Investigating biomarkers in Parkinson's disease using machine learning

by

© Ali Ameli

A Thesis submitted to the

School of Graduate Studies

in partial fulfillment of the

requirements for the degree of

Master of Science

Supervisor: Dr. Hamid Usefi

Co-supervisor: Dr. Lourdes Peña-Castillo

Department of Computer Science

Memorial University of Newfoundland

October 2022

St. John's Newfoundland and Labrador

Abstract

Genome-Wide Association Studies (GWAS) identify genetic variations in individuals affected with diseases such as Parkinson's disease (PD), whose allele or genotype frequencies are significantly different between the affected individuals and individuals who are free of the disease. GWAS data can be used to identify genetic variations associated with the disease of interest. However, GWAS datasets are extensive and contain many more Single Nucleotide Polymorphisms (SNPs pronounced "snips") than individual samples. To address these challenges, we used Singular-Vectors Feature Selection (SVFS) and applied it to PD GWAS datasets. We discovered a group of SNPs that are potentially novel PD biomarkers as we found indirect links between them and PD in the literature but have not directly been associated with PD before. Direct association means that current literature directly links a SNP with PD; while an indirect link means that current literature suggests the involvement of a SNP in a disease other than PD but this other disease co-occurs with PD in a significant number of PD patients. These indirectly-linked SNPs open new potential lines of investigation. Directly-linked SNPs identified by our method are rs11248060, rs239748, rs999473, and rs2313982. One can see the full list of identified SNPs in Section 4.4.

Acknowledgments

I would like to thank my supervisors, who guided me in doing this thesis. They provided me with invaluable advice and helped me in difficult periods. Their motivation and help contributed tremendously to the successful completion of the thesis.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

A	Adenine
AI	Artificial Intelligence
APDGC	Autopsy Confirmed Parkinson Disease GWAS Consortium
API	Application programming interface
BED	Browser Extensible Data
C	Cytosine
CNN	Convolutional Neural Networks
CSF	Cerebrospinal Fluid
CSV	Comma-Separated Value
CV	Cross-Validation
DL	Deep Learning
DNA	Deoxyribonucleic Acid
DTs	Decision Trees
EEG	Electro Encephalogram
ENPP	Enhanced Permutation tests through Multiple Pruning
FAM	Family

FID	Familial ID
G	Guanine
GB	Gradient Boosting
GRCH	Genome Reference Consortium Human
GWAS	Genome-Wide Association Studies
HSIC	Hilbert-Schmidt Independence Criterion
IGR	Intergenic Region
IoT	Internet of Things
KNNs	K-Nearest Neighbour
LD	Linkage Disequilibirum
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Machine Learning
MLPs	Multi-Layered Perceptrons
MRI	Magnetic Resonance Imaging
NA	Not Available
NB	Naive Bayes

NCBI	National Center for Biotechnology Information
NIA	National Institute on Aging
NIH	National Institutes of Health
NN	Neural Network
NRMSA	Normalized Root Mean Square Error
OCT	Optical Coherence Tomography
PCC	the distance from the Origin point to a selected point
PD	Parkinson's Disease
PED	Pedigree format
PPMI	Parkinson's Progression Markers Initiative
QoL	Quality of Life
RF	Random Forests
RNA	Ribonucleic acid
SMC	Simple matching coefficient
SNIPA	Seronegative Inflammatory Polyarthritis
SNPs	Single Nucleotide Polymorphisms
SPECT	Single-Photon Emission Computed Tomography

SRA	Sequence Read Archive
SVFS	Singular-Vectors Feature Selection
SVM	Support Vector Machine
T	Thymine
dbGaP	database of Genotypes and Phenotypes
mRMR	minimum Redundancy Maximum Relevance

Chapter 1

Introduction

Parkinson's disease (PD) is one of the most prevalent neurodegenerative disorders affecting 1-2 persons per 1,000 people and has a prevalence rate of 1% over the age of 60 [1]. Due to a growth in the number of senior individuals and age-standardized incidence rates, the estimated number of people affected with PD in the world more than doubled (from 2.5 million to 6.1 million) between 1990 and 2016 [2]. According to Jankovic [3], PD is a degenerative neurological condition that affects both the motor and non-motor aspects of movement including planning, initiation, and execution [4].

Parkinson's symptoms appear in people who have lost 80% or more of dopamine-producing cells in the substantia nigra region of the brain [5]. To assist with the coordination of the millions of nerve and muscle cells involved in movement, dopamine often works in a careful balance with other neurotransmitters. Without adequate dopamine, this equilibrium is disrupted, leading to the typical symptoms of PD, including tremor (trembling in the hands, arms, legs, and jaw), rigidity (stiffness of the limbs), slowness of movement, and decreased balance and coordination [6]. Patients with PD experience severe impairments in their quality of life (QoL), ability to engage in social activities, and ability to maintain healthy family relationships [7]–[9].

Motor symptoms have historically been used to make PD diagnosis. Although the cardinal indications of PD have been established in clinical examinations, most of the rating scales used to determine the disease severity have not been thoroughly examined and verified [3]. While non-motor symptoms (such as cognitive changes, difficulties with attention and planning, sleep disorders, sensory abnormalities, and olfactory dysfunction) are common in patients before the onset of PD, they lack specificity, are challenging to assess, and/or vary from patient to patient [3], [10], [11]. Non-motor symptoms cannot currently be utilized to diagnose PD on their own, despite some of them being used as supportive diagnostic criteria [12], [13]. Although there is no cure, there are numerous treatment options available, including drugs, lifestyle changes, and surgery. PD is not lethal in and of itself, although significant complications might arise.

In the healthcare industry, the use of Machine Learning (ML) techniques is expanding. As the term "Machine Learning" suggests, it is possible for computer software to learn from data in a semi-automatic way and extract meaningful representations from it. Handwriting patterns ([14], [15]), movement ([16], [17]), neuroimaging

([18]–[20]), speech ([21], [22]), Cerebrospinal Fluid (CSF) ([23], [24]), cardiac scintigraphy ([25]), serum ([26]), and Optical Coherence Tomography (OCT) ([27]) are just a few of the data modalities that have been subjected to ML models for the PD diagnosis. In order to diagnose PD, ML also enables the combination of several modalities, such as Magnetic Resonance Imaging (MRI) and Single-Photon Emission Computed Tomography (SPECT) data [18], [28]. Therefore, we can rely on these alternative measures to diagnose the disease in its preclinical phases or atypical forms to uncover pertinent elements that are not frequently used in the clinical diagnosis of PD.

Genome-Wide Association Studies (GWAS) look for genetic variants (particularly Single Nucleotide Polymorphisms (SNPs)) in people with a particular disease and those without. This information can be utilized to find SNPs linked to the disease of interest. SNPs (pronounced “snips”) are human's most prevalent form of genetic variation [29]. GWAS datasets are massive and include considerably more SNPs than individual samples.

In this thesis, we will focus on finding SNPs that could be used as biomarkers for PD. The term “biomarker” [30], a portmanteau of “biological” and “marker”, refers to a large subclass of molecular and medical signs or objective indicators of health status, which can be quantified precisely and consistently.

We used GWAS data from PD cases and healthy controls (here healthy controls refer to people without PD but they might have other health-related issues). These datasets are described in Section 3.1, and downloaded from the database of Genotypes and Phenotypes (dbGaP) [1]. We worked with five PD datasets, including Tier1 [31], NINDS1 [32], NINDS2 [32], Familial [33], and Autopsy [34] datasets. To prepare the datasets for the feature selection method, we preprocessed and cleaned them. These datasets were all decrypted, then converted into a Comma-Separated Values (CSV) format. After converting the GWAS data into CSV format, we used a data imputation technique, KNNcatimpute [35] to impute Not Available (NA) values.

As it was mentioned before, GWAS data is extensive and building a model on such large datasets takes a long time. So, due to the high dimensionality of the GWAS dataset, we have to reduce dimensionality. The standard technique for dimensionality reduction is feature selection. Feature selection is the process of choosing the most reliable, non-redundant, and pertinent features to include in a model and its primary objectives are to enhance a predictive model's performance and lower modelling's computational expense. We used the Singular-Vectors Feature Selection (SVFS) algorithm [36] as the feature selection technique. Let $D = [A \mid \mathbf{b}]$ be a labelled dataset, with **b** representing the class label and features (attributes) representing columns in matrix A. M. Afshar and H. Usefi showed how the signature matrix $S_A = I - A^\dagger A$ (I is the identity matrix and $A^\dagger$ is the pseudo-inverse of A) can be used to partition A's columns into clusters, with columns in one cluster correlating exclusively with columns in another cluster. The signature matrix S_D of D is used by SVFS to locate the cluster that holds **b**. Afshar and Usefi reduced the size of A by eliminating irrel-

[1] https://www.ncbi.nlm.nih.gov/gap/

evant features from the other clusters. The signature matrix S_A of reduced A is then used by SVFS to partition the remaining features into clusters and select the most important features from each cluster. We also evaluated HSIC-Lasso [37] feature selection as another technique. HSIC Lasso is one of the best techniques for choosing sparse nonlinear features and is based on the Hilbert-Schmidt independence criterion [38]. HSIC Lasso can be regarded as a convex variant of widely used minimum redundancy maximum relevance (mRMR) feature selection algorithm. However, the HSIC-Lasso's results were not as good as SVFS's (see Table 3.7). Consequently, we used SVFS for feature selection for all of the approaches.

We proposed five different approaches to integrate datasets and compare them with the baseline approach of no integration. By integrating datasets, we mean combining independent datasets into a single dataset containing data points from each of the original datasets to be used for generating a machine learning model for detecting PD cases. We used Random Forests (RF) as the classifier for all approaches. We also tested Support Vector Machine (SVM) and Gradient Boosting (GB), but they required a lengthy time to process the datasets and performed less accurately than RF.

In Approach 0, we ran the SVFS feature selection algorithm on each dataset separately to extract the most important features. After 50 rounds, we constructed a dictionary of SNP IDs with the number of times each SNP was selected by the SVFS (henceforth referred to as frequency). For each dataset we obtained the SNPs with the highest frequency and using Cross-Validation (CV), we assessed the classification performance of a model generated using these highly frequent SNPs as features. We considered this approach as the baseline for our analysis. Out of all the approaches, we found that Approach 0 (baseline approach) had the best accuracy on each dataset compared to other approaches. We obtained the highest (87.46%) and lowest (51.92%) accuracy on the NINDS2 and Tier1 datasets, respectively.

In Approach 1, we ran the SVFS feature selection algorithm on the Familial dataset and selected the most frequent SNPs, then performed CV to assess the classification performance of each of the other four datasets. Due to the limited availability of SNPs in Approach 1, we expected Approach 1 to be less accurate.

In Approach 2, we first obtained the intersection of SNPs between the Familial dataset and each of the other four datasets. SNPs not in the intersection were removed. We followed the same steps that had been done for Approach 1 by extracting the most frequent SNPs from the condensed version of the Familial dataset and doing CV on the other datasets. Approaches 1 and 2 had comparable performance with average accuracy (Autopsy = 64.89%, NINDS1 = 51.98%, NINDS2 = 53.04%, and Tier1 = 44.47%) and (Autopsy = 65.40%, NINDS1 = 52.56%, NINDS2 = 50.38%, and Tier1 = 31.50%), respectively.

In Approach 3, to enhance the classification performance, we combined datasets to increase the number of instances (individuals) before feature selection. We obtained the SNPs in the intersection between the Familial dataset and the other four datasets (common SNPs between Familial and other datasets). We ran the SVFS feature selection algorithm on each of the four merged datasets and extracted the most frequent SNPs. Then, performed CV to assess the classification performance of the model generated on each of the merged datasets.

Approach 4 was the same as Approach 3, but an equal number of instances per dataset were merged. The number of cases and healthy controls taken from each dataset was the same. The trend between approaches 3 and 4 was the same, but these two approaches performed better than Approach 1 and 2.

After performing approaches 0 to 4, we had different list of most frequent SNPs from each approach for each dataset. We collected SNPs that are in common among different approaches for the same dataset and among different datasets. There are some SNPs that are in common among at least two approaches or two datasets. We extracted those SNPs and called them the possible biomarkers for PD. Additionally, we investigated the associated phenotypes of selected SNPs to find out any potential link with PD. The main results of this thesis were discussed in Chapter 4. We discovered rs11248060, rs239748, rs999473, and rs2313982 that were directly linked to PD. We presented our candidate list of SNPs associated with PD in Section 4.4. Further clinical investigations are required to validate these findings.

The main contributions of this thesis are:

- Showing that feature selection and machine learning algorithms can be used for identifying SNPs potentially associated with PD.

- Comparing the impact of integrating data sets in the identification of SNPs potentially associated with PD.

- Identifying a number of SNPs as potential biomarkers of PD not yet mentioned in the literature.

The summary of the contents of the subsequent chapters are:

- Chapter 2 - Background and related works: Recent research on applying ML algorithm on GWAS and PD datasets

- Chapter 3 - Methodology: All steps of applying ML approaches on PD datasets

- Chapter 4 - Results and Discussion: Identified SNPs as the biomarkers of PD

- Chapter 5 - Conclusion: Major findings in relation to the objectives and research questions and limitation of the work

Chapter 2

Background and related works

2.1 Background

Deoxyribonucleic acid (DNA) is a long, double-stranded molecule that contains all the instructions needed to develop and direct living things [39]. Each strand of DNA is made up of four chemical units called nucleotides (Adenine (A), Cytosine (C), Guanine (G), and Thymine (T)). These units are, in fact, the letters of the genetic alphabet. The two strands of DNA complement each other and are paired together. The pairing process is such that A is always paired with T, and C is always paired with G.

The complete DNA set in any living thing is called its genome [40]. Because DNA is almost always present in two strands, the length of the genome is measured in base pairs. The genome is stored in long molecules of DNA called chromosomes.

A gene is a region of DNA that is transcribed [41]. The genes are copied into RNA molecules. Some of these RNA molecules contain the instructions to produce proteins.

The entire content of human nuclear DNA is divided into 46 chromosomes [42]. Chromosomes are inherited from each parent to the offspring as a set of 23. Thus, there are two copies of most genes in each cell. These different versions of a single gene are called alleles. Some alleles [43] may cause a particular trait (phenotype) in an organism. The composition and set of alleles that an organism carries are called the genotypes and is often expressed in letters [44]. All visible traits in an organism that result from the interaction of its genotype with the environment are called phenotypes [45]. Examples of phenotypes include the colour, shape, size of the organism, and its behaviour, and susceptibility to certain diseases. An organism's phenotype may change during its life with environmental changes or physiological and morphological changes resulting from ageing.

A locus is a specific location on the genome [46]. SNPs are the most prevalent form of genetic variation in humans. Each SNP is a variation in a single nucleotide. In a specific locus, a SNP might, for instance, swap out the nucleotide cytosine (C) with the nucleotide thymine (T). SNPs typically occur all over a person's DNA. There are around 4 to 5 million SNPs in an individual's genome, which implies they typically occur almost once every 600 to 750 nucleotides (3 billion nucleotides in the human genome divided by 4 or 5 million SNPs). Every single individual has SNPs;

nevertheless, for a variation to be called SNP, it must be present in at least 1% of the population [29]. More than 600 million SNPs have been discovered by researchers in human populations worldwide.

