this multi-modal approach is that the various modalities balance each other out, with some modalities being better at predicting case status and others being better at classifying controls. The model developed improved disease risk prediction, which is an important step in better assessing the risk of Parkinson's disease [42]. There are other interesting studies based on prediction of Parkinson's Disease using data collected remotely from smartphones without having the need to go and visit a

clinician for a physical diagnosis [41, 51].

In this paper, the goal is to understand the diagnostic potential of individual data sources, namely, Clinico-demographic, Biospecimen obtained from the cerebrospinal fluid, and Genetic (SNP) data. We hypothesize that the predictive power would further increase when these different sources would be combined using an early integration approach to create a multimodal dataset. Various machine learning models are used for predictions.

2 Methods

2.1 Study Data

Data from the Parkinson Progression Marker Initiative (PPMI) [44] available at `http://www.ppmi-info.org/` was used for this study. Data from three different modalities were used, namely, Clinico-demographic, Biospecimen, and Genetic. The detailed workflow used can be found in Fig 1.

For the clinical modality, 682 participants were analyzed, with 196 control cases and 486 PD patients. The mean age for the participants was 60.9 years with a standard deviation of 10.2. The following features were considered for clinico-demographic analysis, namely, First_Fam_Num, First_Fam_PD, Other_Fam_Num, Other_Fam_PD, AGE, Socio_Score, BJLO, ESS, GDS, HVLT_RECALL, HVLT_RECOG, HVLT_RETENT, LNS, QUIP_SEC_A, QUIP_SEC_B, QUIP_SEC_C, QUIP_SEC_D, QUIP_SEC_E, RBDSQ, SCOPA_TOTAL, SCOPA_GASTRO, SCOPA_URINARY, SCOPA_CARDIO, SCOPA_THERMO, SCOPA_PUPIL, SCOPA_SEXUAL, SFT, STAI, and UPSIT. A detailed description of the various parameters and their computations is listed in Supplementary Table 1. The features MDS_UPDRS1, MDS_UPDRS2, MDS_UPDRS3, Tremor Score, PIGD score, and MoCA scores (Unadjusted and adjusted) were dropped from the analysis owing to the fact that they are indirectly used by the clinician in diagnosing PD, and would lead to circularity in the setting.

Similarly, for the biospecimen modality, 641 participants were analyzed, with 185 HC and 456 PD cases. The features considered were Aβ42 (958.9 $\pm$ 384.5), αSYN (1614.2 $\pm$ 623.4), P_TAU (15.7 $\pm$ 6.6), T_TAU (182.2 $\pm$ 65.0), and URATE (312.3 $\pm$ 73.5) obtained from the cerebrospinal fluid of the participant. The detailed methodology used to extract these features can be found in Supplementary Table 2.

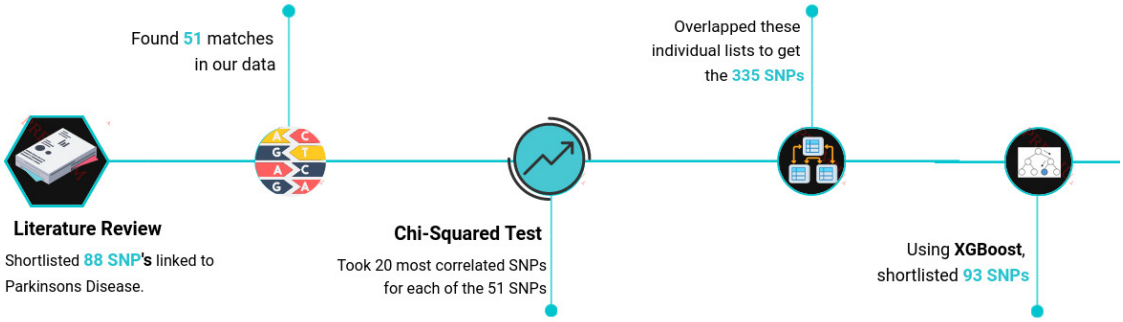

Figure 2: Workflow for Genetic Data Feature Selection

Finally, for the genetic biomarkers modality, we considered 733 participants, with 217 belonging to HC and 516 to PD. Variant Calling Format (VCF) files containing SNPs (hg19) of each participant were analysed to give a total of 6899 unique SNPs in the PPMI data. A biology-based approach was applied to shortlist prognostic SNPs. To do this, a detailed survey of SNPs majorly lying in PD related gene regions was done and 88 prognostic SNPs were identified and annotated to hg19 reference. 51 of these 88 biologically relevant SNPs overlapped with our genetic dataset. Using these 51 SNPs, a correlation analysis was performed with the 6899 and top 20 of the most correlated SNPs (Chi-Squared test, p-value < 0.05), for each prognostic SNP, were selected. This resulted in a set of 335 biologically relevant SNPs, following which XGBoost feature selection method was used to rank the SNPs by feature importance and a final set of 93 prognostic SNPs was obtained. These 93 SNPs, mentioned in Supplementary Table 3, were used for further genetic analysis.

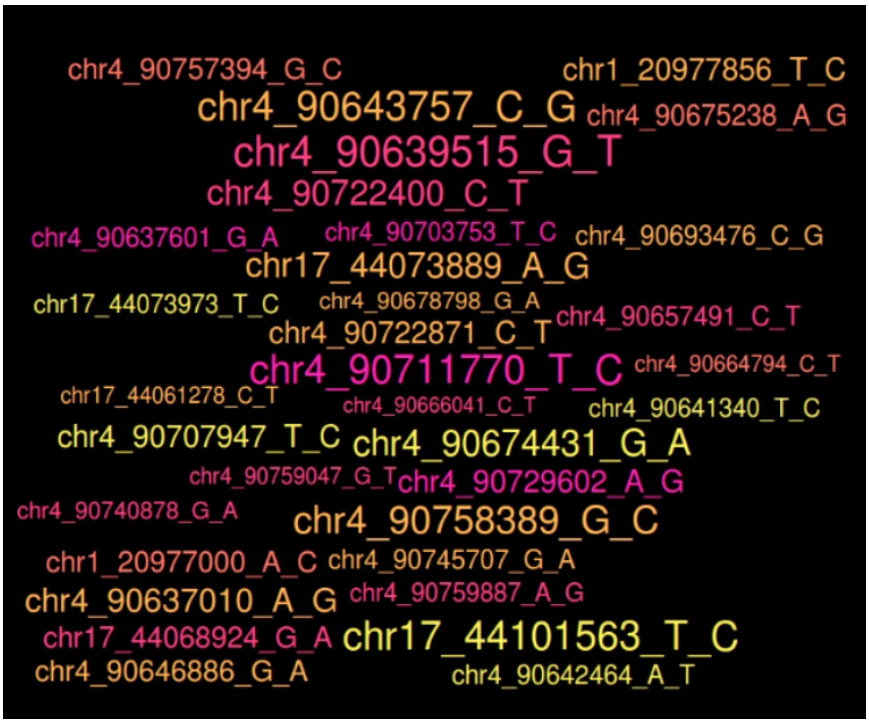

Figure 3: Word Cloud for some of the selected SNPs

2.2 Procedure

The first step was to go over multiple time-points for which a participant's data was collected in each dataset and take an average over them. For each modality, only those participants were considered who did not have missing data for any of the features considered. The entire code for pre-processing of the dataset, including calculation of the various features can be found on the public GitHub repository `https://github.com/SP9144/MutliModal_Prediction_NLD`) For the multi-modal dataset, participants with information for all three individual modalities were considered (n = 598).

Next, to reduce the high dimensionality of the data, we performed Principal Component Analysis (PCA) which is an unsupervised linear transformation technique that is commonly

used for feature extraction and dimensionality reduction. Based on the correlation between features, PCA assists in identifying trends in data and seeks to locate the highest variance directions in high-dimensional data, projecting them onto a new subspace with the same or fewer dimensions as the first. Eigenvalue Decomposition and Singular Value Decomposition(SVD) are used in PCA to minimize dimensionality. The principal components are the eigenvectors that correspond to the largest eigenvalues and are the orthogonal axes of the new (low dimension) subspace, which can be interpreted as the directions of maximum variance. Hence, the original space (with a dimension of the number of features) is reduced to a subspace spanned by a few eigenvectors, with minimal reconstruction error.