2.2 Using ML on GWAS data

There have been many research projects using ML on GWAS data in recent years.

ML applications in GWAS were explored by Nicholls et al., 2020 ([47]). This review article focused on three components: selected models, input features, and output model efficiency. The authors focused on prioritizing complex disease-associated loci and the contributions made by ML to achieving the GWAS end-game, with wide-ranging translational implications. GWAS end-game is a situation in which all common population variation that affects a characteristic has been recognized, offering sound scientific explanations and mechanisms with a dependable translational capacity [47]. According to this paper, many ML algorithms are used for post-GWAS analysis. Still, the most common ones are Gradient Boosting (GB), Random Forests (RF), Support Vector Machine (SVM), Logistic Regression (LR), and Neural Network (NN).

Enhanced Permutation tests through Multiple Pruning (ENPP), proposed by Leem et al., 2020 is a permutation method for GWAS [48]. If the features in each permutation round are found to be non-significant, ENPP prunes them. They used this method to find the association with a non-normally distributed phenotype (fasting plasma glucose) in an actual dataset (Korea Association REsource: KARE [49]) of 327,872 SNPs.

A. Nalls et al., 2014 performed an analysis of PD across 7,893,274 variants, 13,708 cases, and 95,282 healthy controls [50]. They performed a meta-analysis of all existing European ancestry PD GWAS study data. They applied their methods to three datasets. The genomic inflation factor for each dataset was between 0.889 and 1.056 and was calculated for each chromosome separately as well as for the entire genome for the various densities. It was defined as the median of the observed chi-squared test statistics divided by the expected median of the corresponding chi-squared distribution. The $p-$value threshold was 5×10^{-8}. They found 28 independent risk loci for PD.

2.3 Classification on PD patients using ML

We can utilize ML algorithms to uncover pertinent aspects that are not often employed in the clinical diagnosis of PD and rely on these alternative measures to detect the disease in its preclinical stages or in atypical forms [51].

With the objective of enabling improved individualized treatments and evaluating the suggested ones, Artificial Intelligence (AI) and Internet of Things (IoT) technologies can support both the early diagnosis of PD and the monitoring of PD patients [52], adding to the already well-established conventional procedures. Additionally, the evaluation of the efficacy of currently prescribed medications, as well as the optimization of surgical treatments, the prediction of the course of the disease, and the prevention of unfavourable consequences even in real-time, can be facilitated by training ML algorithms on sensory data collected from PD patients [53]. These interventions facilitate the shift from clinic-centric to patient-centric healthcare practices and pave the way for precision therapy in PD and other chronic diseases [54].

The classification of PD patients and healthy controls, which is usually addressed based on inertial signals, is the first issue that will be tackled. In order to accomplish this, gait features have been extracted manually using feature engineering approaches [55]–[59] or automatically using deep Convolutional Neural Networks (CNNs) [60], and these features have been fed into several classification algorithms. SVM, Decision Trees (DTs), RF, K-Nearest Neighbour (KNN), bagged, boosted, and fine trees, LR, Linear Discriminant Analysis (LDA), and Naive Bayes (NB) classifiers, as well as Multi-Layered Perceptrons (MLPs) or other NNs, are some of the deployed algorithms.

There is a wealth of genetic and transcriptome data of patients with PD thanks to high-throughput methodologies, but traditional statistical methods used for data analysis have not produced much in the way of insightful integrated analysis or interpretation of the data [61]. ML has thus been used to evaluate and interpret extensive, extremely complicated genomic and transcriptome data in order to gain better insights into PD. ML models have been created in particular to integrate patient genotype data either alone or in combination with demographic, clinical, neuroimaging, and other data for PD outcome studies. Additionally, they have been applied to discovering PD biomarkers based on transcriptome information, such as gene expression patterns from microarrays [62].

Several studies looking into the role of the gut microbiota in PD found some common microbial population changes in PD patients, such as a decrease in Lachnospiraceae and an increase in Verrucomicrobiaceae families [63]. They analyzed 165 rRNA gene sequencing data from six different studies using three different supervised ML algorithms. They developed a classifier that can predict the pathological status of PD patients compared to healthy controls as a result of this research, and they established a subset of 22 bacterial families that are discriminative for the prediction.

$DEEP_{ENA}$ is a Deep Learning (DL) technique that is used to determine whether or not an individual has PD based on premotor features [64]. Specifically, several indicators were considered in this study to detect PD at an early stage, including rapid eye movement, olfactory loss, cerebrospinal fluid data, and dopaminergic imaging markers. A comparison of the proposed DL model with twelve ML and ensemble learning methods based on relatively limited data, including 183 healthy controls and 401 early onset PD patients, reveals that the built model has the best detection efficiency, with an average accuracy of 96.45%. Their case study was the Parkinson's Progression Markers Initiative (PPMI) [65] database (401 early onset PD patients and 183 healthy controls).

Ahmadi Rastegar et al., 2019 tested 27 inflammatory cytokines and chemokines in serum at baseline and after a year to examine cytokine stability. Cytokines might be useful in ML models for PD progression prediction. The baseline measurements were then combined with ML models to predict longitudinal clinical outcomes after a two-year follow-up [66]. The best prediction models achieved a Normalized Root Mean Square Error (NRMSE) of 0.1123 and 0.1993 in the motor symptom severity scales of Hoehn and Yahr and the unified PD rating scale part three, respectively. Their case study was the Michael J Fox Foundation LRRK2 [67] clinical consortium longitudinal sample collection (serum samples and clinical data from 160 patients for a baseline comparison of LRRK2-PD and idiopathic PD).

Singh Dhami et al., 2017 proposed a ML method that takes as input a specific collection of data from the PPMI study and divides them into two classes: PD cases and healthy controls [68]. They tested their method on 1194 patients obtained from the PPMI, and the results demonstrate that it achieves cutting-edge performance with minimal feature engineering.

There is no standard procedure for making a PD diagnosis, which makes it a challenging and time-consuming task. As a result, numerous investigations have been carried out to identify reliable PD biomarkers. One method that has been applied in the search for biomarkers is the examination of electroencephalogram (EEG) signal features [69]. This study assessed the efficacy of EEG Hjorth features as biomarkers for PD. SVM, KNN, and RF algorithms were used for classification, following a 5-fold CV methodology, using the database that is accessible at the public repository known as The Patient Repository for EEG Data Computational Tools (PRED + CT). With an SVM classifier, the suggested model distinguished between PD patients and healthy controls with an accuracy of 89.56%.

Alex Li and Chenyu Li used ML techniques to create a classifier using gait data from Parkinson'patients and healthy controls [70]. A more precise and affordable diagnostic procedure might be facilitated by the classifier. The Gait in PD dataset, available on PhysioNet [71], [72], is the input to their algorithm. It contains force sensor data used to measure the gait of 214 individuals with idiopathic PD and 92 healthy controls. A classification model of Parkinson's patients and healthy controls was created using a variety of ML approaches, including LR, SVM, DT, and KNN.

A cohort of cognitively healthy controls' s single-cell chromatin accessibility landscapes and three-dimensional chromatin interactions were profiled to create a multi-omic epigenetic atlas of the adult human brain by Corces et al., 2020 ([73]). With the help of a ML classifier, authors were able to integrate this multi-omic framework and predict dozens of functional SNPs for Alzheimer's and Parkinson's disorders, as well as target genes and cell types for hitherto orphaned locations from GWAS.

This Chapter presented the previous promising work in biomarker discovery using ML for classification purposes, and the use of ML and feature selection algorithms to identify SNPs has been studied [74]. However, what made our study different was the use of ML techniques on novel PD datasets. We deployed different feature selection algorithms in our research, contributing to the area of interest.

Chapter 3

Methodology

We will apply the SVFS algorithm [36] and ML classifiers on five GWAS PD datasets to differentiate between PD patients and healthy controls. In this section the methods and datasets used will be described.

3.1 Dataset description

We used five different datasets. We obtained those datasets from Genotype and Phenotype (dbGaP) database [75].

1. Phs000126 (Familial) [33], [76]–[82] dataset combines the results of two major National Institutes of Health (NIH)-funded genetic research aimed at discovering new genes that influence the risk of PD. PROGENI (PI: Tatiana Foroud; R01NS037167) and GenePD (PI: Richard Myers; R01NS036711) have been analyzing and recruiting families with two or more PD affected members for over

eight years. There are almost 1,000 PD families in the total sample.

Type	Source	Platform
Whole Genome Genotyping	Illumina	HumanCNV370v1

Table 3.1: Technology/Platform for genotyping for Familial dataset

2. Phs000394 (Autopsy)-Confirmed Parkinson Disease GWAS Consortium (APDGC) [34] was established to perform a genome-wide association research in people with neuropathologically diagnosed PD and healthy controls. Their study's hypothesis is that by enrolling only cases and healthy controls with neuropathologically proven illness status, diagnostic misclassification will be reduced and power to identify novel genetic connections will be increased.

Type	Source	Platform
Whole Genome Genotyping	Illumina	$HumanOmni1 - Quad\_v1 - 0\_B$

Table 3.2: Technology/Platform for genotyping for Autopsy dataset

3. Phs000089 (NINDS) [32], [83]–[86] repository was created in 2001 with the intention of creating standardised, widely applicable diagnostic and other clinical data as well as a collection of DNA and cell line samples to enhance the field of neurological illness gene discovery. All samples, phenotypic information, and genotypic information are accessible. The collection also includes well-described neurologically healthy controls subjects. This collection served as the founda-

tion for both the expanded investigation by Simon-Sanchez et al. and the first stage study by Fung et al [83]. The laboratories of Dr. Andrew Singleton of the National Institute on Aging (NIA) and Dr. John Hardy of the NIA produced and submitted the genotyping data (NIH Intramural, funding from NIA and NINDS). NINDS dataset is divided into NINDS1 and NINDS2 (NINDS2 is a subset of NINDS1).

Type	Source	Platform
Whole Genome Genotyping	Illumina	HumanHap250Sv1.0

Table 3.3: Technology/Platform for genotyping for NINDS1 & NINDS2 dataset

4. The dbGaP team at NCBI calculated this Genome-Wide Association scan phs000048 (Tier 1) [31], [87]–[89] between genotype and PD status. 443 sibling pairs that were at odds for PD served as the samples. Between June 1996 and May 2004, the sibling pairs were drawn from the Mayo Clinic's Rochester, Minnesota, Department of Neurology's clinical practise. Drs. Maraganore and Rocca used three Perlegen DNA chips per person and 85k SNP markers to provide genotype data.

Type	Source	Platform
Whole Genome Genotyping	Perlegen	PERLEGEN-85K

Table 3.4: Technology/Platform for genotyping for Tier1 dataset

3.2 Data collection

All used data are gathered from the National Center for Biotechnology Information (NCBI) website [90].

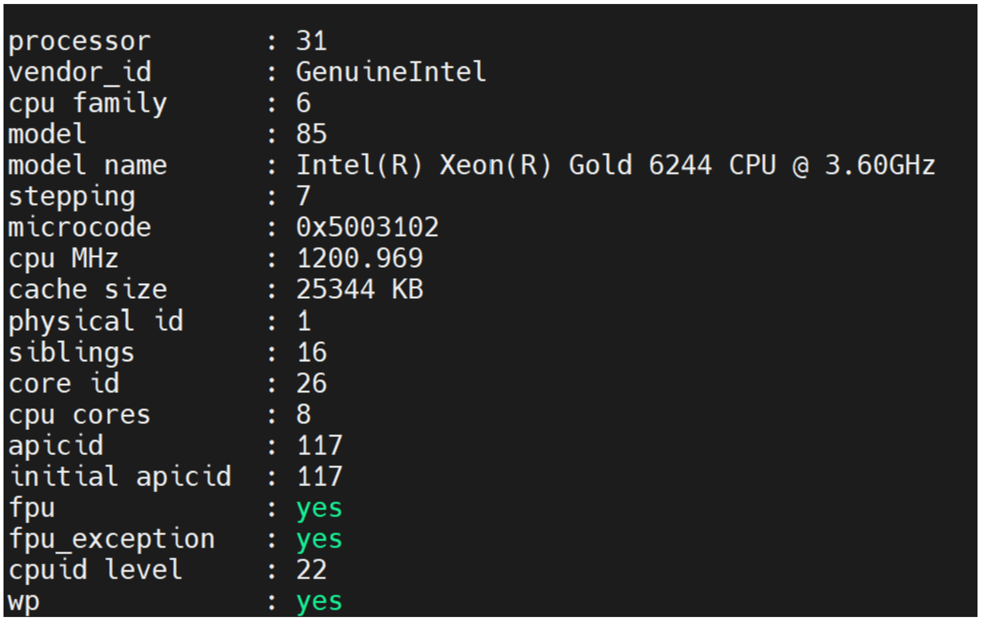
Table 3.5, shows a summary of used datasets. The provided information is extracted from the ped.log of each dataset.

Dataset ID	Samples	Missing Phenotype	Cases	healthy controls	SNPs
1. phs000394 (Autopsy)	1001	24	642	335	1134514
2. phs000126 (Familial)	2082	315	900	867	344301
3. phs000089 (NINDS1)	1741	0	940	801	545066
4. phs000089 (NINDS2)	526	0	263	263	241847
5. phs000048 (Tier1)	886	0	443	443	198345

Table 3.5: Dataset description

3.3 Computing machine hardware details

We used a computing server for all of our computations. The system has two NVIDIA RTX 800 GPUs and around 250G of RAM. The full specifications are listed below:



```
processor       : 31
vendor_id       : GenuineIntel
cpu family      : 6
model           : 85
model name      : Intel(R) Xeon(R) Gold 6244 CPU @ 3.60GHz
stepping        : 7
microcode       : 0x5003102
cpu MHz         : 1200.969
cache size      : 25344 KB
physical id     : 1
siblings        : 16
core id         : 26
cpu cores       : 8
apicid          : 117
initial apicid  : 117
fpu             : yes
fpu_exception   : yes
cpuid level     : 22
wp              : yes
```

Figure 3.1: CPU details

3.4 Data gathering

3.4.1 SRA toolkit installation

Before starting to preprocess the datasets, we need to decrypt and convert the datasets into PED [91] format. PED format is a standard format among genomic datasets. After downloading the whole datasets, one would perform the following steps:

1. Install SRA toolkit [92] on the server:

    - Download the installation file from: `https://github.com/ncbi/sra-tools/wiki/02.-Installing-SRA-Toolkit`

    - After downloading the toolkit, extract it with this command:

```
$ tar -vxzf sratoolkit.tar.gz
```

- For convenience (and to show where the binaries are), append the path to the binaries to the PATH environment variable:

  ```
  $ export PATH=$PATH:$PWD/sratoolkit.2.
  4.0-1.mac64/bin$
  ```

- Verify that the shell will find the binaries:

  ```
  $ which fastq-dump
  ```

- This should produce output similar to:

  ```
  $ /Users/JoeUser/sratoolkit.2.4.0-1.m
  ac64/bin/fastq-dump$
  ```

- Proceed:

  https://github.com/ncbi/sra-tools/wiki/Quick-Toolkit-Configuration

- Test that the toolkit is functional:

  ```
  $ fastq-dump --stdout SRR390728 | head -n 8
  ```

  Within a few seconds, the command should produce this exact output (and nothing else):

  ```
  $ @SRR390728.1 1 length=72 CAT\\
  TCTTCACGTAGTTCTCGAGCCTTGGTTTTCAG
  C GATGGAGAATGACTTTGACAAGCTGAGAGA
  ```

```
AGNTNC +SRR390728.1 1 length=72

;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;66514

2;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&

(@SRR390728.2 2 length=72 AAGTAGGTC

TCGTCTGTGTTTTCTACGAGCTTGTG TTCCAGCTG

ACCCACTCCCTGGGTGGGGGGACTGGGT +SRR390

728.2 2 length=72 ;;;;;;;;;;;;;;;;;4

;;;;3;393.1+4&&5&& ;;;;;;;;;;;;;;;;;

;;;;<9;<;;;;;464262$
```

3.4.2 Dataset decryption

We need to decrypt the “matrixfmt” file for our research. To do that, we performed the following steps:

- Decrypt the file named: phg000233.v1.CIDR_AutopsyPD.genotype-calls matrixfmt.c1.ARU.tar.ncbi_enc. As an example, the commands are showing the steps for the Autopsy dataset.