After this data analysis, we then built predictive models using supervised machine learning for PD diagnosis using the three data modalities individually and then on the combined multi-modal data using an early integration method. For each of the data types, we trained Random Forest classifiers, Gaussian Naive Bayes, Support Vector Machines (SVM), Logistic Regression, Stochastic Gradient Descent classifier (SGD), and Artificial Neural Networks (ANN). To get accurate and reliable predictions, Random Forest was used. They create several decision trees and merge them together. Each tree in the random forest produces a class prediction, and the class with the most votes becomes the prediction of the model [11]. A probabilistic classifier, Naive Bayes, was also used, which is based on Bayes' theorem and assumes a strong independence between features. By assuming a Gaussian distribution, Naive Bayes can be applied to real-valued (continuous) attributes. Gaussian Naive Bayes is an extension of naive Bayes [55]. To ensure classes are differentiated, Support vector machines or SVMs were used. They work by maximising the margin that is maximising the distance between the classes.The SVM algorithm's goal is to find a hyperplane in the original space (with a dimension equal to the number of features) that categorizes the data points clearly [15]. Logistic regression statistically models the likelihood of the train data point belonging to one of the many output classes using a sigmoid activation function [46]. SGD Classifier is a linear classifier optimized by Stochastic Gradient Descent which uses only a subset of training samples for each epoch, and therefore is faster than the general Gradient Descent (GD) [34]. Artificial Neural Networks (ANNs) structure comprises an input, one or more hidden, and an output layer. Input layer nodes pass information to hidden layer nodes by firing activation functions (like sigmoid or ReLU) and this process continues till the output layer. One or more nodes of the output layer are activated and these indicate the class of the train data point. For all models, except for ANN, we performed an 80:20, train-test split and ensured that the class distribution in both sets remains similar. To train the different models, we performed a Randomized Search over the tunable hyperparameters and performed a 5-fold Cross-Validation to further validate our results. Randomised search was used for optimizing hyperparameters of a given model. Although this search explores the same parameter space, it does not explore all parameter combinations, a certain number of randomly picked combinations are tried. This allows for a reliable estimate of parameters in a shorter span of time. As the Randomized Search may not give the most optimal set of parameters as Grid Search, each model was trained multiple times and the best performing models were saved for further downstream analysis. In k-fold Cross Validation (CV), the training set is split into k sets. For each k (k=5) set of subsamples (fold), the model is trained using k-1 of the folds as training data, and the remaining data is used to validate the resulting model. The average of the values computed in the loop becomes the output metric stated by k-fold cross-validation. The model with

the highest values of 'AUC-ROC' was chosen to be the best performing one, as accuracy is not a very good predictor of model performance in a class-imbalanced dataset like ours. For the ANN, however, the data was split into 60:20:20 for train, validation and test sets respectively and equal class distribution in the three sets was ensured. The ANN was trained using the Keras library, which is a high-level API built on top of Tensorflow. Activation functions used included sigmoid and ReLU activation. Hyperparameters were tuned manually using trial and error, with a binary cross-entropy loss function optimized using the Adam optimizer. Adaptive Moment Estimation (Adam) is an optimization algorithm that iteratively updates network weights based on training data. It's a stochastic gradient descent approach based on adaptive first-order and second-order moment estimation [35]. For all the different data modalities including the early integrated multi-modal data, all of these steps were repeated and final results were obtained.

All code relating to the trained models, including the models and their weights, is also present in the github repository under the train_code folder.

3 Results

3.1 Exploratory Data Visualizations

3.1.1 Clinical Data

As mentioned above, clinical data for 682 participants, with 196 control cases and 486 PD patients was obtained from PPMI. The dataset was first analyzed for correlations and the Heatmap thus obtained can be seen in Fig 4. It can be clearly observed, that features like UPSIT, SCOPA, RBDSQ and STAI seem to be more correlated with the final diagnosis (STATUS).

The Box plot, Violin plot and the Swarm plot for the clinical modality can be seen in Fig 5, 6, and 7. From the Violin plot, with Kernel Density Estimation which is a non-parametric way to estimate the probability distribution function of a random variable, we can conclude the predictive power of a feature using the differential distribution functions for the two classes (Green for HC, and Orange for PD). The feature First_Fam_PD which is the number of PD affected people in the first family of the participant, presents a negligible spike in the HC class, whereas a distinct spike can be seen for the PD class. This is clearly indicative of the fact that a person with already affected members in his first family is more susceptible to PD. The peaks for RBDSQ feature also occur differently for HC and PD, with the peak for PD occuring at a higher value, which can also be confirmed with the Swarm plot, with most of the HC class participants (Red), having a lower value of RBDSQ than the PD class (aqua). Other features with differential values in the two classes are GDS (lesser in HC), SCOPA_TOTAL (lesser in HC), SCOPA_GASTRO (lesser in HC), SCOPA_CARDIO (lesser in HC), STAI (lesser in HC) and UPSIT (greater in HC). All of this can be confirmed with both the Violin and the Swarm plots which therefore act as very powerful visualization tools.

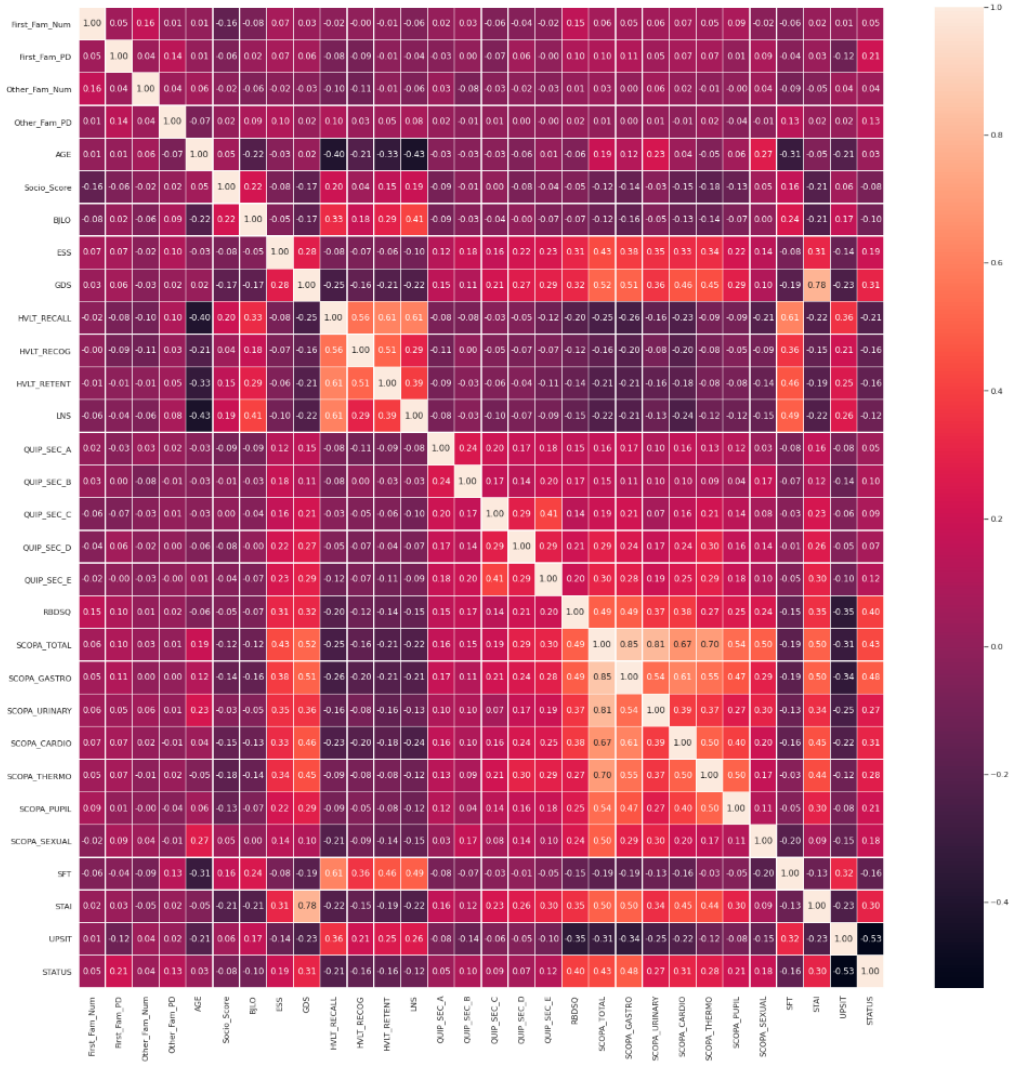

Figure 4: Heatmap for Clinical Data

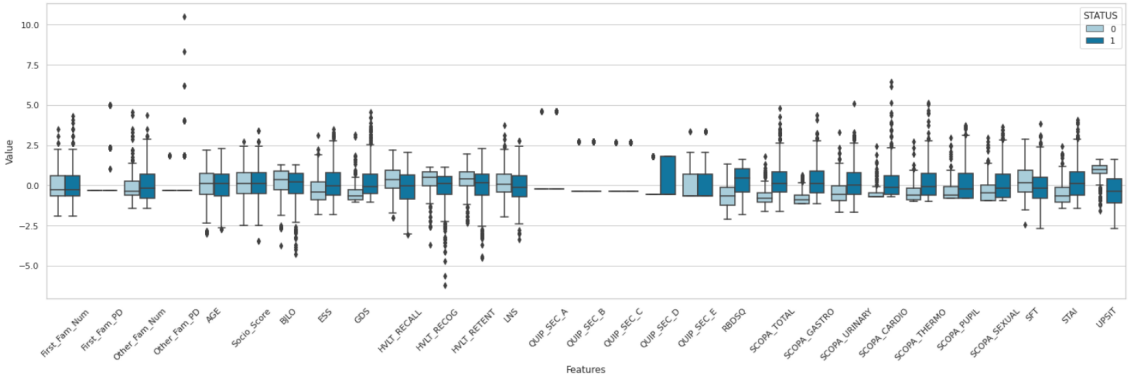

Figure 5: Box plot for Clinical Data

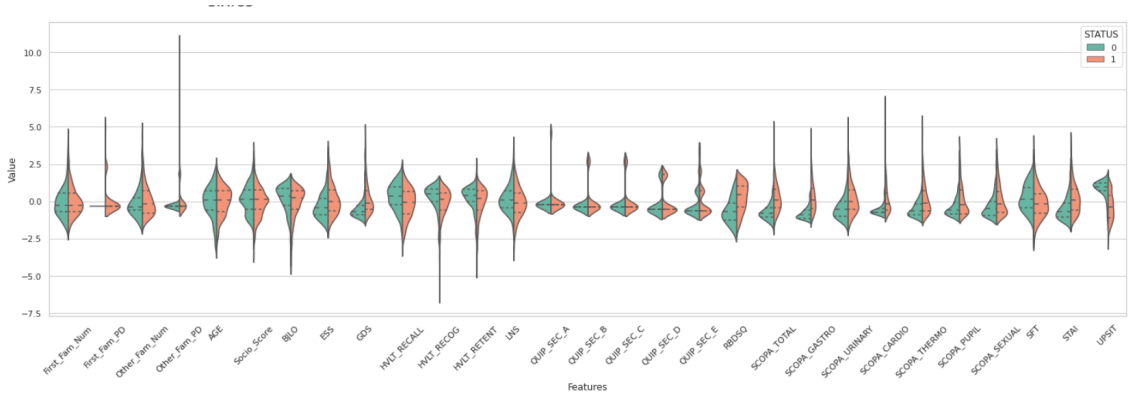

Figure 6: Violin plot for Clinical Data

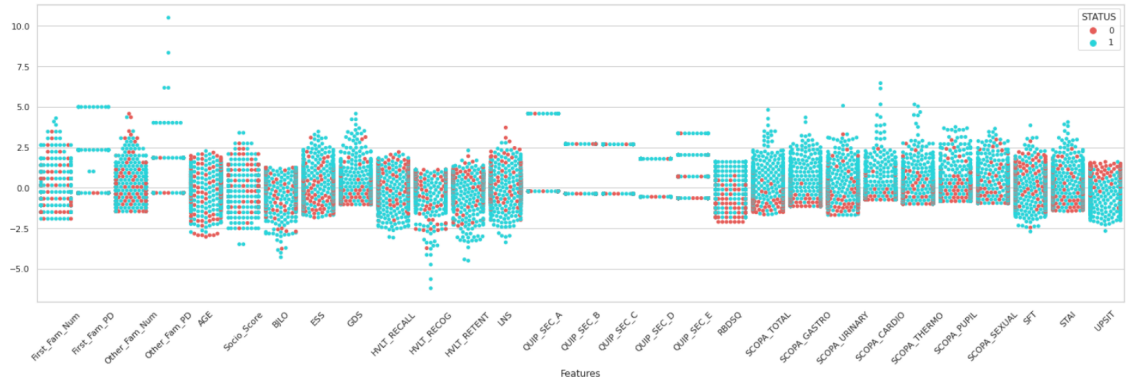

Figure 7: Swarm plot for Clinical Data

Next we perfomed dimensionality reduction on the data using Principal Component Analysis. 41.4% of the variance in the data was explained by PC1, and 23% by PC2. On sorting the loading scores in descending order of magnitude for the features contributing the most to PC1, STAI, SCOPA_TOTAL, UPSIT, SFT, HVLT_RECALL GDS, and RBDSQ were returned. The Scree and PCA plots, coloured by classes, can be seen in Fig 8.

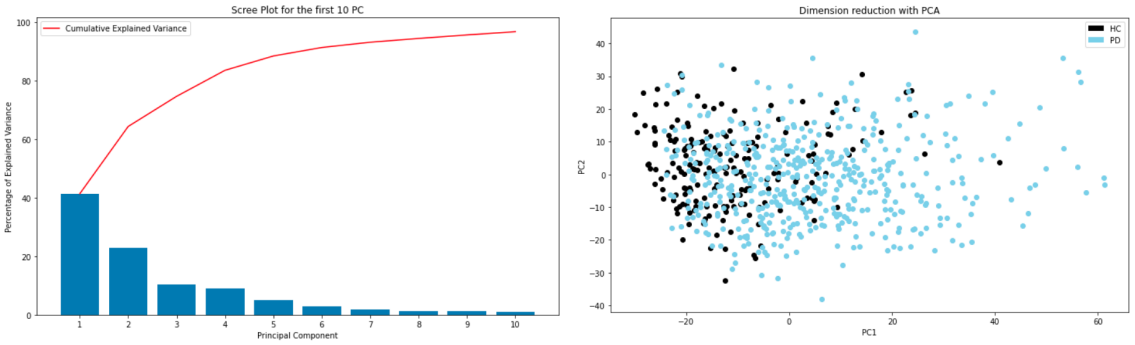

Figure 8: Principal Component Analysis on Clinical Data

3.1.2 Biospecimen Data

The Heatmap for correlations amongst the features in Biospecimen modality, namely, ABETA42, A_SYN, P_TAU, T_TAU, URATE, and the diagnosis (STATUS) is shown in Fig 9. It is clear that none of the other input features correlate with URATE. On the other hand a strong correlation can be seen for the features P_TAU and T_TAU. The correlation for A_SYN with P_TAU and T_TAU is also significant. None of the features seem to have a strong correlation with the diagnosis. This is also supported by other visualizations, like the similar probability distribution functions in the Violin Plot and a random distribution of HC and PD participants in the Swarm Plot.

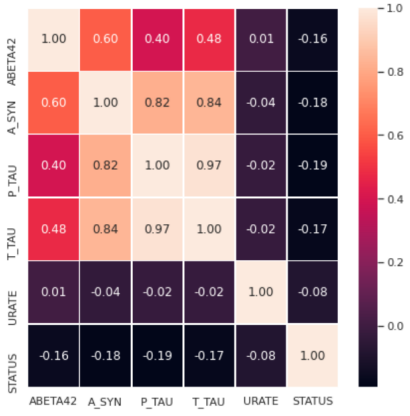

Figure 9: Heatmap

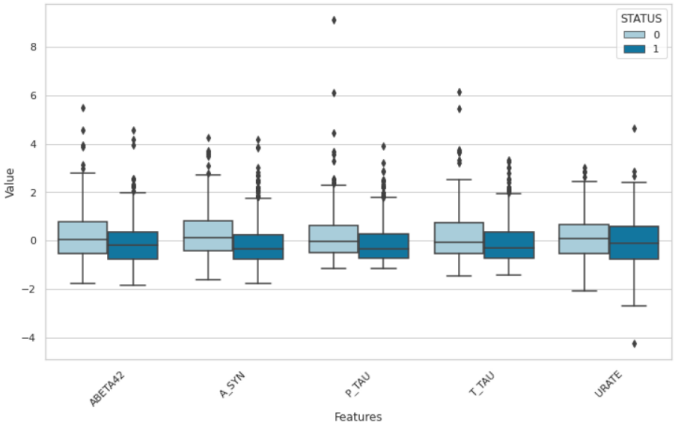

Figure 10: Box plot for Biospecimen Markers

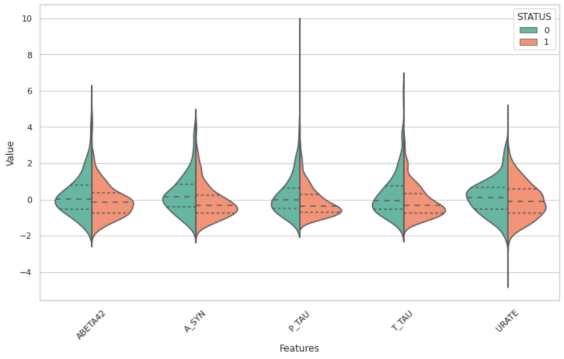

Figure 11: Violin Plot

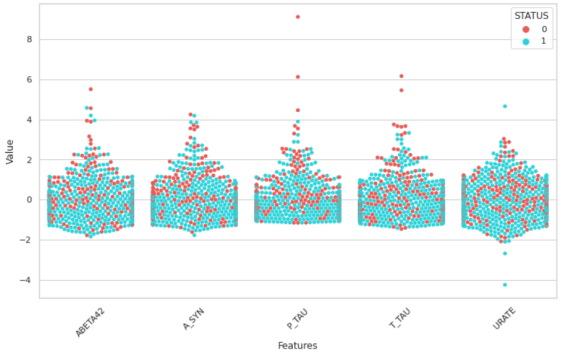

Figure 12: Swarm plot

We next plotted detailed correlation plots for the pairs of features that were found to be highly correlated in Fig 13. The plots for P_TAU and T_TAU, A_SYN and P_TAU, A_SYN and T_TAU and, P_TAU and URATE clearly show that the former 3 have high correlation (closely mapping to the regression line), whereas the last pair has a very poor correlation.