  This file is encrypted with a.ngc key value.

- Read this guideline before applying the decryption command on the datasets: `https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=toolkit_doc&f=vdb-decrypt`

- Run this command for decryption:

```
$ /gpfs/home/aameli/sratoolkit.2.11.0-
ubuntu64/bin/vdb-decrypt --ngc /
research/ project/cs-genomics/
PD/Datasets/68660/prj_21073.ngc
/research/project/ cs-genomics/
PD/Datasets/68660/ PhenoGenoty
peFiles/RootStudyConsentSet _
phs000394.CIDR_AutopsyPD.v1.p1
.c1.ARU/ GenotypeFiles/phg0
00233.v1.CIDR_AutopsyPD .genoty
pe-calls-matrixfmt.c1.ARU.tar.ncbi_enc$
```

- The output of this command is a folder that contains three files: ".bed", ".bim", ".fam".

3.4.3 What are BED, BIM, FAM, PED, MAP files?

The genotyping data is contained in a binary file called the BED [93] file. The SNP names and MAP coordinates are contained in the BIM [94] file.

A BIM file contains the following six fields [95]:

- Chromosome code (either an integer, or "X"/"Y"/"XY"/"MT"; "0" indicates unknown) or name

- Variant identifier
- Position in morgans or centimorgans (safe to use dummy value of “0”)
- Base-pair coordinate (1-based; limited to 231-2)
- Allele 1 (corresponding to clear bits in .bed; usually minor)
- Allele 2 (corresponding to set bits in .bed; usually major)

The FAM [96] file contains the Familial structure, where the affection status for each individual in the 6^{th} column should be specified.

- Familial ID (“FID”)
- Within-Familial ID (“IID”; cannot be “0”)
- Within-Familial ID of father (“0” if father is not in dataset)
- Within-Familial ID of mother (“0” if mother is not in dataset)
- Sex code (“1” = male, “2” = female, “0” = unknown)
- Phenotype value (“1” = control, “2” = case, “-9”/“0”–out/non-numeric = missing data if case/control)

For sample pedigree information and genotyping calls, the PED format is an original standard text format. Usually requires a “.map” file to be included.

- CHR Chromosome code

- SNP Variant identifier
- BETA Regression slope for real data. Only present with "–qfam emp-se".
- EMP_BETA Sample mean of permutation regression slopes. Only present with "–qfam emp-se".
- EMP_SE Sample stdev of permutation regression slopes. Only present with "–qfam emp-se".
- EMP1 Empirical $p-$value (pointwise), or lower $p-$value permutation count
- NP Number of permutations performed for this variant

A variant information file in MAP [97] format is included with a ".ped" text pedigree and genotyping table.

- Chromosome code: PLINK [98] 1.9 also permits contig names here, but most older programs do not.
- Variant identifier
- Position in morgans or centimorgans (optional; also safe to use dummy value of "0")
- Base-pair coordinate

3.4.4 Dataset conversion to PED format with PLINK

We made sure PLINK [98] is installed on the server properly. If it is not installed on one's operating system, one can obtain PLINK at:

`https://zzz.bwh.harvard.edu/plink/`

- Create a text file which has:

  allfiles.txt

  CIDR AutopsyPD Top sample level. bed

  CIDR AutopsyPD Top sample level. bim

  CIDR AutopsyPD Top sample level. fam

  Note: Sometimes, these three files are zipped. So, extract them before going to the next step. The file names mentioned are from the Autopsy dataset.

- Use PLINK to convert these three files into PED format:

```
$ plink - -bfile dbGaP_AutopsyPD_filter

- -merge-list allFiles.txt - -recode

- -allele1234 - -out PD_5 - -noweb$
```

- After running the mentioned command, there will be some outputs. The outputs which we are looking for are:

  PD 5. map

  PD 5. ped

  PD 5. fam

Now, we can run the preprocessing R® script on PED and FAM files to convert them into CSV format.

3.5 Data preprocessing

One can see the flowchart of the whole analysis process in Figure 3.2. Details of specific steps are given in Figures 3.3 to 3.6.

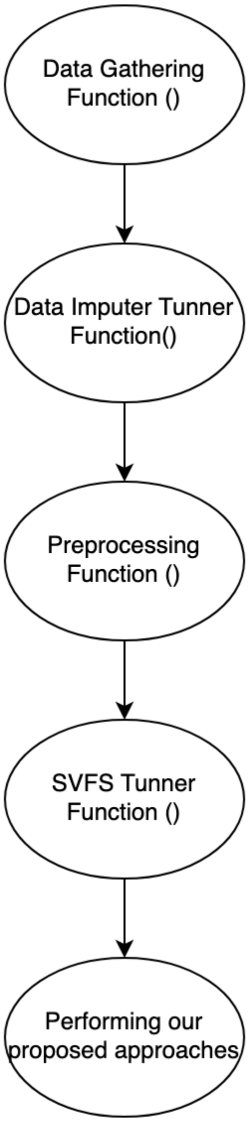

Data Gathering Function ()
Data Imputer Tunner Function()
Preprocessing Function ()
SVFS Tunner Function ()
Performing our proposed approaches

Figure 3.2: Flowchart of whole process

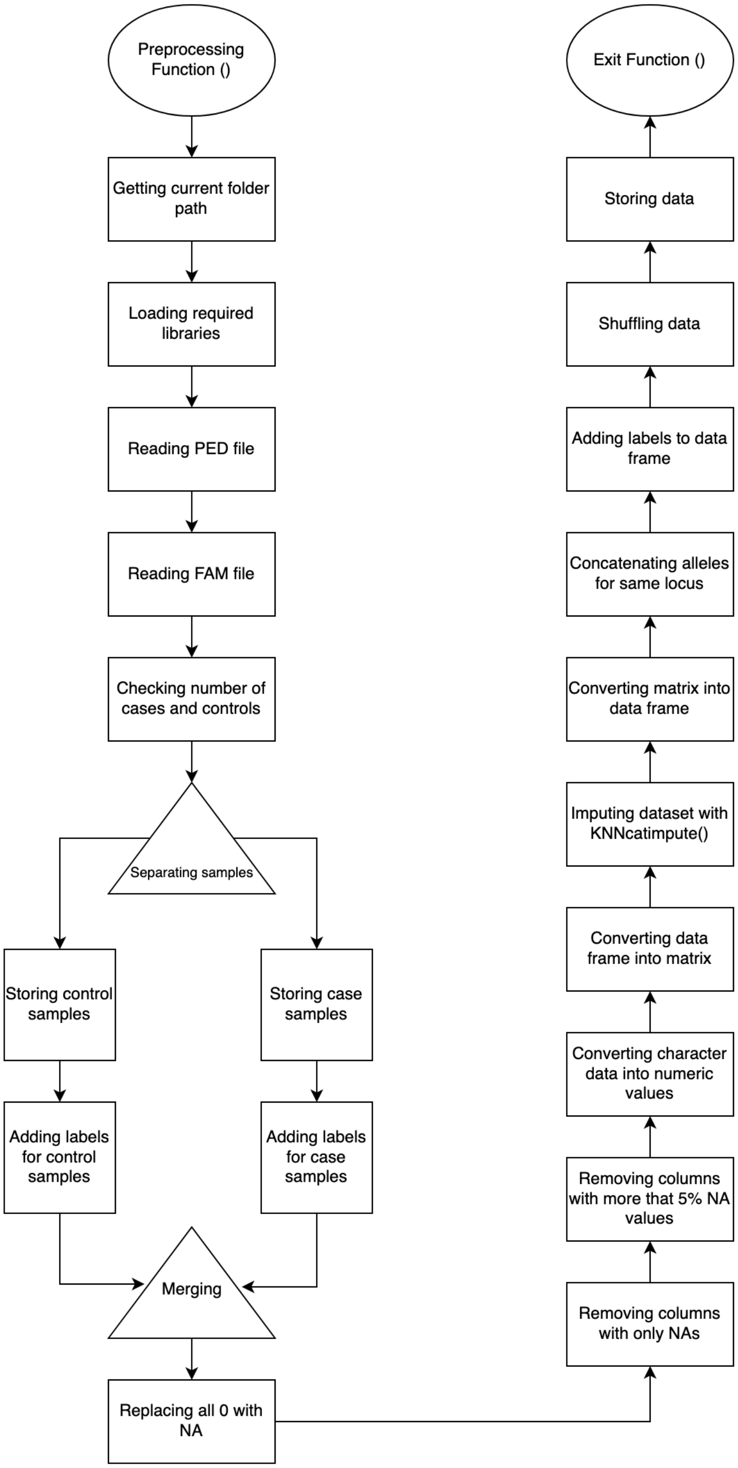

Preprocessing Function ()
Getting current folder path
Loading required libraries
Reading PED file
Reading FAM file
Checking number of cases and controls
Separating samples
Storing control samples
Storing case samples
Adding labels for control samples
Adding labels for case samples
Merging
Replacing all 0 with NA
Removing columns with only NAs
Removing columns with more that 5% NA values
Converting character data into numeric values
Converting data frame into matrix
Imputing dataset with KNNcatimpute()
Converting matrix into data frame
Concatenating alleles for same locus
Adding labels to data frame
Shuffling data
Storing data
Exit Function ()

Figure 3.3: Flowchart of data preprocessing steps (refer to Section 3.5 for more details)

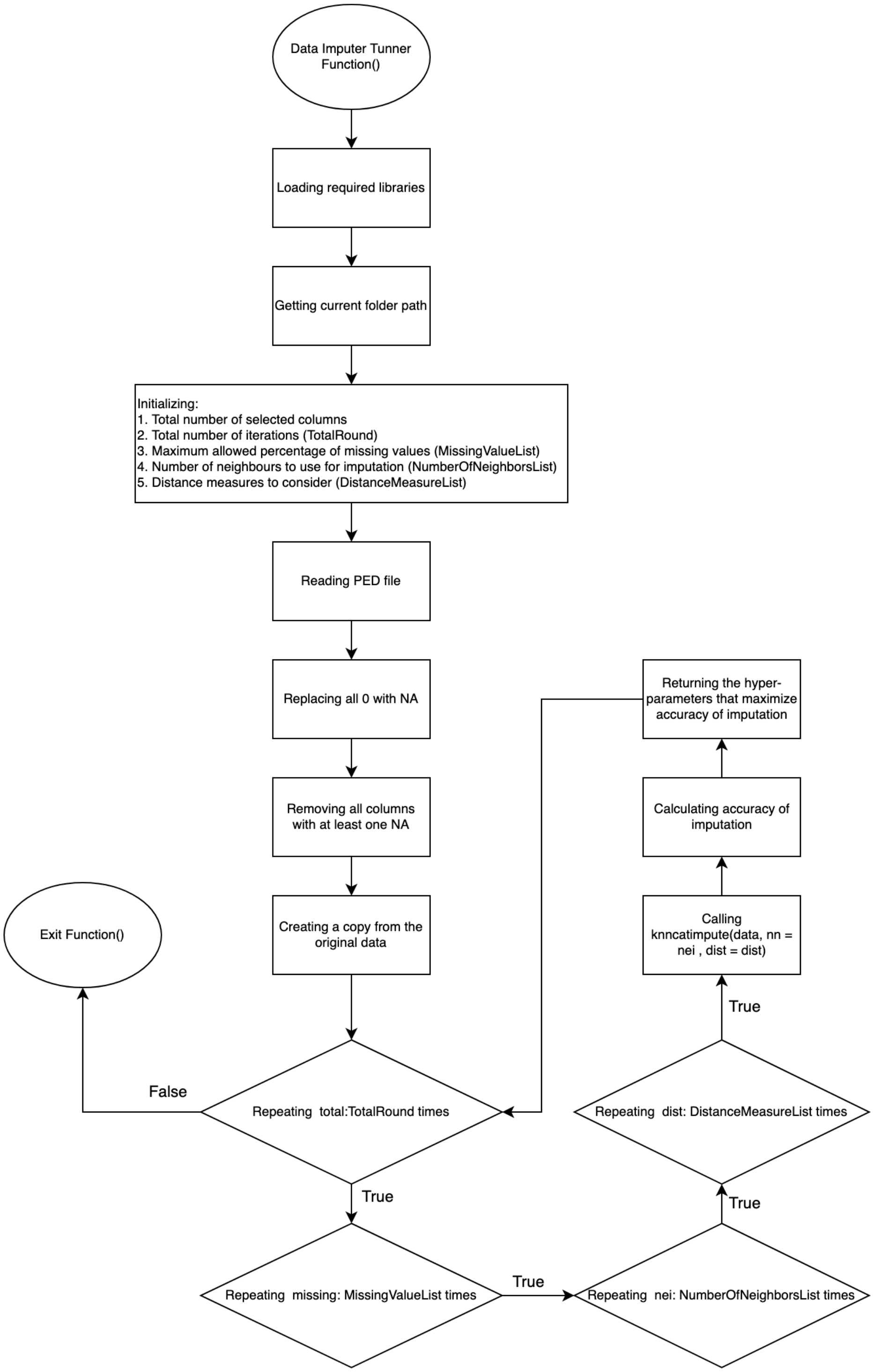

Data Imputer Tunner Function()
Loading required libraries
Getting current folder path
Initializing:
1. Total number of selected columns
2. Total number of iterations (TotalRound)
3. Maximum allowed percentage of missing values (MissingValueList)
4. Number of neighbours to use for imputation (NumberOfNeighborsList)
5. Distance measures to consider (DistanceMeasureList)
Reading PED file
Replacing all 0 with NA
Removing all columns with at least one NA
Creating a copy from the original data
Exit Function()
False
Repeating total:TotalRound times
True
Repeating missing: MissingValueList times
True
Repeating nei: NumberOfNeighborsList times
True
Repeating dist: DistanceMeasureList times
True
Calling knncatimpute(data, nn = nei , dist = dist)
Calculating accuracy of imputation
Returning the hyper-parameters that maximize accuracy of imputation

Figure 3.4: Flowchart of hyper parameter optimization for knncatimpute (refer to Section 3.5.1 for more details)

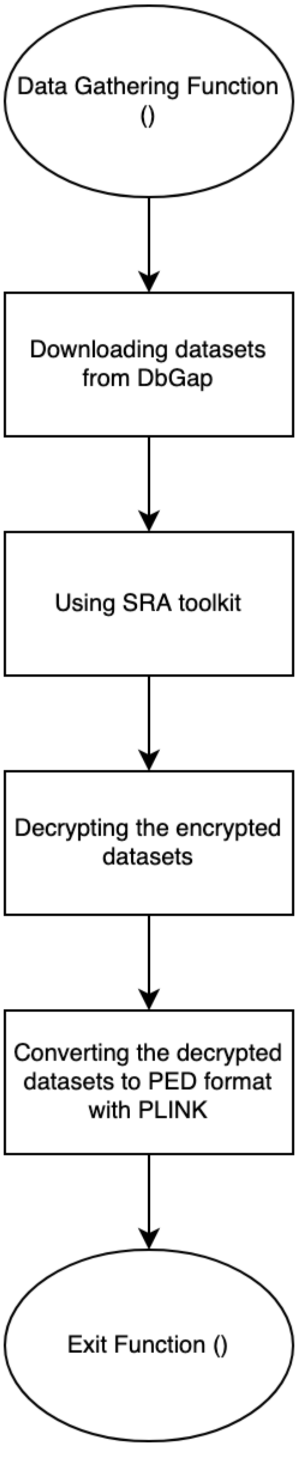

Data Gathering Function ()
Downloading datasets from DbGap
Using SRA toolkit
Decrypting the encrypted datasets
Converting the decrypted datasets to PED format with PLINK
Exit Function ()

Figure 3.5: Flowchart of data gathering steps (refer to Section 3.4 for more details)

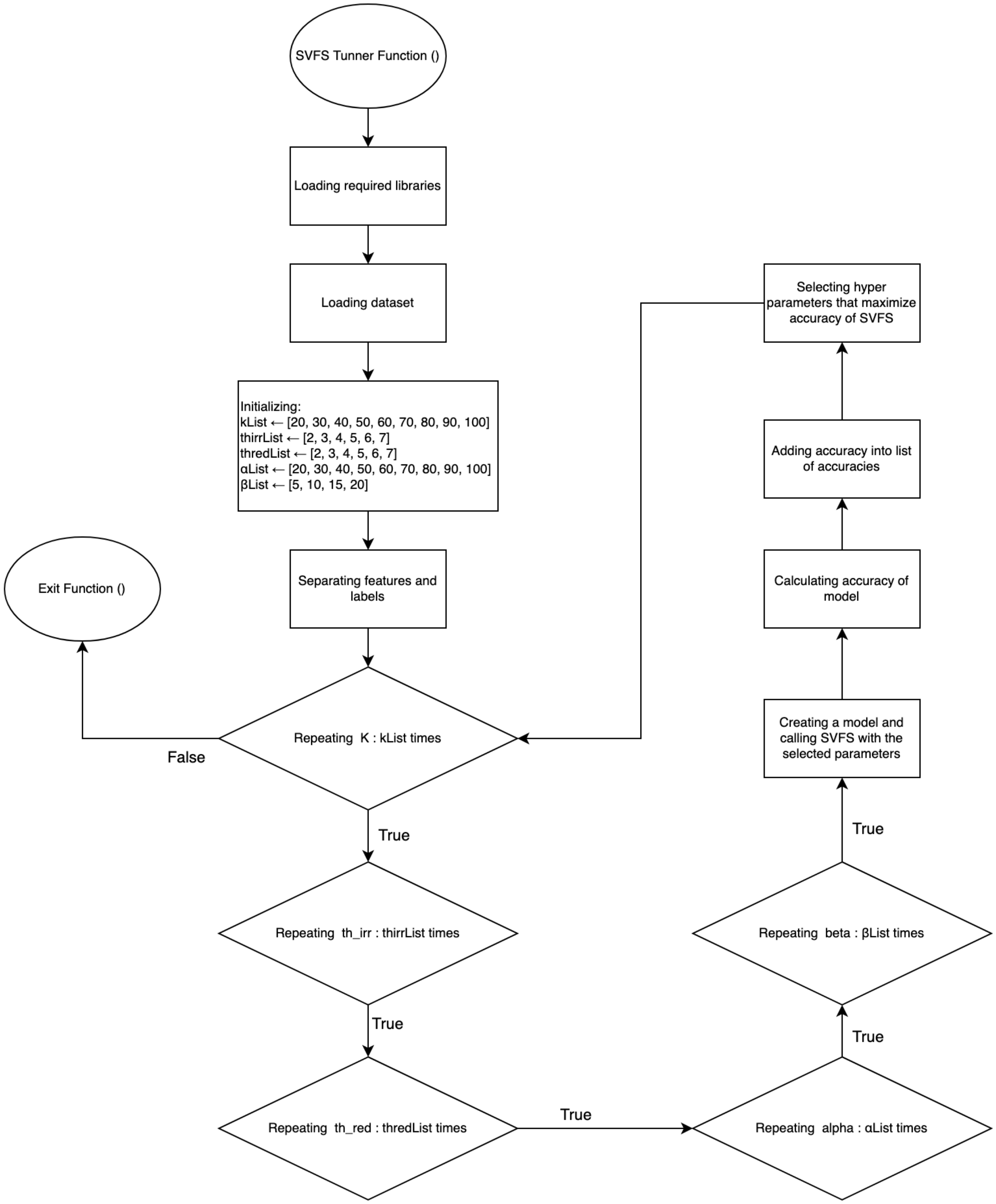

SVFS Tunner Function ()
Loading required libraries
Loading dataset
Initializing:
kList ← [20, 30, 40, 50, 60, 70, 80, 90, 100]
thirrList ← [2, 3, 4, 5, 6, 7]
thredList ← [2, 3, 4, 5, 6, 7]
αList ← [20, 30, 40, 50, 60, 70, 80, 90, 100]
βList ← [5, 10, 15, 20]
Exit Function ()
Separating features and labels
Repeating K : kList times
False
True
Repeating th_irr : thirrList times
True
Repeating th_red : thredList times
True
Repeating alpha : αList times
True
Repeating beta : βList times
True
Creating a model and calling SVFS with the selected parameters
Calculating accuracy of model
Adding accuracy into list of accuracies
Selecting hyper parameters that maximize accuracy of SVFS

Figure 3.6: Flowchart of depicting SVFS tunning steps (refer to Section 3.6.1 for more details)

3.5.1 Tunning data imputer

Because the genomic dataset has lots of NA values, we applied data imputation techniques such as KNNcatimputer [35]. There are two important files for our preprocessing. FAM and PED files. The label of our instances is stored in the Fam file. We used the "Cohen" distance and n = 20 as the parameters for the imputer. The mentioned parameters were all tunned. We used Algorithm 1 for finding the optimal value for each parameter (see also Figure 3.4). We selected 10 random portion of the original datasets. This means that we repeated our algorithm 10 times on 10 different portions. Each portion contained 10,000 features. We defined a list of possible values for percentage of missing values, number of neighbours, and distance measure. Percentage of missing values indicates what percentage of the values in each column is allowed to be NA.

KNNcatimputer is:

```
knncatimpute(x, dist = NA, nn = 3, weights = TRUE)
```

This function accepts four parameters [99] as the input:

- x: a numeric matrix containing missing values. For our project, we passed the datasets that is converted to a numeric matrix to this function.

- dist: a character string naming the distance measure or a distance matrix. This parameter can be "SMC", "PCC", or "Cohen". For our project we considered all these values and the optimized one was "Cohen".

- nn: an integer specifying the number of nearest neighbors used in the imputation of the missing values. We chose nn=20 as this was the optimal value for the datasets.

- weights: should weighted KNN be used to impute the missing values? If true, the vote of each nearest neighbor is weighted by the reciprocal of its distance to the observation or variable when the missing values of this observation or variable, respectively, are replaced.

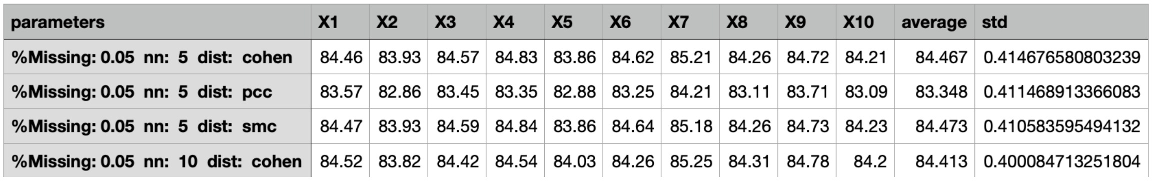
So, for using the mentioned imputer we need below libraries in R®:

```
library(data.table)

library(dplyr)

library(tidyr)

library(scrime)

library(missForest)
```

Algorithm 1 Finding the best parameters for knncatimpute function

```
Require:
  $n \leftarrow 10,000$
  $TotalRound \leftarrow 10$
  $missingValuesList \leftarrow c(0.05, 0.1, 0.15, 0.2)$
  $numberOfNeighborsList \leftarrow c(5, 10, 15, 20)$
  $distanceList \leftarrow c(``cohen", ``pcc", ``smc")$
Ensure: $data \leftarrow knncatimpute(dataset, nn = nei, dist = dist)$
  for roundNo in 1:TotalRound do
    for miss in missingValuesList do
      Generating missing values in the selected partition
      extract the location of NA values, and keep the values of each NA cell
      for nei in numberOfNeighborsList do
        for dist in distanceList do
          $data \leftarrow knncatimpute(data_t, nn = nei, dist = dist)$
          return accuracy with selected parameters
        end for
      end for
    end for
  end for
```

After applying Algorithm 1 on each dataset, we obtained a CSV file with each combination of the parameters values and their corresponding accuracy per round, average accuracy and standard deviation. Figure 3.7 shows a sample of the results obtained.



parameters	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	average	std
%Missing: 0.05 nn: 5 dist: cohen	84.46	83.93	84.57	84.83	83.86	84.62	85.21	84.26	84.72	84.21	84.467	0.414676580803239
%Missing: 0.05 nn: 5 dist: pcc	83.57	82.86	83.45	83.35	82.88	83.25	84.21	83.11	83.71	83.09	83.348	0.411468913366083
%Missing: 0.05 nn: 5 dist: smc	84.47	83.93	84.59	84.84	83.86	84.64	85.18	84.26	84.73	84.23	84.473	0.410583595494132
%Missing: 0.05 nn: 10 dist: cohen	84.52	83.82	84.42	84.54	84.03	84.26	85.25	84.31	84.78	84.2	84.413	0.400084713251804

Figure 3.7: Data imputer accuracy for each combination of the parameters values

We utilized the “Cohen” distance and n = 20 as the imputer's settings across all datasets to ensure a consistent outcome. Additionally, we counted the number of NA values in each dataset and realized that more than 5% of some columns' values were NA. Thus, we eliminated any column with more than 5% NA values. We used the threshold of 5% for missing values because we randomly selected a subset of the dataset with 0% NA values and experimented with different values for the percentage of missing values. The best results were obtained with 5% as the missing values. To ensure this value was stable, we repeated this procedure 10 times.

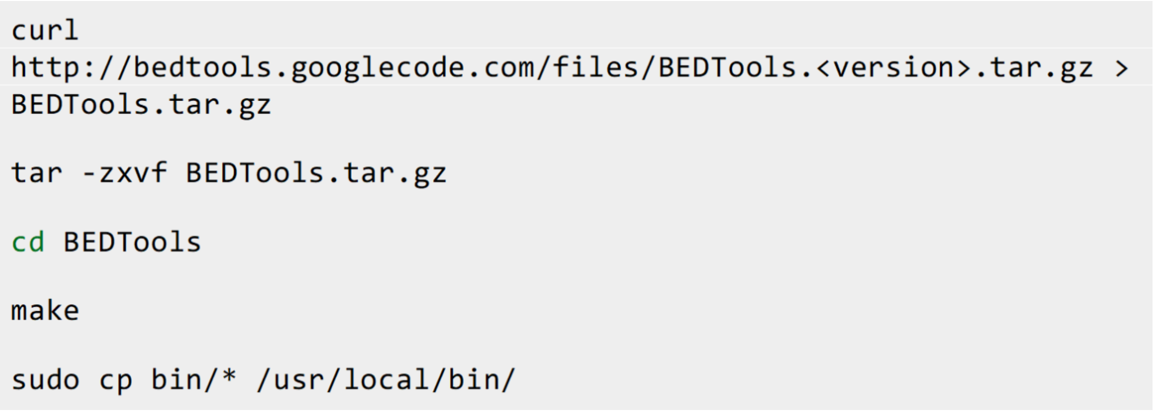
3.5.2 Common SNPs among datasets

In section 3.7, we will need the intersection of the Familial dataset with other datasets. That is why we found the SNPs in common with the Bedtools [100] toolkit.

We get all the information we need from the BIM file. The columns of the BIM file are these:

- Chromosome code (either an integer, or “X”/“Y”/“XY”/“MT”; “0” indicates unknown) or name
- Variant identifier
- Position in morgens or centimorgans (safe to use dummy value of “0”)
- Base-pair coordinate (1-based; limited to 231-2)
- Allele 1 (corresponding to clear bits in. bed; usually minor)
- Allele 2 (corresponding to set bits in. bed; usually major)

For the Bedtools installation in Linux we need to use another toolkit to find the number of common SNPs between the available datasets. One can easily install Bedtools as below:



```
curl
http://bedtools.googlecode.com/files/BEDTools.<version>.tar.gz >
BEDTools.tar.gz

tar -zxvf BEDTools.tar.gz

cd BEDTools

make

sudo cp bin/* /usr/local/bin/
```

Figure 3.8: Bedtools installation

For finding the intersection between SNPs of those datasets, we need to convert the BIM file of each dataset into BED format.

1. 1st column of the BIM file will be the 1st column of the BED file.

2. 4th column of the BIM file -1 will be the 2nd column of the BED file.

3. 4th column of the BIM file will be the 3rd column of the BED file.

4. 2nd column of the BIM file will be the 4th column of the BED file.

For finding the intersection between the datasets, we used this command:

```
bedtools intersect -a SNPs_file1.bed -b SNPs_file2.bed
```

We calculated the common SNPs between Familial dataset and other datasets. Because in some of the approaches described in Section 3.7 we used the intersected version of datasets. As a starting point, we used the Familial dataset and considering it as the base dataset. Considering that it has the most number of instances and features. On the Familial dataset, our model accuracy is also higher.

DatasetID 1	DatasetID 2	Number of SNPs in common
phs000126 (Familial)	phs000394 (Autopsy)	209,730
phs000126 (Familial)	phs000089 (NINDS1)	305,812
phs000126 (Familial)	phs000089 (NINDS2)	4,034
phs000126 (Familial)	phs000048 (Tier1)	28,646

Table 3.6: Number of SNPs in common between Familial and all other four datasets.