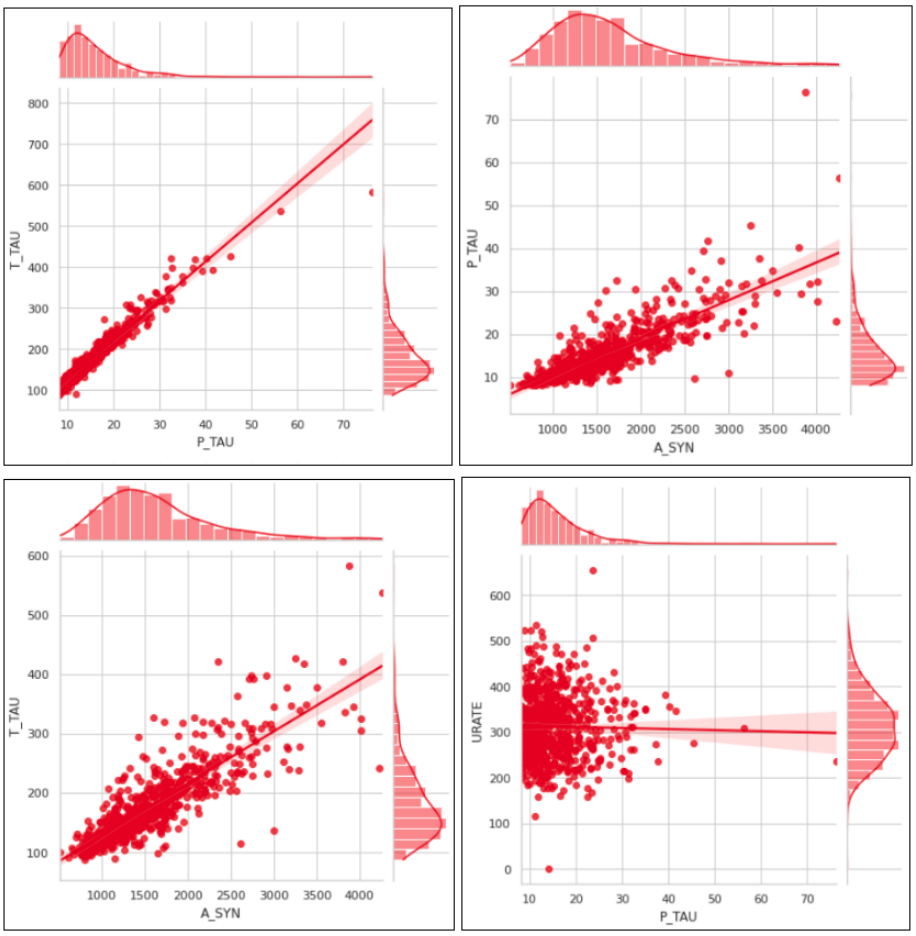

Figure 13: Correlation plots for a) P_TAU and T_TAU b) A_SYN and P_TAU c) A_SYN and T_TAU d) P_TAU and URATE in the Biospecimen Markers

Principal Component Analysis performed on these 5 features showed an expected high amount of variance getting captured in the first principal component (84.1%) and (14.7%) in the second principal component. URATE as expected had the smallest loading score by magnitude in the first principal component.

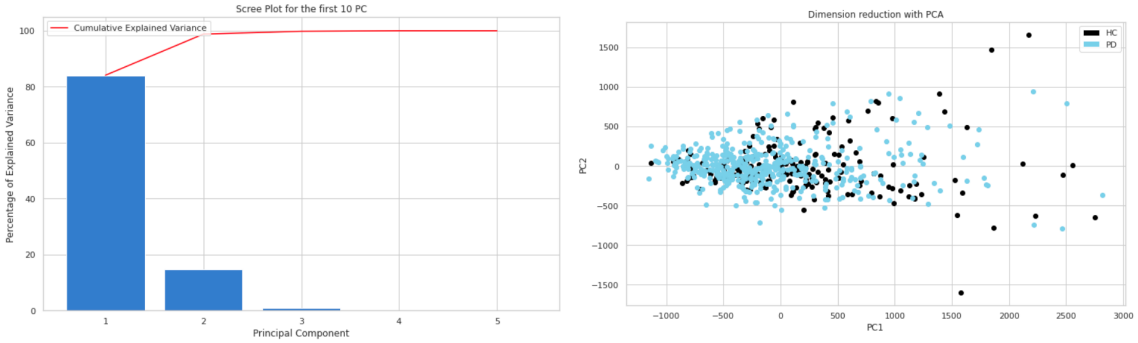

Figure 14: Principal Component Analysis on Biospecimen Markers

3.1.3 Genetic Data

Genetic data related to the presence/absence of 93 SNPs was collected from 733 patients (217 HC, 516 PD). As the data was categorical in nature, and was already selected using XGBoost feature selection method, we did not perform other visual analysis for it. PCA, however, was conducted on the dataset, and the results are shown in Fig 15. chr4_90753960_C_T, chr4_90741519_G_A, chr4_90755939_A_G, chr4_90759047_G_T, and chr4_90763260_A_G among others had the highest loading score for the first principal component. Only 21.4% and 13.0% variance could be captured by the first two principal components.

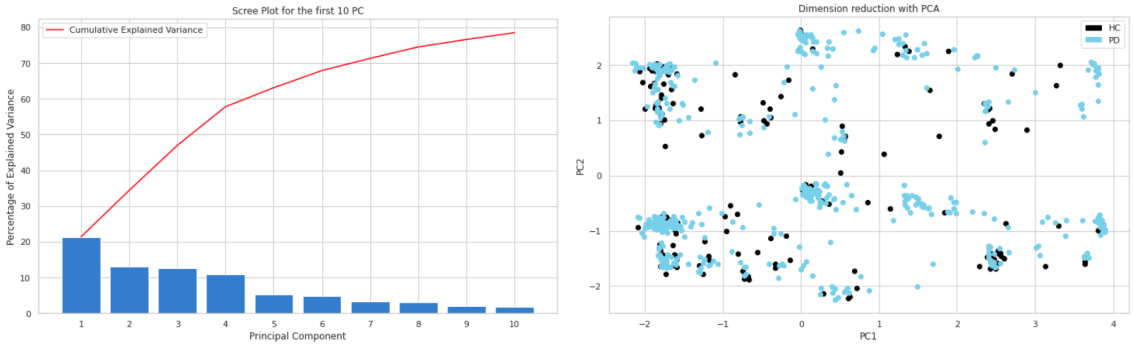

Figure 15: Principal Component Analysis on Genetic SNP data

3.1.4 Multimodal Data

Data from the three modalities was combined using an early integration method, wherein, all those participants data was dropped which had incomplete information for any of the three. After this integration, 598 participants remained, with 171 belonging to the control case and 427 belonging to the diseased case. Principal Component Analysis on this combined dataset, did not reveal a very clear separation for the two classes as the features contribution the most to the largest eigenvalue were all related to Biospecimen data, which in itself isn't a very good diagnostic predictor.

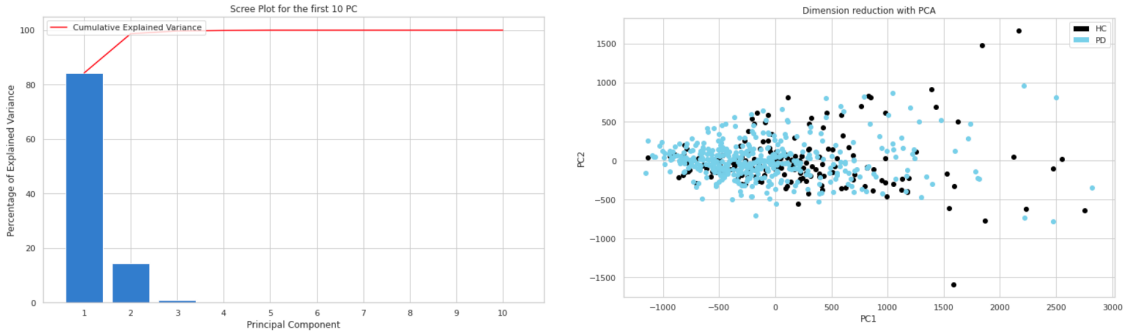

Figure 16: Principal Component Analysis on the combined Multimodal data

3.2 Model Training and Evaluation

For all the data modalities, multiple supervised machine learning models were trained, namely, Random Forests (RF), Gaussian Naive Bayes (GNB), Support Vector Machines (SVM), Logistic Regression (LR), Stochastic Gradient Descent classifier (SGD) and Artificial Neural Networks (ANN). The code relating to the training along with model parameters is available in the github folder. Details about model parameters have also been mentioned in the Supplementary Table 4.