3.6 Feature selection

In this step of our project, we used the SVFS algorithm [36] as the feature selection technique. After converting the datasets into CSV format, we can apply the SVFS feature selection algorithm to them. Next we describe briefly how SVFS works. Let $D = [A \mid \mathbf{b}]$ be a labelled dataset, with $\mathbf{b}$ representing the class label and features (attributes) representing columns in matrix A. M. Afshar and H. Usefi showed how the signature matrix $S_A = I - A^{\dagger}A$ (I is the identity matrix and $A^{\dagger}$ is the pseudo-inverse of A) can be used to partition A's columns into clusters, with columns in one cluster correlating exclusively with columns in another cluster. The signature matrix S_D of D is used by SVFS to locate the cluster that holds $\mathbf{b}$. Afshar and Usefi reduced the size of A by eliminating irrelevant features from the other clusters. The signature matrix S_A of reduced A is then used by SVFS to partition the remaining features into clusters and select the most important features from each cluster.

The parameters used for SVFS feature selection algorithm are:

- K = The number of selected features
- Th_{irr} = The threshold set to filter out the irrelevant features
- Th_{red} = Maps the weak feature correlations to zero.
- α = The parameter α is used when facing significant clusters to divide the clusters into sub-clusters with α members.
- β = The parameter β is the number of features selected from each of the sub-clusters with β members.

As a second option, we used the HSIC-Lasso [37], [38] feature selection algorithm with the default parameters [101] (default parameters are B=20 and M=3, B is the block parameter and M is the permutation parameter). However, this feature selection algorithm's performance was worse than SVFS. Thus, it has been confirmed by [36] that SVFS outperformed all other feature selection algorithms. Table 3.7 shows the comparison between the SVFS and HSIC-Lasso feature selection performance on the mentioned datasets.

We used the SVFS implementation available at `https://github.com/Majid1292/SVFS` and the HSIC-Lasso implementation available at `https://github.com/riken-aip/pyHSICLasso`.

Dataset ID	Feature Selection Algorithm	Accuracy±sd
phs000394 (Autopsy)	SVFS	68.09% ± 0.93
	HSIC-Lasso	65.71% ± 0.09
phs000126 (Familial)	SVFS	100% ± 0
	HSIC-Lasso	51.10% ± 0.38
phs000089 (NINDS1)	SVFS	64.10% ± 1.73
	HSIC-Lasso	53.27% ± 0.95
phs000089 (NINDS2)	SVFS	75.72% ± 2.91
	HSIC-Lasso	49.71% ± 0.53
phs000048 (Tier1)	SVFS	51.11% ± 2.38
	HSIC-Lasso	47.46% ± 2.97

Table 3.7: Comparison between SVFS & HSIC-Lasso on baseline approach 0

3.6.1 Tunning SVFS

The SVFS algorithm parameters need to be tunned. We defined five lists for possible values of K, Th_{irr}, Th_{red}, α, and β. The initial values for these parameters were selected according to Majid Afshar and Hamid Usefi's paper [36]. We used Algorithm 2 to find the optimal parameter values from the specified list (see also Figure 3.6).

Algorithm 2 Finding the best parameters for SVFS algorithm

```
Require:
  $kList \leftarrow [20, 30, 40, 50, 60, 70, 80, 90, 100]$
  $thirrList \leftarrow [2, 3, 4, 5, 6, 7]$
  $thredList \leftarrow [2, 3, 4, 5, 6, 7]$
  $\alpha List \leftarrow [20, 30, 40, 50, 60, 70, 80, 90, 100]$
  $\beta List \leftarrow [5, 10, 15, 20]$
Ensure: $fs \leftarrow SVFS(train_x, train_y, irr, 1.7, red, k, \alpha, \beta)$
  for $k : kList$ do
    for $irr : thirrList$ do
      for $red : thredList$ do
        for $alpha : alphaList$ do
          for $beta : betaList$ do
            $fs \leftarrow SVFS(train_x, train_y, irr, 1.7, red, k, alpha, beta)$
            return accuracy with selected parameters
          end for
        end for
      end for
    end for
  end for
```

The best accuracy belonged to:

$K = 50$

$Th_{irr} = 3$

$Th_{red} = 4$

$\alpha = 50$

$\beta = 5$

3.7 Approaches to integrate datasets

In this thesis, we proposed five different approaches to integrate datasets. For each approach, we defined two modes (A and B).

We used the preproccessed version of the supplied datasets, which were described in detail in Section 3.4, for all approaches. We repeated 10 times 5-fold CV on the datasets for each approach. For all methods, we used Random Forests (RF) as the ML classifier with the following settings: n_estimator = 100, criteria = “gini”, max depth = None, and min_samples_split = 2. We also tested Support Vector Machine (SVM) and Gradient Boosting (GB), but they required a lengthy time to process the datasets and performed less accurately than RF. The parameters of RF were tuned. We used the implementation available in scikit-learn (version 1.2) for all the classifiers.

3.7.1 Approach 0

This is our baseline approach. In approach 0, we ran the SVFS feature selection algorithm on each datasets separately to extract the most important features. SVFS feature selection algorithm ran 50 rounds (5-fold CV for 10 times) for each dataset.

In each round, SVFS selected some SNPs. After 50 rounds, we constructed a dictionary of SNP ID with number of times each SNP was selected by SVFS (henceforth referred to as frequency). We chose a threshold for the frequency. We set five as the frequency, selected the SNPs with at least a frequency of 5 and named them the most common SNPs. We chose five as the frequency since, in some datasets, using different thresholds resulted in a sharp decline in the number of common SNPs. Having a small number of common SNPs (features) caused issues when obtaining SNPs in common between dataset in subsequent approaches (see below).

For each dataset we obtained the common SNPs and using CV we assess the classification performance of a model generated using the common SNPs as features and random forest as the classifier. This was part A of this approach.

For part B, we extended the list of most common SNPs by finding SNPs in linkage disequilibirum (LD) with the most common SNPs. LD is when nearby variants are associated within a population more often than if they were unlinked [102]. There are multiple websites to get SNPs in LD. We used SNIPA [103] and Ensemble Rest API [104]. Both websites' results were similar regarding the accuracy of built models. So, we chose SNIPA for its usability.

To get SNPs in LD with a list of SNP IDs, we must define some values for specified parameters. These parameters are genome assembly, variant set, population, and genome annotation. The variant set is the main criteria for comparing the given SNPs list with other SNPs. The population can be chosen from African, American, European, East Asian, and South Asian.

The mentioned parameters for all our approaches were set to GRCH-37, 1000 genomes phase 3 V5, European, and Ensemble 87, respectively. We chose GRCH-37, 1000 genomes phase 3 V5, and Ensemble 87, because these values were the most up to date configuration for our processing. The population for all of 5 datasets is European that is why we chose European. So, we extended each dataset's most common SNPs list with LD and again tried to make a model according to the new list of SNPs and did CV.

3.7.2 Approach 1

In approach 1, we selected the important features from the Familial dataset and performed CV to assess the classification performance of a model generated using each of the other datasets. We ran the SVFS feature selection algorithm on the Familial dataset and extracted the most common SNPs. This time, our features were the selected most common SNPs from the Familial dataset obtained the most common SNPs in common with each of the other datasets and carried out 10-fold CV on each of the other datasets. As with approach 0, we extended the common SNPs with LD and did the same steps in part B.

3.7.3 Approach 2

In approach 2, we first obtained the intersection of SNPs between the Familial dataset and each of the other four datasets. SNPs not in the intersection were removed from the datasets. We did the same steps that have been done for approach 1 by extracting the most common SNPs from the condensed version of the Familial dataset and doing CV on the other datasets. As before, for part B SNPs were extended with LD as well.

3.7.4 Approach 3

In approach 3, we increased the number of instances (individuals) by merging datasets, before doing feature selection. Four merged datasets were created: Familial and Autopsy, Familial and NINDS1, Familial and NINDS2, and Familial and Tier1. We got the SNPs in the intersection between the Familial dataset and the other 4 datasets. We ran the SVFS feature selection algorithm on each of the four merged datasets and extracted the most common SNPs. Then, performed CV to assess the classification performance of a model generated using each of the other datasets.

To allow for a direct comparison with the other approaches, in this approach we calculated the accuracy per each dataset in addition to the accuracy on the merged dataset. To do this, we added another column called datasetID to indicate the original dataset of every instance. In this approach we obtained three accuracies: one overall for the merged dataset and one for the instances of each merged dataset (Table 4.4).

3.7.5 Approach 4

This approach is the same as approach 3, but equal number of instances per dataset were merged. The number of cases and healthy controls taken from each dataset is the same. The cases and healthy controls to include on the merged dataset from the dataset with the higher number of cases and healthy controls were selected randomly.

3.7.6 Compare SNPs ID & gene names between approaches and datasets

We obtained the percentage of SNP IDs and genes in common selected as the most common SNPs among approaches and among datasets.

For getting the genes name associated with a SNP ID we used Biomart platform [105]. See Figures 3.9 and 3.10.

`http://useast.ensembl.org/biomart/martview/0850cfdd1575058045850fddafc913b9`

Once opened the link, set the below configuration:

1. Choose database: "Ensemble Variation 107"

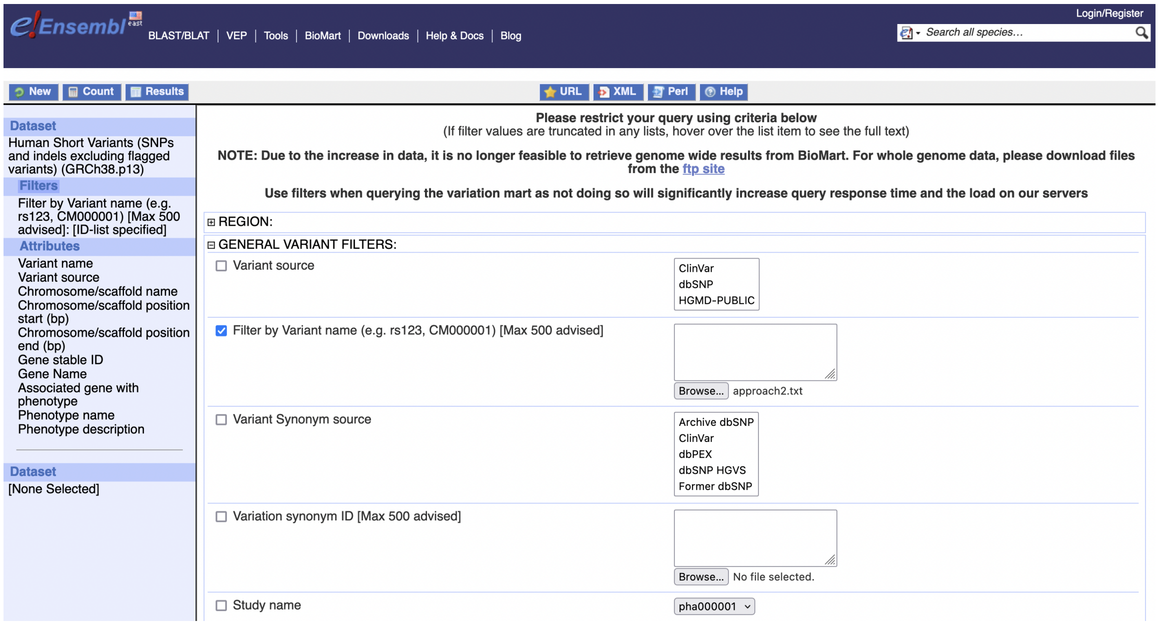
2. Choose dataset: "Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p13)". Then click on attributes

3. "VARIANT ASSOCIATED INFORMATION": In phenotype annotation sec-

tion, select:

- Associated gene with phenotype
- Phenotype name
- Phenotype description

4. Click on “GENE ASSOCIATED INFORMATION” and select:
   - Gene stable ID
   - Gene Name
5. Click on filters. In “GENERAL VARIANT FILTERS”, click on “Filter by Variant name” and pass the SNPs list.
6. Click on “count” and then “results”.
7. Download the result in any format.



Dataset
Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p13)
Filters
Filter by Variant name (e.g. rs123, CM000001) [Max 500 advised]: [ID-list specified]
Attributes
Variant name
Variant source
Chromosome/scaffold name
Chromosome/scaffold position start (bp)
Chromosome/scaffold position end (bp)
Gene stable ID
Gene Name
Associated gene with phenotype
Phenotype name
Phenotype description

Dataset
[None Selected]

Please restrict your query using criteria below
(If filter values are truncated in any lists, hover over the list item to see the full text)

NOTE: Due to the increase in data, it is no longer feasible to retrieve genome wide results from BioMart. For whole genome data, please download files from the ftp site

Use filters when querying the variation mart as not doing so will significantly increase query response time and the load on our servers

⊞ REGION:

⊟ GENERAL VARIANT FILTERS:

□ Variant source — ClinVar / dbSNP / HGMD-PUBLIC

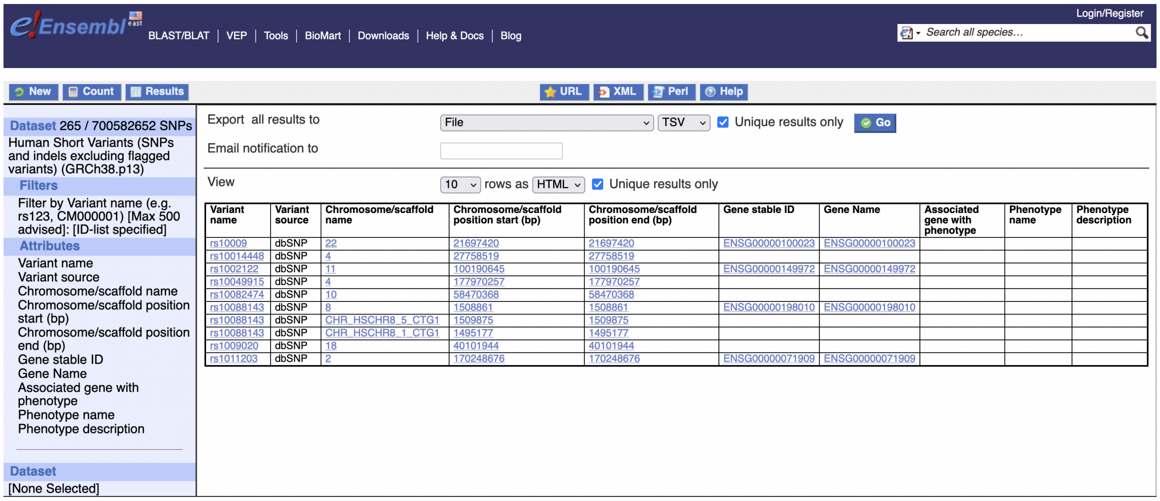
☑ Filter by Variant name (e.g. rs123, CM000001) [Max 500 advised] — Browse... approach2.txt

□ Variant Synonym source — Archive dbSNP / ClinVar / dbPEX / dbSNP HGVS / Former dbSNP

□ Variation synonym ID [Max 500 advised] — Browse... No file selected.

□ Study name — pha000001

Figure 3.9: Biomart input



Dataset 265 / 700582652 SNPs
Human Short Variants (SNPs and indels excluding flagged variants) (GRCh38.p13)
Filters
Filter by Variant name (e.g. rs123, CM000001) [Max 500 advised]: [ID-list specified]
Attributes
Variant name
Variant source
Chromosome/scaffold name
Chromosome/scaffold position start (bp)
Chromosome/scaffold position end (bp)
Gene stable ID
Gene Name
Associated gene with phenotype
Phenotype name
Phenotype description

Dataset
[None Selected]

Export all results to File TSV ☑ Unique results only

Email notification to

View 10 rows as HTML ☑ Unique results only

Variant name	Variant source	Chromosome/scaffold name	Chromosome/scaffold position start (bp)	Chromosome/scaffold position end (bp)	Gene stable ID	Gene Name	Associated gene with phenotype	Phenotype name	Phenotype description
rs10009	dbSNP	22	21697420	21697420	ENSG00000100023	ENSG00000100023			
rs10014448	dbSNP	4	27758519	27758519					
rs1002122	dbSNP	11	100190645	100190645	ENSG00000149972	ENSG00000149972			
rs10049915	dbSNP	4	177970257	177970257					
rs10082474	dbSNP	10	58470368	58470368					
rs10088143	dbSNP	8	1508861	1508861	ENSG00000198010	ENSG00000198010			
rs10088143	dbSNP	CHR_HSCHR8_5_CTG1	1509875	1509875					
rs10088143	dbSNP	CHR_HSCHR8_1_CTG1	1495177	1495177					
rs1009020	dbSNP	18	40101944	40101944					
rs1011203	dbSNP	2	170248676	170248676	ENSG00000071909	ENSG00000071909			

Figure 3.10: Biomart result

3.7.7 Gene-level agreement

Note that when using the Biomart platform for getting the gene names according to SNP IDs, there might be some SNP IDs that do not have a gene name. This will happen when a SNP is an intergenic variant. An intergenic region (IGR) is a stretch of DNA sequences located between genes. Refer to Tables 3.8, 3.9, 3.10, and 3.11 for more details on percentage of intergenic and non-intergenic variants selected as the most frequent SNPs per approach and dataset.

Approach	Percentage of genes identified	Percentage of intergenic variants (between genes)
Approach 0	37.5%	62.5%
Approach 1	34.50%	65.49%
Approach 2	38.70%	61.29%
Approach 3	40.61%	59.38%
Approach 4	37.52%	62.47%

Table 3.8: Percentage of most frequent SNPs located in genes and in intergenic regions for Autopsy dataset

Approach	Percentage of genes identified	Percentage of intergenic variants (between genes)
Approach 0	40.18%	59.81%
Approach 1	37.66%	62.33%
Approach 2	37.41%	62.58%
Approach 3	39.81%	60.18%
Approach 4	42.21%	57.78%

Table 3.9: Percentage of most frequent SNPs located in genes and in intergenic regions for NINDS1 dataset

Approach	Percentage of genes identified	Percentage of intergenic variants (between genes)
Approach 0	39.02%	60.97%
Approach 1	0%	100%
Approach 2	27.77%	72.22%
Approach 3	36.87%	63.12%
Approach 4	46.51%	53.48%

Table 3.10: Percentage of most frequent SNPs located in genes and in intergenic regions for NINDS2 dataset

Approach	Percentage of genes identified	Percentage of intergenic variants (between genes)
Approach 0	36.88%	63.11%
Approach 1	54.54%	45.45%
Approach 2	41.97%	58.02%
Approach 3	37.69%	62.30%
Approach 4	40.24%	59.75%

Table 3.11: Percentage of most frequent SNPs located in genes and in intergenic regions for Tier1 dataset

Chapter 4

Results and Discussion

In this section, we present the CV accuracies for each of the five approaches described in Section 3. Tables 4.1, 4.2, 4.3, 4.4, and 4.5 show the results for approaches per dataset respectively.

The computational time for each approach is around 3 to 5 hours depending on the input dataset's size (the bigger the dataset, the longer the execution time). For all experiments we have available 250 Gb of RAM.

Table 4.1 shows that there are more available SNPs in approach 0 than other approaches. Out of all the approaches, we found that approach 0 had the best accuracy. Parts A and B for all approaches both have similar accuracy levels. Therefore, the LD had little effect on the performance.

The least number of selected SNPs are presented in Approach 1 (Table 4.2). We used the SNPs we extracted from the Familial dataset in Autopsy, NINDS1, NINDS2, and Tier1 datasets in approach 1. Due to the limited availability of SNPs in approach 1, we expected approach 1 to be less accurate.

Approaches 1 and 2 had comparable performance (Tables 4.2, 4.3) with average accuracy (Autopsy = 64.89%, NINDS1 = 51.98%, NINDS2 = 53.04%, and Tier1 = 44.47%) and (Autopsy = 65.40%, NINDS1 = 52.56%, NINDS2 = 50.38%, and Tier1 = 31.50%), respectively. The trend between approaches 3 and 4 is the same, but this trend performs better (Tables 4.4 and 4.5) than that of Approach 1 and 2. In all below tables, number of available SNPs means number of SNPs included in to the approach.

In general, we recommend other researchers to use Approaches 3 and 4 in studies similar to ours. In the mentioned two approaches, we increased the number of samples. As a result, we achieved higher classification performance in terms of accuracy. Our approaches can be applied to other SNPs data of different disease like breast cancer, lung cancer and etc. The input data should be in tabular format along with the SNPs ID as the columns.

The results are dependent on the selected features. If one selects other features, the accuracy will be changed. Note that, the selected features are the features that are common among at least 2 approaches for a consistent result.

In all tables, by the number of Available SNPs we mean number of SNPs included in to the corresponding approach.

Approach		Dataset	Number of Available SNPs	Number of Samples	Accuracy±sd
0	A. Same dataset (Most Common SNPs)	Autopsy	404	Total: 977 Case (2): 642 Control (1): 335	70.83% ± 1.74
		NINDS1	1723	Total: 1741 Case (2): 940 Control (1): 801	77.14% ± 2.12
		NINDS2	792	Total: 526 Case (2): 263 Control (1): 263	87.46% ± 3.30
		Tier1	218	Total: 886 Case (2): 443 Control (1): 443	51.92% ± 3.12
	B. Same dataset (Most Common SNPs + LD)	Autopsy	507	Total: 977 Case (2): 642 Control (1): 335	70.11% ± 1.69
		NINDS1	2575	Total: 1741 Case (2): 940 Control (1): 801	75.88% ± 1.52
		NINDS2	988	Total: 526 Case (2): 263 Control (1): 263	84.41% ± 0.49
		Tier1	422	Total: 886 Case (2): 443 Control (1): 443	48.31% ± 1.35

Table 4.1: Approach 0

Approach		Dataset	Number of Available SNPs	Number of Samples	Accuracy±sd
1	A. Familial dataset (Most Common SNPs)	Autopsy	143	Total: 977 Case (2): 642 Control (1): 335	64.89% ± 0.20
		NINDS1	238	Total: 1741 Case (2): 940 Control (1): 801	51.98% ± 2.87
		NINDS2	1	Total: 526 Case (2): 263 Control (1): 263	53.04% ± 4.83
		Tier1	23	Total: 886 Case (2): 443 Control (1): 443	44.47% ± 2.13
	B. Familial dataset (Most Common SNPs + LD)	Autopsy	190	Total: 977 Case (2): 642 Control (1): 335	64.28% ± 0.56
		NINDS1	414	Total: 1741 Case (2): 940 Control (1): 801	51.69% ± 1.83
		NINDS2	97	Total: 526 Case (2): 263 Control (1): 263	48.68% ± 7.33
		Tier1	148	Total: 886 Case (2): 443 Control (1): 443	33.53% ± 4.89

Table 4.2: Approach 1

Approach		Dataset	Number of Available SNPs	Number of Samples	Accuracy±sd
2	A. Intersection (Most Common SNPs)	$Familial \cap Autopsy$	278	Total: 977 Case (2): 642 Control (1): 335	65.40% ± 0.47
		$Familial \cap NINDS1$	290	Total: 1741 Case (2): 940 Control (1): 801	52.56% ± 2.02
		$Familial \cap NINDS2$	32	Total: 526 Case (2): 263 Control (1): 263	50.38% ± 2.86
		$Familial \cap Tier1$	317	Total: 886 Case (2): 443 Control (1): 443	31.50% ± 7.31
	B. Intersection (Most Common SNPs + LD)	$Familial \cap Autopsy$	318	Total: 977 Case (2): 642 Control (1): 335	65.61% ± 0.79
		$Familial \cap NINDS1$	388	Total: 1741 Case (2): 940 Control (1): 801	50.66% ± 2.50
		$Familial \cap NINDS2$	32	Total: 526 Case (2): 263 Control (1): 263	50.94% ± 3.50
		$Familial \cap Tier1$	327	Total: 886 Case (2): 443 Control (1): 443	32.28% ± 2.25

Table 4.3: Approach 2

Approach		Dataset	Number of Available SNPs	Number of Samples		*Accuracy* ± *sd*	
3	A. The intersection of SNPs & Union of the Individual (Most Common SNPs)	$\cup(Familial \cap Autopsy)$	452	Total: 2744 Case (2): 1542 Control (1): 1202	Familial: 1767 Case (2): 900 Control (1): 867	87.86% ± 0.68	Familial: 100% ± 0
					Autopsy: 977 Case (2): 642 Control (1): 335		Autopsy: 65.94% ± 1.3
		$\cup(Familial \cap NINDS1)$	858	Total: 3508 Case (2): 1840 Control (1): 1668	Familial: 1767 Case (2): 900 Control (1): 867	80.67% ± 0.99	Familial: 100% ± 0
					NINDS1: 1741 Case (2): 940 Control (1): 801		NINDS1: 61.06% ± 1.85
		$\cup(Familial \cap NINDS2)$	164	Total: 2293 Case (2): 1163 Control (1): 1130	Familial: 1767 Case (2): 900 Control (1): 867	90.27% ± 0.33	Familial: 99.72% ± 0.18
					NINDS2: 526 Case (2): 263 Control (1): 263		NINDS2: 58.56% ± 0.55
		$\cup(Familial \cap Tier1)$	196	Total: 2653 Case (2): 1343 Control (1): 1310	Familial: 1767 Case (2): 900 Control (1): 867	80.25% ± 2.04	Familial: 100% ± 0
					Tier1: 886 Case (2): 443 Control (1): 443		Tier1: 40.88% ± 5.05
	B. The intersection of SNPs & Union of Individual (Most Common SNPs + LD)	$\cup(Familial \cap Autopsy)$	570	Total: 2744 Case (2): 1542 Control (1): 1202	Familial: 1767 Case (2): 900 Control (1): 867	87.75% ± 0.39	Familial: 100% ± 0
					Autopsy: 977 Case (2): 642 Control (1): 335		Autopsy: 65.43% ± 3.06
		$\cup(Familial \cap NINDS1)$	1184	Total: 3508 Case (2): 1840 Control (1): 1668	Familial: 1767 Case (2): 900 Control (1): 867	80.10% ± 1.97	Familial: 100% ± 0
					NINDS1: 1741 Case (2): 940 Control (1): 801		NINDS1: 59.97% ± 3.10
		$\cup(Familial \cap NINDS2)$	165	Total: 2293 Case (2): 1163 Control (1): 1130	Familial: 1767 Case (2): 900 Control (1): 867	89.88% ± 1.38	Familial: 99.72% ± 0.17
					NINDS2: 526 Case (2): 263 Control (1): 263		NINDS2: 57.02% ± 3.22

		$\cup(Familial \cap Tier1)$	212	Total: 2653
				Case (2): 1343
				Control (1): 1310
					Familial: 1767
					Case (2):
					Control (1):
						81.27% ± 1.03	Familial: 100% ± 0

					Tier1: 886 Case (2): 443 Control (1): 443		Tier1: 43.93% ± 1.74%

Table 4.4: Approach 3

Approach		Dataset	Number of Available SNPs	Number of Samples		$Accuracy \pm sd$	
4	A. The intersection of SNPs & Union of the Individual (Balanced Data & Most Common SNPs)	$\cup(Familial \cap Autopsy)$	525	Total: 1954 Case (2): 1284 Control (1): 670	Familial: 977 Case (2): 642 Control (1): 335	83.26% ± 1.79	Familial: 100% ± 0
					Autopsy: 977 Case (2): 642 Control (1): 335		Autopsy: 66.56% ± 3.16
		$\cup(Familial \cap NINDS1)$	806	Total: 3402 Case (2): 1800 Control (1): 1602	Familial: 1701 Case (2): 900 Control (1): 801	79.89% ± 1.24	Familial: 100% ± 0
					NINDS1: 1701 Case (2): 900 Control (1): 801		NINDS1: 59.79% ± 2.31
		$\cup(Familial \cap NINDS2)$	125	Total: 1052 Case (2): 526 Control (1): 526	Familial: 526 Case (2): 263 Control (1): 263	81.56% ± 1.01	Familial: 100% ± 0
					NINDS2: 526 Case (2): 263 Control (1): 263		NINDS2: 63.02% ± 2.36
		$\cup(Familial \cap Tier1)$	168	Total: 1772 Case (2): 886 Control (1): 886	Familial: 886 Case (2): 443 Control (1): 443	71.61% ± 2.04	Familial: 100% ± 0
					Tier1: 886 Case (2): 443 Control (1): 443		Tier1: 43.20% ± 3.36
	B. The intersection of SNPs & Union of Individual (Balanced Data & Most Common SNPs + LD)	$\cup(Familial \cap Autopsy)$	676	Total: 1954 Case (2): 1284 Control (1): 670	Familial: 977 Case (2): 642 Control (1): 335	82.81% ± 2.29	Familial: 100% ± 0
					Autopsy: 977 Case (2): 642 Control (1): 335		Autopsy: 65.57% ± 4.60
		$\cup(Familial \cap NINDS1)$	1123	Total: 3402 Case (2): 1800 Control (1): 1602	Familial: 1701 Case (2): 900 Control (1): 801	79.78% ± 0.87	Familial: 100% ± 0
					NINDS1: 1701 Case (2): 900 Control (1): 801		NINDS1: 59.47% ± 2.89
		$\cup(Familial \cap NINDS2)$	127	Total: 1052 Case (2): 526 Control (1): 526	Familial: 526 Case (2): 263 Control (1): 263	81.18% ± 1.52	Familial: 100% ± 0
					NINDS2: 526 Case (2): 263 Control (1): 263		NINDS2: 62.35% ± 2.88
		$\cup(Familial \cap Tier1)$	184	Total: 1772 Case (2): 886 Control (1): 886	Familial: 886 Case (2): 443 Control (1): 443	73.48% ± 2.04	Familial: 100% ± 0
					Tier1: 886 Case (2): 443 Control (1): 443		Tier1: 46.86% ± 3.90

Table 4.5: Approach 4

4.1 Comparison between approaches & datasets

In the previous section, we have seen the results per approach. In this section, we aggregate and summarize the results for further comparison between approaches and datasets. We achieved the highest accuracy for all datasets with approach 0 (Table 4.6). We anticipated that approach 0 achieves the highest accuracy, because we train and assess performance on the same datasets. Approaches 1 & 2,and approaches 3 & 4 have similar accuracy. Approaches 3 and 4 achieved higher accuracy than approaches 1 and 2 for Autopsy, NINDS1, and NINDS2. As we have discussed before extending the list of most common SNPs by LD did not improve accuracy (Table 4.7).