The data was split 80:20 into train-test, and for all the other models except for the ANN, a 5-fold cross validation on this train set was performed. In the ANN, the train set was further divided into 60:20 for train and validation sets respectively. As all the datasets were class-imbalanced, class-weights were used. All the training was performed with a GPU accelerator on Google Colab and the notebooks have been made publicly available along with the datasets through the github repository. Randomized search, which was used for all the other models except for ANN, is much faster compared to a Grid Search, but is known to not give optimal results in a single run. Owing to this, all the models were retrained multiple times and only the best performing models (based on AUC-ROC), were saved.

For Random Forests, the hyperparameters, which were tuned through Randomized Search included: Bootstrapping (True, False), Criterion (Gini, Entropy), Max Depth, Max features, Minimum samples in the leaf, Minimum samples split, Number of estimators, Warm start (True, False), and Class weights. For Gaussian Naive Bayes, the hyperparameter was Variable Smoothing. Support Vector Machines were hyperparameter tuned for Kernels (Linear, Polynomial, RBF and Sigmoid), Gamma (Scale, Auto), Shrinking (True, False), Class weights, Decision Function Shape and the penalty trade off parameter C. Logistic Regression had the following hyperparameters: C, Penalty (L1, L2, Elasticnet), Solver (Newton-cg, Lbfgs, Liblinear, Sag, Saga), Class weights, Multi class (Multinomial, OVR, Auto), Warm Start (True, False), and L1 Ratio. For Stochastic Gradient Classifiers, Alpha, L1 Ratio, Loss and Class weights were tuned. ANN however was manually hyperparameter tuned by trial and error and choosing the model best performing on the validation set. The number of hidden layers, and number of neurons in the layers were changed and the training was performed for different batch sizes and different number of epochs. The ReLU (Rectified Linear Unit) activation function was used for the hidden layers, and a sigmoid activation function was used for the output layer. Adam optimizer was used to optimize over a binary cross-entropy loss function.

3.2.1 Clinical Data

Best performing models for the different models along with their ROC curves is shown. Different metrics are computed to compare between the models.

	RF	GNB	SVM	LR	SGD	ANN
Precision	0.93	0.90	**0.94**	0.92	0.93	0.88
Recall	0.85	0.71	0.91	0.88	0.91	**0.93**
F1 score	0.89	0.80	**0.92**	0.90	**0.92**	0.90
Accuracy	0.85	0.74	**0.89**	0.86	0.88	0.86
AUC-ROC	0.91	0.86	**0.95**	**0.95**	**0.95**	0.93

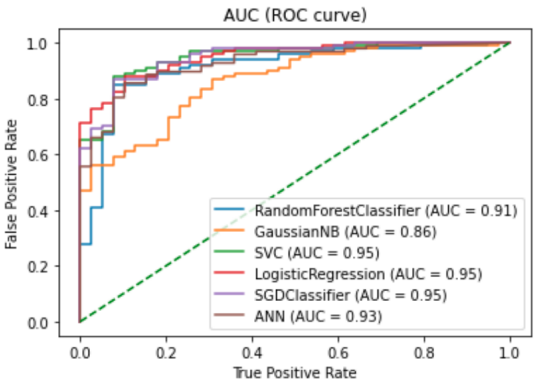

Figure 17: Model evaluation for Clinical Data

All the three SVM, LR and SGD give the highest AUC-ROC of 0.95. However, SVM also gives the highest precision (0.94), high recall (0.91), highest F1 score (0.92) and the highest Accuracy (89%), thereby making it the best performing model for clinical data. The details about the best performing model parameters can be found in Supplementary Table 4. On performing XGBoost based feature importance analysis, SCOPA_TOTAL, UPSIT, SCOPA_GASTRO, SCOPA_CARDIO and First_Fam_PD were found to be the most important features for diagnosis.

3.2.2 Biospecimen Data

	RF	GNB	SVM	LR	SGD	ANN
Precision	0.80	0.76	**0.84**	0.81	0.81	0.80
Recall	0.79	**0.99**	0.61	0.63	0.62	0.63
F1 score	0.79	**0.86**	0.70	0.71	0.70	0.70
Accuracy	0.70	**0.77**	0.63	0.62	0.61	0.61
AUC-ROC	0.65	**0.74**	0.70	0.67	0.69	0.62

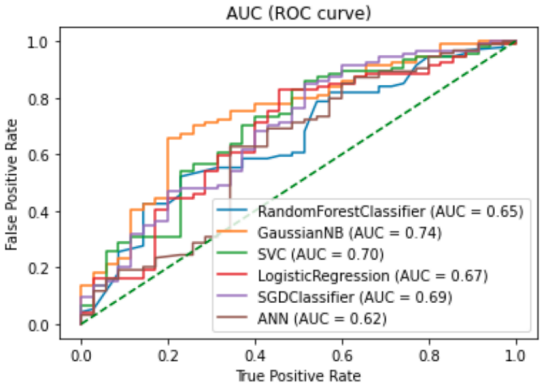

Figure 18: Model evaluation for Biospecimen Markers

None of the models are able to accurately diagnose PD using biospecimen data, however, Gaussian Naive Bayes, has a a very high recall of 0.99, while having a decent accuracy (77%) and AUC-ROC (0.74). Clearly, biospecimen data alone is unable to diagnose PD accurately, but the high recall for GNB suggests that it is still reducing the chances of misdiagnosis for actual PD patients, i.e, the number of false negatives is very low. A_SYN was found to be the most important feature using XGBoost, followed by the derived variables, P_TAU/T_TAU, P_TAU/ABETA42, and T_TAU/ABETA42.

3.2.3 Genetic Data

	RF	GNB	SVM	LR	SGD	ANN
Precision	**0.78**	0.78	0.82	0.79	0.82	0.76
Recall	0.38	0.64	0.59	**0.65**	0.55	0.54
F1 score	0.51	0.70	0.69	**0.72**	0.66	0.63
Accuracy	0.48	0.61	0.61	**0.63**	0.59	0.55
AUC-ROC	0.62	0.60	**0.65**	**0.65**	**0.65**	0.63

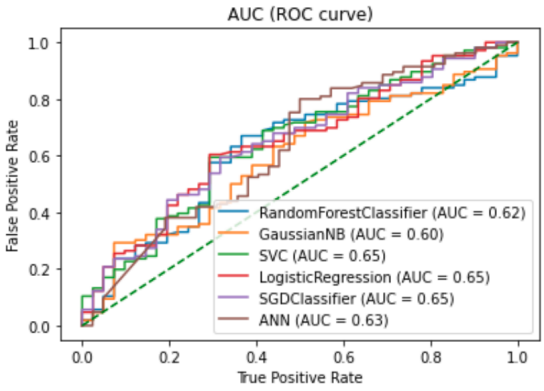

Figure 19: Model evaluation for Genetic Markers

Again, none of the models are able to diagnose PD using genetic data alone. Logistic Regression, is performing the best on this modality, with an AUC-ROC of (0.65), and an accuracy of (63%). Clearly, there is a definite need to augment the predictive power of this modality with other modalities. Feature importance analysis revealed chr4_90678541_G_A, chr4_90697979_T_C, chr4_90753960_C_T, chr4_90711770_T_C, and chr4_90741773_A_G to be the SNPs with highest predictive potential.

3.2.4 Early Integrated Multimodal Data

	RF	GNB	SVM	LR	SGD	ANN
Precision	0.95	0.95	**0.99**	0.97	0.95	0.94
Recall	**0.90**	0.70	0.87	0.70	0.87	0.85
F1 score	0.92	0.81	**0.93**	0.81	0.91	0.89
Accuracy	0.89	0.76	**0.90**	0.77	0.88	0.85
AUC-ROC	0.94	0.90	**0.98**	0.95	0.93	0.93

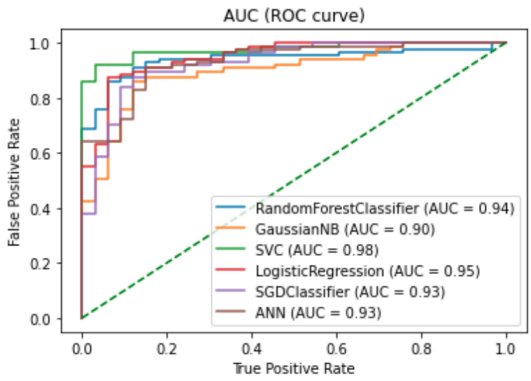

Figure 20: Model evaluation for Multimodal Data

It can be very clearly seen that SVM is performing the best on this dataset, with an accuracy of 90%, AUC-ROC of 0.98, Precision 0.99, Recall 0.87 and F1 score of 0.93. On doing a feature importance analysis using the XGBoost machine learning algorithm, UPSIT, SCOPA_GASTRO, chr4_90755939_A_G, chr4_90646886_G_A, and SCOPA_TOTAL were found to be the most important features for accurate diagnosis.

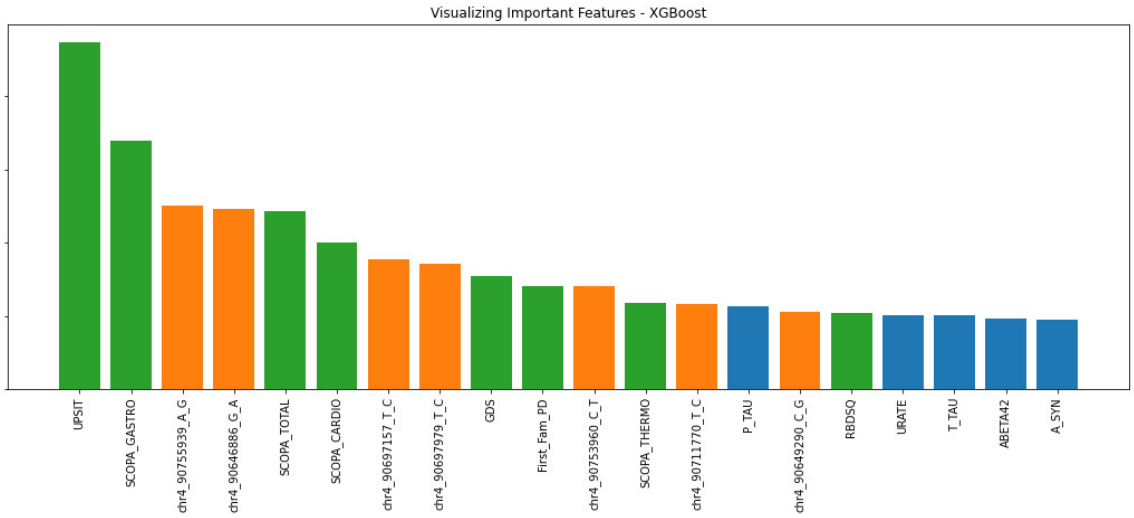

Figure 21: XGBoost Feature Importance for Multimodal data

The most interesting observation here is that, as expected, the predictive power of the models increased on using a combination of the datasets. Clinical data which alone had a maximum AUC-ROC of 0.95, now had both an increased accuracy and an increased AUC-ROC on integration with lesser predictive Biospecimen and Genetic datasets.

	Model	Precision	Recall	F1 Score	Accuracy	AUC-ROC
Clinical	SVM	0.94	0.91	0.92	0.89	0.95
Biospecimen	GNB	0.76	**0.99**	0.86	0.77	0.74
Genetic	LR	0.79	0.65	0.72	0.63	0.65
Multimodal	SVM	**0.99**	0.87	**0.93**	**0.90**	**0.98**

Figure 22: Performance comparison of best models for different data modalities

4 Discussion

In the care, treatment, and study of complex heterogeneous diseases, predicting disease and disease course is a major challenge. However, in the current age of big data, getting hetereogenous sources of data is possible. It is therefore, imperative that machine learning based predictive models step up and incorporate these diverse datasets to augment their diagnostic potential. To help in the diagnosis of PD, we make use of a multimodal data-based machine learning model.

We created a very comprehensive dataset of features extracted from the PPMI study including a wide-list of parameters for the Clinico-Demographic data. To create the genetic dataset, we applied biology-inspired feature selection methods using machine learning, to not just select the most correlated features amongst the 6899 avaliable SNPs, but to also

incorporate "causation" into our model. The SNPs were obtained after a detailed literature review of available information followed by a correlation analysis using the XGBoost machine learning algorithm. Exploratory Data analysis was performed on all the data modalities, using various visualization tools, namely, Correlation Heat maps, Box plots, Violin plots, and Swarm plots. Unsupervised Learning Approach of Principal Component Analysis was also applied to better understand the feature space. Results from the Principal Component Analysis suggested that SCOPA_TOTAL, UPSIT, STAI are important in dimensionality reduction for Clinical data, and certain SNPs in the chromosome 4 for Genetic data.

We then performed rigorous training on these individual data modalities using various supervised machine learning methods (RF, GNB, SVM, LR, SGD and ANNs). The clinical data was best modelled using Support Vector Machines with the highest AUC-ROC of 0.95 and an accuracy of 90%. SCOPA_TOTAL, UPSIT and SCOPA_GASTRO again appeared to be the most important features for predictive modelling on this dataset. Biospecimen and Genetic data alone, were not able to provide an accurate diagnosis of PD, and even their best performing models GNB for Biospecimen (AUC-ROC: 0.74, Accuracy: 77%) and LR for Genetic (AUC-ROC: 0.65, Accuracy: 63%) were below par and not very useful. However, on combining these different modalities using an early integration apporach, an integrated Multimodal datset was created. SVM, which performed the best on this dataset, was able to achieve an AUC-ROC of 0.98 and an accuracy of 90% which was greater than all the individual models. An XGBoost machine learning algorithm based feature importance analysis revealed that UPSIT, SCOPA_GASTRO, chr4_90755939_A_G, chr4_90646886_G_A, and SCOPA_TOTAL were the most important features for accurate diagnosis.

It is interesting to note that while Genetic data alone was a very poor predictor of PD, on combination with Biospecimen and Clinical data modalities, its predictive power increased, so much so, that 2 of the SNPs even appear in the top 5 important features for diagnosis in the multimodal dataset. It is also important to appreciate that the while Biospecimen and Genetic data themselves fail to model the diagnosis well, they still augment the Clinical models with additional value by supplementing more information and thereby compensating for erroneous predictions.

Machine-learning methods to predict PD have been developed in the past and according to a survey on these ML methods for PD prediction [8], most of them involve usage of speech signals and voice recording data, with accuracies in the range of 85% to 95%. However, one study included in the survey used non-motor symptoms to discriminate PD from healthy controls and reported an accuracy of 72% to 92% [4]. Some studies that made use of MRI data for the same purpose reported an average accuracy of 70% and one study had an accuracy greater than 90%. Faghri et al. [21] proposed using RFE for feature selection followed by Logistic Regression to identify patient subtypes and to predict PD progression based on baseline and year 1 data and reported an average AUC of 0.956. They used the data of only 328 patients from the PPMI cohort and included MDS-UPDRS as a feature, along with other clinical and biospecimen features, this was avoided by us to prevent circularity. A recent study [42] made use of data from PPMI, for training and validation, and PDBP [56], for testing, to predict PD. They trained individual models for genetics, clinico-demographics and transcriptomics data along with a model for combined data and reported maximum AUC for validation as 0.69, 0.89, 0.78 and 0.87 respectively. Performance on the combined model tested on PDBP test set gave an AUC of 0.85 and an

accuracy of 75%. Our models gave better results in comparison to these past studies, the model trained on combined data had an AUC of 0.98 and a high accuracy of 90%. The accuracy of our approach can be further improved by augmenting with other modalities like speech data of patients among other attributes.

5 Conclusion

In this paper, we propose a multimodal machine learning model, which can help in the diagnosis of Parkinson's Disease. Our proposed model produces more accurate results as compared to existing methods. It was observed that the model trained on multimodal data shows better performance than those trained on single data modalities and highlights unexpected features that could be prognostic biomarkers of Parkinson's Disease. The multimodality adds to the robustness as well as generalizability of the model with the various data sources compensating and supplementing each other, giving us a more holistic view. This paper tries to provide an insight into the broader application of machine learning models on multimodal data sources to improve disease diagnosis in the domain of healthcare. For diseases like Cancer, this becomes even more imperative considering its complex heterogenous presentations amongst different patients. Machine learning approaches have the power of capturing complex relationships between features, and an integrated heterogenous data source could further utilize this potential to build even more accurate predictive models.

References

[1] D. Aleksovski, D. Miljkovic, D. Bravi, and A. Antonini. Disease progression in parkinson subtypes: the ppmi dataset. *Neurological Sciences*, 39, 11 2018.

[2] A. S. Allen and G. A. Satten. SNPs in CAST are associated with parkinson disease: A confirmation study. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 9999B:n/a–n/a, 2010.

[3] G. Alves, J. P. Larsen, M. Emre, T. Wentzel-Larsen, and D. Aarsland. Changes in motor subtype and risk for incident dementia in parkinson's disease. *Movement Disorders*, 21(8):1123–1130, 2006.