Approach	$Autopsy\ Accuracy \pm sd$	$NINDS1\ Accuracy \pm sd$	$NINDS2\ Accuracy \pm sd$	$Tier1\ Accuracy \pm sd$
Approach 0 (A)	70.83% ± 1.74	77.14% ± 2.12	87.46% ± 3.30	51.92% ± 3.12
Approach 1 (A)	64.89% ± 0.20	51.98% ± 2.87	53.04% ± 4.83	44.47% ± 2.13
Approach 2 (A)	65.40% ± 0.47	52.56% ± 2.02	50.38% ± 2.86	31.50% ± 7.31
Approach 3 (A)	65.94% ± 1.30	61.06% ± 1.85	58.56% ± 0.55	40.88% ± 5.05
Approach 4 (A)	66.56% ± 3.16	59.79% ± 2.31	63.02% ± 2.36	43.20% ± 3.36

Table 4.6: Comparison of datasets accuracy for part A per approach

Approach	$Autopsy\ Accuracy \pm sd$	$NINDS1\ Accuracy \pm sd$	$NINDS2\ Accuracy \pm sd$	$Tier1\ Accuracy \pm sd$
Approach 0 (B)	70.11% ± 1.69	75.88% ± 1.52	84.41% ± 0.49	48.31% ± 1.35
Approach 1 (B)	64.28% ± 0.56	51.69% ± 1.83	48.68% ± 7.33	33.53% ± 4.89
Approach 2 (B)	65.61% ± 0.79	50.66% ± 2.50	50.94% ± 3.50	32.28% ± 2.25
Approach 3 (B)	65.43% ± 3.06	59.97% ± 3.10	57.02% ± 3.22	43.93% ± 1.74
Approach 4 (B)	65.57% ± 4.60	59.47% ± 2.89	62.35% ± 2.88	46.86% ± 3.90

Table 4.7: Comparison of datasets accuracy for part B per approach

Tables 4.8, 4.9, 4.10, and 4.11 show the percentage of SNPs and genes in common between approaches. There are a small number of SNPs and genes that are common between approach 0 and other approaches. This is expected because, in approach 0 we selected the most common SNPs from each datasets separately. However, in approach 1 and 2 we selected the most common SNPs from Familial dataset and in approaches 3 and 4 from a merged dataset that included Familial. As a result, they have more SNPs and genes in common.

As the features available vary between approaches and datasets, the features selected also vary and thus the accuracy achieved also varies. Thus our results provide a range of the expected accuracy that can be achieved when using GWAS data to identify SNPs associated with PD. We expect that a SNP strongly associated with PD would be selected multiple times and thus all those SNPs identified by more than one approach or in more than one dataset have stronger support to be a PD biomarker and should be prioritized for further investigation. Often there are multiple SNPs in a single gene, so considering whether a gene is identified multiple times via different SNPs might be useful to understand which genes might be involved in the development of PD. In the Table below, we see that indeed the gene-level agreement is higher than the SNP-level agreement between approaches.

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Approach0$ $\beta = Approach1$	SNPs ID: 0.24% Gene Name: 1.48%	SNPs ID: 0.69% Gene Name: 4.08%	SNPs ID: 0.18% Gene Name: 1.09%
$\alpha = Approach0$ $\beta = Approach2$	SNPs ID: 0.24% Gene Name: 2.96%	SNPs ID: 0.35% Gene Name: 3.73%	SNPs ID: 0.14% Gene Name: 1.68%
$\alpha = Approach0$ $\beta = Approach3$	SNPs ID: 2.47% Gene Name: 9.62%	SNPs ID: 2.21% Gene Name: 7.22%	SNPs ID: 1.18% Gene Name: 4.30%
$\alpha = Approach0$ $\beta = Approach4$	SNPs ID: 8.41% Gene Name: 13.33%	SNPs ID: 6.47% Gene Name: 9%	SNPs ID: 3.79% Gene Name: 5.67%
$\alpha = Approach1$ $\beta = Approach2$	SNPs ID: 72.02% Gene Name: 79.59%	SNPs ID: 37.05% Gene Name: 36.44%	SNPs ID: 32.38% Gene Name: 33.33%
$\alpha = Approach1$ $\beta = Approach3$	SNPs ID: 35.66% Gene Name: 34.69%	SNPs ID: 11.28% Gene Name: 9.44%	SNPs ID: 9.37% Gene Name: 8.01%
$\alpha = Approach1$ $\beta = Approach4$	SNPs ID: 32.86% Gene Name: 32.65%	SNPs ID: 8.95% Gene Name: 8%	SNPs ID: 7.56% Gene Name: 6.86%
$\alpha = Approach2$ $\beta = Approach3$	SNPs ID: 31.65% Gene Name: 36.44%	SNPs ID: 19.46% Gene Name: 21.66%	SNPs ID: 13.70% Gene Name: 15.72%
$\alpha = Approach2$ $\beta = Approach4$	SNPs ID: 31.65% Gene Name: 36.44%	SNPs ID: 16.76% Gene Name: 19.5%	SNPs ID: 12.30% Gene Name: 14.55%
$\alpha = Approach3$ $\beta = Approach4$	SNPs ID: 74.55% Gene Name: 71.11%	SNPs ID: 64.19% Gene Name: 64%	SNPs ID: 52.65% Gene Name: 50.79%

Table 4.8: Percentage of common SNPs & genes between approaches for Autopsy

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Approach0$ $\beta = Approach1$	SNPs ID: 0.05% Gene Name: 2.43%	SNPs ID: 0.42% Gene Name: 13.63%	SNPs ID: 0.05% Gene Name: 2.11%
$\alpha = Approach0$ $\beta = Approach2$	SNPs ID: 0.05% Gene Name: 3.86%	SNPs ID: 0.34% Gene Name: 17.11%	SNPs ID: 0.04% Gene Name: 3.25%
$\alpha = Approach0$ $\beta = Approach3$	SNPs ID: 7.77% Gene Name: 14.63%	SNPs ID: 15.61% Gene Name: 23.52%	SNPs ID: 5.47% Gene Name: 9.91%
$\alpha = Approach0$ $\beta = Approach4$	SNPs ID: 7.60% Gene Name: 16.66%	SNPs ID: 16.25% Gene Name: 26.62%	SNPs ID: 5.46% Gene Name: 11.42%
$\alpha = Approach1$ $\beta = Approach2$	SNPs ID: 90.75% Gene Name: 90.90%	SNPs ID: 74.48% Gene Name: 72.07%	SNPs ID: 69.23% Gene Name: 67.22%
$\alpha = Approach1$ $\beta = Approach3$	SNPs ID: 23.10% Gene Name: 36.36%	SNPs ID: 6.41% Gene Name: 10.45%	SNPs ID: 5.28% Gene Name: 8.83%
$\alpha = Approach1$ $\beta = Approach4$	SNPs ID: 22.26% Gene Name: 32.95%	SNPs ID: 6.57% Gene Name: 9.41%	SNPs ID: 5.34% Gene Name: 7.90%
$\alpha = Approach2$ $\beta = Approach3$	SNPs ID: 20% Gene Name: 31.53%	SNPs ID: 6.75% Gene Name: 11.43%	SNPs ID: 5.32% Gene Name: 9.16%
$\alpha = Approach2$ $\beta = Approach4$	SNPs ID: 20% Gene Name: 33.33%	SNPs ID: 7.19% Gene Name: 12.01%	SNPs ID: 5.58% Gene Name: 9.68%
$\alpha = Approach3$ $\beta = Approach4$	SNPs ID: 68.64% Gene Name: 75.16%	SNPs ID: 73.07% Gene Name: 74.67%	SNPs ID: 54.79% Gene Name: 59.89%

Table 4.9: Percentage of common SNPs & genes between approaches for NINDS1

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Approach0$ $\beta = Approach1$	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%
$\alpha = Approach0$ $\beta = Approach2$	SNPs ID: 0% Gene Name: 0.38%	SNPs ID: 0% Gene Name: 9.09%	SNPs ID: 0% Gene Name: 0.37%
$\alpha = Approach0$ $\beta = Approach3$	SNPs ID: 0.12% Gene Name: 2.31%	SNPs ID: 0.60% Gene Name: 10.34%	SNPs ID: 0.10% Gene Name: 1.92%
$\alpha = Approach0$ $\beta = Approach4$	SNPs ID: 0.37% Gene Name: 2.31%	SNPs ID: 2.4% Gene Name: 9.83%	SNPs ID: 0.32% Gene Name: 1.91%
$\alpha = Approach1$ $\beta = Approach2$	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%
$\alpha = Approach1$ $\beta = Approach3$	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%
$\alpha = Approach1$ $\beta = Approach4$	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%
$\alpha = Approach2$ $\beta = Approach3$	SNPs ID: 15.62% Gene Name: 27.27%	SNPs ID: 3.04% Gene Name: 5.17%	SNPs ID: 2.61% Gene Name: 4.54%
$\alpha = Approach2$ $\beta = Approach4$	SNPs ID: 15.62% Gene Name: 36.36%	SNPs ID:4% Gene Name: 6.55%	SNPs ID: 3.28% Gene Name: 5.88%
$\alpha = Approach3$ $\beta = Approach4$	SNPs ID: 15.85% Gene Name: 29.31%	SNPs ID: 20.8% Gene Name: 27.86%	SNPs ID: 9.88% Gene Name: 16.66%

Table 4.10: Percentage of common SNPs & genes between approaches for NINDS2

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Approach0$ $\beta = Approach1$	SNPs ID: 0% Gene Name: 2.43%	SNPs ID: 0% Gene Name: 15.38%	SNPs ID: 0% Gene Name: 2.15%
$\alpha = Approach0$ $\beta = Approach2$	SNPs ID: 0% Gene Name: 4.87%	SNPs ID: 0% Gene Name: 2.98%	SNPs ID: 0% Gene Name: 1.88%
$\alpha = Approach0$ $\beta = Approach3$	SNPs ID: 3.21% Gene Name: 3.65%	SNPs ID: 3.57% Gene Name: 4.10%	SNPs ID: 1.71% Gene Name: 1.97%
$\alpha = Approach0$ $\beta = Approach4$	SNPs ID: 1.83% Gene Name: 6.09%	SNPs ID: 2.38% Gene Name: 7.46%	SNPs ID: 1.04% Gene Name: 3.47%
$\alpha = Approach1$ $\beta = Approach2$	SNPs ID: 86.95% Gene Name: 92.30%	SNPs ID: 6.30% Gene Name: 8.95%	SNPs ID: 6.25% Gene Name: 8.88%
$\alpha = Approach1$ $\beta = Approach3$	SNPs ID: 30.43% Gene Name: 38.46%	SNPs ID: 3.57% Gene Name: 6.84%	SNPs ID: 3.30% Gene Name: 6.17%
$\alpha = Approach1$ $\beta = Approach4$	SNPs ID: 13.04% Gene Name: 23.07%	SNPs ID: 1.78% Gene Name: 4.47%	SNPs ID: 1.59% Gene Name: 3.89%
$\alpha = Approach2$ $\beta = Approach3$	SNPs ID: 26.49% Gene Name: 29.10%	SNPs ID: 42.85% Gene Name: 53.42%	SNPs ID: 19.58% Gene Name: 23.21%
$\alpha = Approach2$ $\beta = Approach4$	SNPs ID: 19.55% Gene Name: 23.13%	SNPs ID:36.90% Gene Name: 46.26%	SNPs ID: 14.65% Gene Name: 18.23%
$\alpha = Approach3$ $\beta = Approach4$	SNPs ID: 35.71% Gene Name: 41.09%	SNPs ID: 41.66% Gene Name: 44.77%	SNPs ID: 23.80% Gene Name: 27.27%

Table 4.11: Percentage of common SNPs & genes between approaches for Tier1

Tables 4.12, 4.13, 4.14, 4.15,and 4.16 show the comparison of the percentage of SNP IDs and gene names common among different datasets for approach 0, approach 1, approach 2, approach 3, and approach 4 respectively.

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Autopsy$ $\beta = NINDS1$	SNPs ID: 0.49% Gene Name: 8.88%	SNPs ID: 0.11% Gene Name: 2.43%	SNPs ID: 0.09% Gene Name: 1.95%
$\alpha = Autopsy$ $\beta = NINDS2$	SNPs ID: 0% Gene Name: 9.62%	SNPs ID: 0% Gene Name: 5.01%	SNPs ID: 0% Gene Name: 3.41%
$\alpha = Autopsy$ $\beta = Tier1$	SNPs ID: 0% Gene Name: 1.48%	SNPs ID: 0% Gene Name: 2.43%	SNPs ID: 0% Gene Name: 0.93%
$\alpha = NINDS1$ $\beta = NINDS2$	SNPs ID: 4.41% Gene Name: 14.43%	SNPs ID: 9.59% Gene Name: 27.41%	SNPs ID: 3.11% Gene Name: 10.44%
$\alpha = NINDS1$ $\beta = Tier1$	SNPs ID: 0% Gene Name: 2.23%	SNPs ID: 0% Gene Name: 13.41%	SNPs ID: 0% Gene Name: 1.95%
$\alpha = NINDS2$ $\beta = Tier1$	SNPs ID: 0% Gene Name: 2.31%	SNPs ID: 0% Gene Name: 7.31%	SNPs ID: 0% Gene Name: 1.79%

Table 4.12: Percentage of common SNPs & genes between datasets for approach 0

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Autopsy$ $\beta = NINDS1$	SNPs ID: 92.30% Gene Name: 95.91%	SNPs ID: 55.46% Gene Name: 53.40%	SNPs ID: 53.01% Gene Name: 52.22%
$\alpha = Autopsy$ $\beta = NINDS2$	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%
$\alpha = Autopsy$ $\beta = Tier1$	SNPs ID: 9.79% Gene Name: 12.24%	SNPs ID: 60.86% Gene Name: 46.15%	SNPs ID: 9.21% Gene Name: 10.71%
$\alpha = NINDS1$ $\beta = NINDS2$	SNPs ID: 0.42% Gene Name: 0%	SNPs ID: 100% Gene Name: 0%	SNPs ID: 0.42% Gene Name: 0%
$\alpha = NINDS1$ $\beta = Tier1$	SNPs ID: 9.24% Gene Name: 14.77%	SNPs ID: 95.65% Gene Name: 100%	SNPs ID: 9.20% Gene Name: 14.77%
$\alpha = NINDS2$ $\beta = Tier1$	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%	SNPs ID: 0% Gene Name: 0%