[4] R. Armañanzas, C. Bielza, K. R. Chaudhuri, P. Martinez-Martin, and P. Larrañaga. Unveiling relevant non-motor parkinson's disease severity symptoms using a machine learning approach. *Artificial Intelligence in Medicine*, 58(3):195–202, July 2013.

[5] M. Asgari and I. Shafran. Extracting cues from speech for predicting severity of parkinson's disease. In *2010 IEEE International Workshop on Machine Learning for Signal Processing*, pages 462–467. IEEE, 2010.

[6] R. H. Benedict, D. Schretlen, L. Groninger, and J. Brandt. Hopkins verbal learning test–revised: Normative data and analysis of inter-form and test-retest reliability. *The Clinical Neuropsychologist*, 12(1):43–55, 1998.

[7] A. L. Benton, N. R. Varney, and K. d. Hamsher. Visuospatial judgment: A clinical test. *Archives of neurology*, 35(6):364–367, 1978.

[8] S. Bind, A. K. Tiwari, and A. K. Sahani. A survey of machine learning based approaches for parkinson disease prediction, 2015.

[9] R. Biundo, L. Weis, and A. Antonini. Cognitive decline in parkinson's disease: the complex picture. *npj Parkinson's Disease*, 2(1), Sept. 2016.

[10] F. D. Bowman, D. F. Drake, and D. E. Huddleston. Multimodal imaging signatures of parkinson's disease. *Frontiers in Neuroscience*, 10, Apr. 2016.

[11] L. Breiman. *Machine Learning*, 45(1):5–32, 2001.

[12] D. J. Burn. Motor subtype and cognitive decline in parkinson's disease, parkinson's disease with dementia, and dementia with lewy bodies. *Journal of Neurology, Neurosurgery & Psychiatry*, 77(5):585–589, May 2006.

[13] S. J. Chung. Human genetic variation and parkinson's disease. *Journal of Movement Disorders*, 3(1):1–5, Apr. 2010.

[14] S. J. Chung, S. M. Armasu, J. M. Biernacka, T. G. Lesnick, D. N. Rider, S. J. Lincoln, A. I. Ortolaza, M. J. Farrer, J. M. Cunningham, W. A. Rocca, and D. M. Maraganore. Common variants in PARK loci and related genes and parkinson's disease. *Movement Disorders*, 26(2):280–288, Dec. 2010.

[15] N. Cristianini and E. Ricci. Support vector machines. In *Encyclopedia of Algorithms*, pages 928–932. Springer US, 2008.

[16] J. Dalrymple-Alford, M. MacAskill, C. Nakas, L. Livingston, C. Graham, G. Crucian, T. Melzer, J. Kirwan, R. Keenan, S. Wells, et al. The moca: well-suited screen for cognitive impairment in parkinson disease. *Neurology*, 75(19):1717–1725, 2010.

[17] C. L. das Chagas Campêlo and R. H. Silva. Genetic variants in SNCA and the risk of sporadic parkinson's disease and clinical outcomes: A review. *Parkinson's Disease*, 2017:1–11, 2017.

[18] R. C. Deo. Machine learning in medicine. *Circulation*, 132(20):1920–1930, 2015.

[19] T. Duka, V. Duka, J. N. Joyce, and A. Sidhu. α-synuclein contributes to gsk-3β-catalyzed tau phosphorylation in parkinson's disease models. *The FASEB Journal*, 23(9):2820–2830, 2009.

[20] T. L. Edwards, W. K. Scott, C. Almonte, A. Burt, E. H. Powell, G. W. Beecham, L. Wang, S. Züchner, I. Konidari, G. Wang, C. Singer, F. Nahab, B. Scott, J. M. Stajich, M. Pericak-Vance, J. Haines, J. M. Vance, and E. R. Martin. Genome-wide association study confirms SNPs inSNCAand theMAPTRegion as common risk factors for parkinson disease. *Annals of Human Genetics*, 74(2):97–109, Mar. 2010.

[21] F. Faghri, S. H. Hashemi, H. Leonard, S. W. Scholz, R. H. Campbell, M. A. Nalls, and A. B. Singleton. Predicting onset, progression, and clinical subtypes of parkinson disease using machine learning. *bioRxiv*, page 338913, 2018.

[22] Z. Galaz, Z. Mzourek, J. Mekyska, Z. Smekal, T. Kiska, I. Rektorova, J. R. Orozco-Arroyave, and K. Daoudi. Degree of parkinson's disease severity estimation based on speech signal processing. In *2016 39th International Conference on Telecommunications and Signal Processing (TSP)*, pages 503–506. IEEE, 2016.

[23] N. Genain, M. Huberth, and R. Vidyashankar. Predicting parkinson's disease severity from patient voice features. 2014.

[24] B. I. Giasson, M. S. Forman, M. Higuchi, L. I. Golbe, C. L. Graves, P. T. Kotzbauer, J. Q. Trojanowski, and V. M.-Y. Lee. Initiation and synergistic fibrillization of tau and alpha-synuclein. *Science*, 300(5619):636–640, 2003.

[25] J. A. Gladsjo, C. C. Schuman, J. D. Evans, G. M. Peavy, S. W. Miller, and R. K. Heaton. Norms for letter and category fluency: demographic corrections for age, education, and ethnicity. *Assessment*, 6(2):147–178, 1999.

[26] C. G. Goetz, B. C. Tilley, S. R. Shaftman, G. T. Stebbins, S. Fahn, P. Martinez-Martin, W. Poewe, C. Sampaio, M. B. Stern, R. Dodel, et al. Movement disorder society-sponsored revision of the unified parkinson's disease rating scale (mds-updrs): scale presentation and clinimetric testing results. *Movement disorders: official journal of the Movement Disorder Society*, 23(15):2129–2170, 2008.

[27] S. Grover, S. Bhartia, A. Yadav, K. Seeja, et al. Predicting severity of parkinson's disease using deep learning. *Procedia computer science*, 132:1788–1794, 2018.

[28] J. L. Guo, D. J. Covell, J. P. Daniels, M. Iba, A. Stieber, B. Zhang, D. M. Riddle, L. K. Kwong, Y. Xu, J. Q. Trojanowski, et al. Distinct α-synuclein strains differentially promote tau inclusions in neurons. *Cell*, 154(1):103–117, 2013.

[29] E. Hartelt, R. Scherbaum, M. Kinkel, R. Gold, S. Muhlack, and L. Tönges. Parkinson's disease multimodal complex treatment (pd-mct): Analysis of therapeutic effects and predictors for improvement. *Journal of clinical medicine*, 9(6):1874, 2020.

[30] P. C. Hawkins, A. G. Skillman, and A. Nicholls. Comparison of shape-matching and docking as virtual screening tools. *Journal of medicinal chemistry*, 50(1):74–82, 2007.

[31] Z. Huang, H. Lei, Y. Zhao, F. Zhou, J. Yan, A. Elazab, and B. Lei. Longitudinal and multi-modal data learning for parkinson's disease diagnosis. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1411–1414. IEEE, 2018.

[32] J.-H. Kang, B. Mollenhauer, C. S. Coffey, J. B. Toledo, D. Weintraub, D. R. Galasko, D. J. Irwin, V. Van Deerlin, A. S. Chen-Plotkin, C. Caspell-Garcia, et al. Csf biomarkers associated with disease heterogeneity in early parkinson's disease: the parkinson's progression markers initiative study. *Acta neuropathologica*, 131(6):935–949, 2016.

[33] V. Kelly, C. Johnson, E. McGough, A. Shumway-Cook, F. Horak, K. Chung, A. Espay, F. Revilla, J. Devoto, C. Wood-Siverio, S. Factor, B. Cholerton, K. Edwards, A. Peterson, J. Quinn, T. Montine, C. Zabetian, and J. Leverenz. Association of cognitive domains with postural instability/gait disturbance in parkinson's disease. *Parkinsonism & Related Disorders*, 21(7):692–697, July 2015.

[34] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, Sept. 1952.

[35] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[36] C. Klein and A. Westenberger. Genetics of parkinson's disease. *Cold Spring Harbor Perspectives in Medicine*, 2(1):a008888–a008888, Jan. 2012.

[37] R. G. Knight, H. J. Waal-Manning, and G. F. Spears. Some norms and reliability data for the state-trait anxiety inventory and the zung self-rating depression scale. *British Journal of Clinical Psychology*, 22(4):245–249, 1983.

[38] M. Knop, P. G. Sämann, and M. E. Keck. Parkinson-komplextherapie am max-planck-institut für psychiatrie: Ein multimodales, flexibles stationäres therapieprogramm für kritische krankheitsphasen. *Psychiatr. Neurol*, 5:24–29, 2017.

[39] A. E. Lang and A. M. Lozano. Parkinson's disease. *New England Journal of Medicine*, 339(16):1130–1143, Oct. 1998.

[40] B. Lei, S. Chen, D. Ni, and T. Wang. Discriminative learning for alzheimer's disease diagnosis via canonical correlation analysis and multimodal fusion. *Frontiers in aging neuroscience*, 8:77, 2016.

[41] W. Li, W. Zhu, E. R. Dorsey, and J. Luo. Predicting parkinson's disease with multimodal irregularly collected longitudinal smartphone data. In *2020 IEEE International Conference on Data Mining (ICDM)*, pages 1106–1111. IEEE, 2020.

[42] M. B. Makarious, H. L. Leonard, D. Vitale, H. Iwaki, L. Sargent, A. Dadu, I. Violich, E. Hutchins, D. Saffo, S. Bandres-Ciga, J. J. Kim, Y. Song, M. Bookman, W. Nojopranoto, R. H. Campbell, S. H. Hashemi, J. A. Botia, J. F. Carter, M. Maleknia, D. W. Craig, K. V. Keuren-Jensen, H. R. Morris, J. A. Hardy, C. Blauwendraat, A. B. Singleton, F. Faghri, and M. A. Nalls. Multi-modality machine learning predicting parkinson's disease. Mar. 2021.

[43] K. Marek, S. Chowdhury, A. Siderowf, S. Lasch, C. S. Coffey, C. Caspell-Garcia, T. Simuni, D. Jennings, C. M. Tanner, J. Q. Trojanowski, et al. The parkinson's progression markers initiative (ppmi)–establishing a pd biomarker cohort. *Annals of clinical and translational neurology*, 5(12):1460–1477, 2018.

[44] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, et al. The parkinson progression marker initiative (ppmi). *Progress in neurobiology*, 95(4):629–635, 2011.

[45] A. Michell, S. Lewis, T. Foltynie, and R. Barker. Biomarkers and parkinson's disease. *Brain*, 127(8):1693–1705, 2004.

[46] S. P. Morgan and J. D. Teachman. Logistic regression: Description, examples, and comparisons. *Journal of Marriage and the Family*, 50(4):929, Nov. 1988.

[47] M. A. Nalls, C. Y. McLean, J. Rick, S. Eberly, S. J. Hutten, K. Gwinn, M. Sutherland, M. Martinez, P. Heutink, N. M. Williams, et al. Diagnosis of parkinson's disease on the

basis of clinical and genetic classification: a population-based modelling study. *The Lancet Neurology*, 14(10):1002–1009, 2015.

[48] Z. S. Nasreddine, N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow. The montreal cognitive assessment, moca: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*, 53(4):695–699, 2005.

[49] W. G. Ondo, K. D. Vuong, H. Khan, F. Atassi, C. Kwak, and J. Jankovic. Daytime sleepiness and other sleep disorders in parkinson's disease. *Neurology*, 57(8):1392–1396, 2001.

[50] G. Pahuja, T. Nagabhushan, B. Prasad, and R. Pushkarna. Early detection of parkinson's disease through multimodal features using machine learning approaches. *International Journal of Signal and Imaging Systems Engineering*, 11(1):31, 2018.

[51] A. Papadopoulos, D. Iakovakis, L. Klingelhoefer, S. Bostantjopoulou, K. R. Chaudhuri, K. Kyritsis, S. Hadjidimitriou, V. Charisis, L. J. Hadjileontiadis, and A. Delopoulos. Unobtrusive detection of parkinson's disease from multi-modal and in-the-wild sensor data using deep learning techniques. *Scientific Reports*, 10(1):1–13, 2020.

[52] C.-h. Park, P. H. Lee, S.-K. Lee, S. J. Chung, and N.-Y. Shin. The diagnostic potential of multimodal neuroimaging measures in parkinson's disease and atypical parkinsonism. *Brain and behavior*, 10(11):e01808, 2020.

[53] Z. Ping, W. Xiaomu, X. Xufang, C. Wenfeng, S. Liang, and W. Tao. GAPDH rs1136666 SNP indicates a high risk of parkinson's disease. *Neuroscience Letters*, 685:55–62, Oct. 2018.

[54] M. Poletti, D. Frosini, C. Pagni, F. Baldacci, V. Nicoletti, G. Tognoni, C. Lucetti, P. D. Dotto, R. Ceravolo, and U. Bonuccelli. Mild cognitive impairment and cognitive-motor relationships in newly diagnosed drug-naive patients with parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 83(6):601–606, Apr. 2012.

[55] I. Rish. An empirical study of the naïve bayes classifier. *IJCAI 2001 Work Empir Methods Artif Intell*, 3, 01 2001.

[56] L. S. Rosenthal, D. Drake, R. N. Alcalay, D. Babcock, F. D. Bowman, A. Chen-Plotkin, T. M. Dawson, R. B. Dewey, D. C. German, X. Huang, B. Landin, M. McAuliffe, V. A. Petyuk, C. R. Scherzer, C. S. Hillaire-Clarke, B.-A. Sieber, M. Sutherland, C. Tarn, A. West, D. Vaillancourt, J. Zhang, and K. G. and. The NINDS parkinson's disease biomarkers program. *Movement Disorders*, 31(6):915–923, Oct. 2015.

[57] M. R. Salmanpour, M. Shamsaei, A. Saberi, S. Setayeshi, I. S. Klyuzhin, V. Sossi, and A. Rahmim. Optimized machine learning methods for prediction of cognitive outcome in parkinson's disease. *Computers in Biology and Medicine*, 111:103347, Aug. 2019.

[58] A. H. Shahid and M. P. Singh. A deep learning approach for prediction of parkinson's disease progression. *Biomedical Engineering Letters*, 10(2):227–239, Apr. 2020.

[59] A. Sharma, N. Osato, H. Liu, S. Asthana, T. C. Dakal, G. Ambrosini, P. Bucher, I. Schmitt, and U. Wüllner. Common genetic variants associated with parkinson's disease display widespread signature of epigenetic plasticity. *Scientific Reports*, 9(1), Dec. 2019.

[60] T. Simuni, C. Caspell-Garcia, C. S. Coffey, D. Weintraub, B. Mollenhauer, S. Lasch, C. M. Tanner, D. Jennings, K. Kieburtz, L. M. Chahine, et al. Baseline prevalence and longitudinal evolution of non-motor symptoms in early parkinson's disease: the ppmi cohort. *Journal of Neurology, Neurosurgery & Psychiatry*, 89(1):78–88, 2018.

[61] M. Skorvanek, J. G. Goldman, M. Jahanshahi, C. Marras, I. Rektorova, B. Schmand, E. van Duijn, C. G. Goetz, D. Weintraub, G. T. Stebbins, et al. Global scales for cognitive screening in parkinson's disease: Critique and recommendations. *Movement Disorders*, 33(2):208–218, 2018.

[62] A. Smith. *Symbol digit modalities test*. Western Psychological Services Los Angeles, 1973.

[63] K. Stiasny-Kolster, G. Mayer, S. Schäfer, J. C. Möller, M. Heinzel-Gutenbrunner, and W. H. Oertel. The rem sleep behavior disorder screening questionnaire—a new diagnostic instrument. *Movement disorders*, 22(16):2386–2393, 2007.

[64] A. Tsanas, M. Little, P. McSharry, and L. Ramig. Accurate telemonitoring of parkinson's disease progression by non-invasive speech tests. *Nature Precedings*, pages 1–1, 2009.

[65] M. Visser, J. Marinus, A. M. Stiggelbout, and J. J. Van Hilten. Assessment of autonomic dysfunction in parkinson's disease: the scopa-aut. *Movement disorders: official journal of the Movement Disorder Society*, 19(11):1306–1312, 2004.

[66] M. Visser, J. Marinus, A. M. Stiggelbout, and J. J. Van Hilten. Assessment of autonomic dysfunction in parkinson's disease: the scopa-aut. *Movement disorders: official journal of the Movement Disorder Society*, 19(11):1306–1312, 2004.

[67] D. Wechsler. Wechsler adult intelligence scale–. 1955.

[68] D. Weintraub, S. Hoops, J. A. Shea, K. E. Lyons, R. Pahwa, E. D. Driver-Dunckley, C. H. Adler, M. N. Potenza, J. Miyasaki, A. D. Siderowf, et al. Validation of the questionnaire for impulsive-compulsive disorders in parkinson's disease. *Movement disorders: official journal of the Movement Disorder Society*, 24(10):1461–1467, 2009.

[69] D. Weintraub, K. A. Oehlberg, I. R. Katz, and M. B. Stern. Test characteristics of the 15-item geriatric depression scale and hamilton depression rating scale in parkinson disease. *The American journal of geriatric psychiatry*, 14(2):169–175, 2006.

[70] J. Wills, J. Jones, T. Haggerty, V. Duka, J. N. Joyce, and A. Sidhu. Elevated tauopathy and alpha-synuclein pathology in postmortem parkinson's disease brains with and without dementia. *Experimental neurology*, 225(1):210–218, 2010.