Table 4.13: Percentage of common SNPs & genes between datasets for approach 1

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Autopsy$ $\beta = NINDS1$	SNPs ID: 38.84% Gene Name: 42.99%	SNPs ID: 37.24% Gene Name: 41.44%	SNPs ID: 23.47% Gene Name: 26.74%
$\alpha = Autopsy$ $\beta = NINDS2$	SNPs ID: 0% Gene Name: 1.86%	SNPs ID: 0% Gene Name: 18.18%	SNPs ID: 0% Gene Name: 1.72%
$\alpha = Autopsy$ $\beta = Tier1$	SNPs ID: 10.43% Gene Name: 14.01%	SNPs ID: 9.14% Gene Name: 11.19%	SNPs ID: 5.12% Gene Name: 6.63%
$\alpha = NINDS1$ $\beta = NINDS2$	SNPs ID: 0% Gene Name: 1.80%	SNPs ID: 0% Gene Name: 18.18%	SNPs ID: 0% Gene Name: 1.66%
$\alpha = NINDS1$ $\beta = Tier1$	SNPs ID: 7.58% Gene Name: 16.21%	SNPs ID: 6.94% Gene Name: 13.43%	SNPs ID: 3.76% Gene Name: 7.92%
$\alpha = NINDS2$ $\beta = Tier1$	SNPs ID: 3.12% Gene Name: 18.18%	SNPs ID: 0.31% Gene Name: 1.49%	SNPs ID: 0.28% Gene Name: 1.39%

Table 4.14: Percentage of common SNPs & genes between datasets for approach 2

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Autopsy$ $\beta = NINDS1$	SNPs ID: 15.26% Gene Name: 22.22%	SNPs ID: 8.04% Gene Name: 13.07%	SNPs ID: 5.56% Gene Name: 8.96%
$\alpha = Autopsy$ $\beta = NINDS2$	SNPs ID: 0.66% Gene Name: 3.33%	SNPs ID: 1.82% Gene Name: 10.34%	SNPs ID: 0.48% Gene Name: 2.58%
$\alpha = Autopsy$ $\beta = Tier1$	SNPs ID: 1.54% Gene Name: 3.33%	SNPs ID: 3.57% Gene Name: 8.21%	SNPs ID: 1.09% Gene Name: 2.42%
$\alpha = NINDS1$ $\beta = NINDS2$	SNPs ID: 0.69% Gene Name: 2.94%	SNPs ID: 3.65% Gene Name: 15.51%	SNPs ID: 0.59% Gene Name: 2.53%
$\alpha = NINDS1$ $\beta = Tier1$	SNPs ID: 1.51% Gene Name: 3.26%	SNPs ID: 6.63% Gene Name: 13.69%	SNPs ID: 1.24% Gene Name: 2.71%
$\alpha = NINDS2$ $\beta = Tier1$	SNPs ID: 0.60% Gene Name: 5.17%	SNPs ID: 0.51% Gene Name: 4.10%	SNPs ID: 0.27% Gene Name: 2.34%

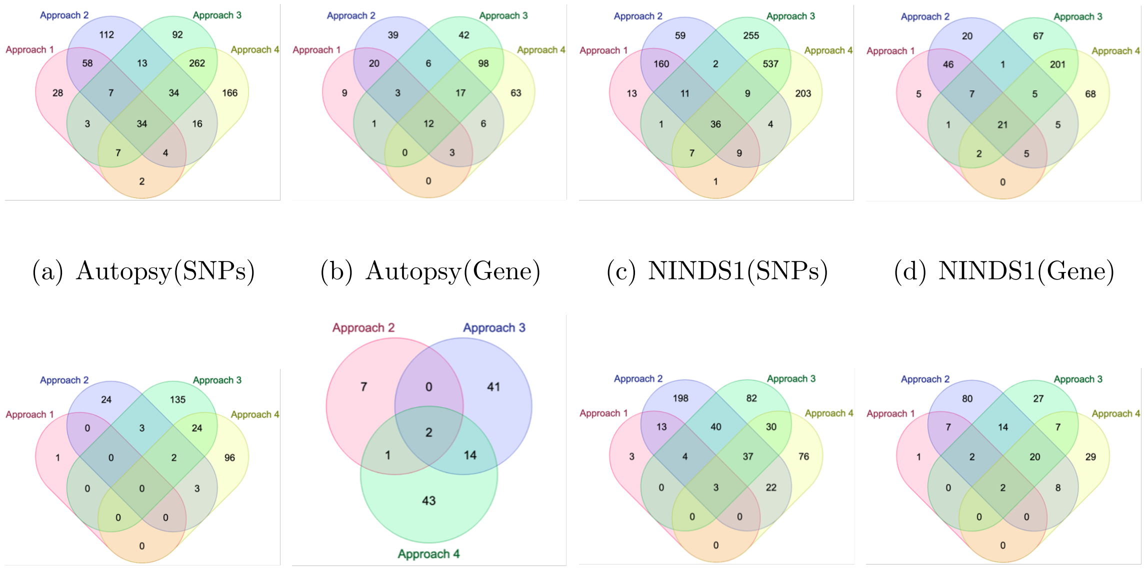
Table 4.15: Percentage of common SNPs & genes between datasets for approach 3

	$method1 = \frac{length(\alpha \cap \beta)}{length(\alpha)} \times 100$	$method2 = \frac{length(\alpha \cap \beta)}{length(\beta)} \times 100$	$method3 = \frac{length(\alpha \cap \beta)}{length(\alpha \cup \beta)} \times 100$
$\alpha = Autopsy$ $\beta = NINDS1$	SNPs ID: 12.95% Gene Name: 24%	SNPs ID: 8.43% Gene Name: 15.58%	SNPs ID: 5.38% Gene Name: 10.43%
$\alpha = Autopsy$ $\beta = NINDS2$	SNPs ID: 0.76% Gene Name: 3%	SNPs ID: 3.2% Gene Name: 9.83%	SNPs ID: 0.61% Gene Name: 2.35%
$\alpha = Autopsy$ $\beta = Tier1$	SNPs ID: 1.90% Gene Name: 4.5%	SNPs ID: 5.95% Gene Name: 13.43%	SNPs ID: 1.46% Gene Name: 3.48%
$\alpha = NINDS1$ $\beta = NINDS2$	SNPs ID: 0.86% Gene Name: 3.57%	SNPs ID: 5.60% Gene Name: 18.03%	SNPs ID: 0.75% Gene Name: 3.07%
$\alpha = NINDS1$ $\beta = Tier1$	SNPs ID: 1.86% Gene Name: 4.22%	SNPs ID: 8.92% Gene Name: 19.40%	SNPs ID: 1.56% Gene Name: 3.59%
$\alpha = NINDS2$ $\beta = Tier1$	SNPs ID: 0% Gene Name: 3.27%	SNPs ID: 0% Gene Name: 2.98%	SNPs ID: 0% Gene Name: 1.58%

Table 4.16: Percentage of common SNPs & genes between datasets for approach 4

4.2 Venn diagram for different approaches and same dataset

In Figure 4.1, we show the number of SNPs and genes that are in common among different approaches for the same dataset. It should be noted that there are some SNPs or genes that are common among at least two approaches or two datasets. We extracted those SNPs and genes and called them the possible biomarkers for PD. The reason that approach 0 is excluded from the Venn diagram is that the SNPs that were chosen in this approach were from different datasets (Autopsy, NINDS1, NINDS2, and Tier1), whereas in the other approaches, the SNPs that were chosen came from the Familial dataset. Diagram (f) only includes 3 approaches because there are no common SNPs between approach 1 and approach 2, 3, 4.



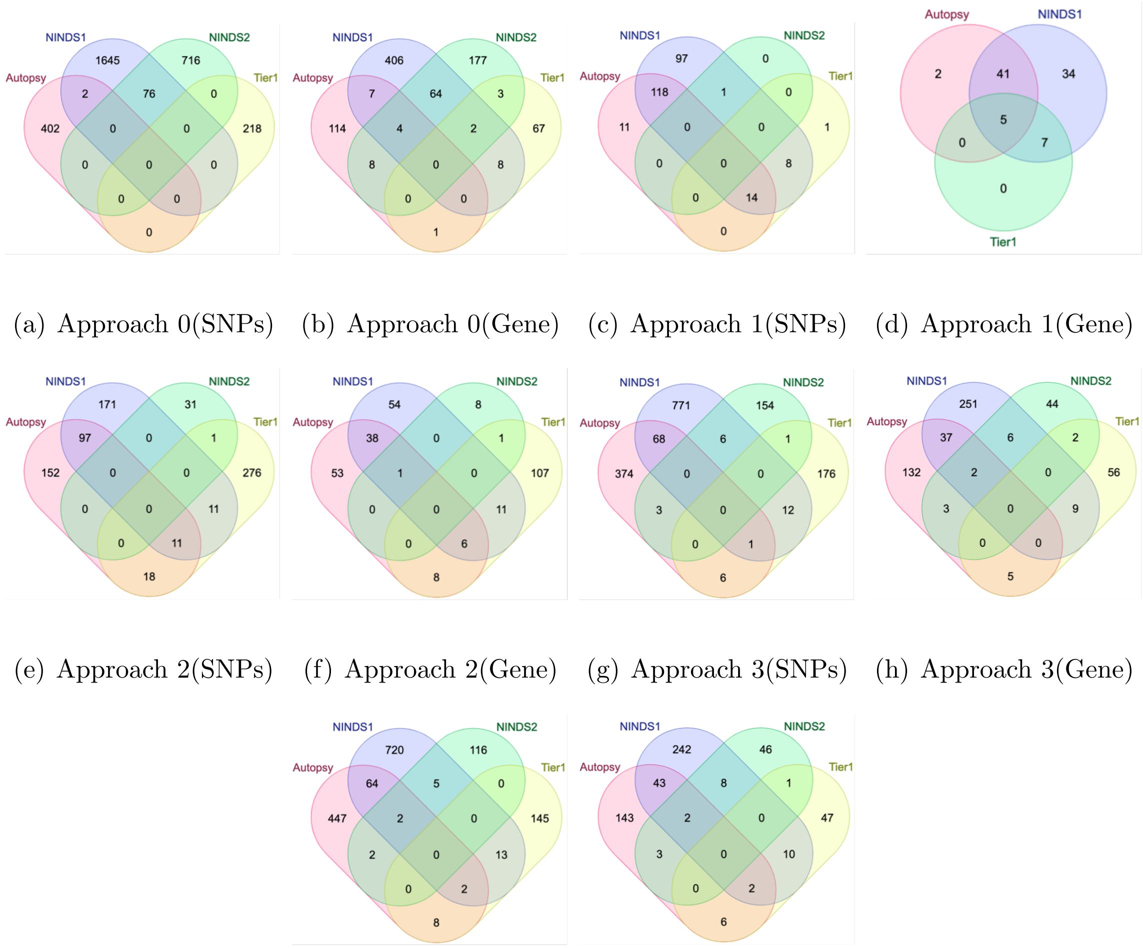
(a) Autopsy(SNPs) (b) Autopsy(Gene) (c) NINDS1(SNPs) (d) NINDS1(Gene)

(e) NINDS2(SNPs) (f) NINDS2(Gene) (g) Tier1(SNPs) (h) Tier1(Gene)

Figure 4.1: Venn diagram for different approaches and same dataset

4.3 Venn diagram for different datasets and same approach

Figure 4.2 shows the number of SNPs and genes that are in common among different datasets for each approach. It is clear that the number of SNPs overlap between datasets fell dramatically when we compare the findings using different datasets but the same approach.



(a) Approach 0(SNPs) (b) Approach 0(Gene) (c) Approach 1(SNPs) (d) Approach 1(Gene)

(e) Approach 2(SNPs) (f) Approach 2(Gene) (g) Approach 3(SNPs) (h) Approach 3(Gene)

(i) Approach 4(SNPs) (j) Approach 4(Gene)

Figure 4.2: Venn diagram for different datasets and same approach

4.4 Biomarkers and associated phenotypes

We selected the SNPs identified by at least two datasets or two approaches and listed their associated phenotype in Tables 4.17, 4.18, 4.19, 4.20, 4.21, 4.22 ,and 4.23 (we have several small tables because it is difficult to fit all this information into one single table). The process of obtaining each SNP phenotype was explained in Section 3.7.6. The third column lists the approaches and datasets were the corresponding SNP was in the list of frequent SNPs. Some of these SNPs were linked to other diseases, however the diseases highlighted were linked to PD indirectly.

We identified four SNP IDs (11248060, rs239748, rs999473, and rs231398) that have a direct link with PD (see table 4.22). Additional research demonstrates that 50 identified SNP IDs (rs13006682, rs1037100, rs1367445, rs2827784, rs4409785, rs11727767, rs2551043, rs7039377, rs4984406, rs1420956, rs2070762, rs4794665, rs6088520, rs2240308, rs2298632, rs3892715, rs4077636, rs7152906, rs1950829, rs1801274, rs12490036, rs12643013, rs6749972, rs1801274, rs1919309, rs2284178, rs901273, rs10894032, rs4300072, rs7554436, rs7646765, rs130423, rs12659814, rs252139, rs194933, rs934178, rs1870676, rs6590810, rs1007415, rs385893, rs11076194, rs4771493, rs4886755, rs799160, rs7026582, rs706779, rs799160, rs9303277, rs9409664, and rs10117) are indirectly associated with PD. These indirectly associated SNPs can be further investigated as potential biomarkers for PD.

Direct association means that current literature directly links a SNP with PD; while an indirect link means that current literature suggests the involvement of a SNP in a disease other than PD but this other disease co-occurs with PD in a significant number of PD patients.

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs13006682 rs1037100	PR interval (The time between atrial depolarization and ventricular depolarization)	Approach 0 (NINDS1 and NINDS2) Approach 4 (Autopsy and NINDS1) NINDS1 (Approach 2 and Approach 3) NINDS1 (Approach 2 and Approach 4) NINDS1 (Approach 3 and Approach 4)	[106]
rs1367445	Osteoporosis Lumbar Spine BMD(bone density)	Approach 0 (NINDS1 and NINDS2)	[107][108][109][110]
rs2827784	C-reactive protein	Approach 0 (NINDS1 and NINDS2)	[111][112]
rs4409785	Rheumatoid arthritis	Approach 0 (NINDS1 and NINDS2)	[113][114][115][116]
	Vitiligo	Approach 0 (NINDS1 and NINDS2)	[117]
	Sex hormone-binding globulin levels	Approach 0 (NINDS1 and NINDS2)	[118][119]
	Myasthenia gravis	Approach 0 (NINDS1 and NINDS2)	[120]
	Multiple sclerosis and Low Density Lipoprotein (LDL) levels	Approach 0 (NINDS1 and NINDS2)	[121][122]
	Multiple sclerosis	Approach 0 (NINDS1 and NINDS2)	[123]
	Medication use thyroid preparations	Approach 0 (NINDS1 and NINDS2)	[124]
	Hypothyroidism,	Approach 0 (NINDS1 and NINDS2)	[125],[126]
	Graves disease	Approach 0 (NINDS1 and NINDS2)	[127]
	Eosinophil counts	Approach 0 (NINDS1 and NINDS2)	[128][129]
	Autoimmune traits	Approach 0 (NINDS1 and NINDS2)	[130][131]
	Autoimmune thyroid diseases (Graves'disease or Hashimoto's thyroiditis)	Approach 0 (NINDS1 and NINDS2)	[132][133]

Table 4.17: SNPs in association with phenotypes - 1

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs11727767 rs2551043 rs7039377 rs4984406 rs1420956	Obesity-related traits	Approach 1 (Autopsy and NINDS1) Approach 1 (NINDS1 and Tier1) Approach 2 (Autopsy and NINDS1) Approach 4 (Autopsy and NINDS1) Autopsy (Approach 1 and Approach 2) Autopsy (Approach 2 and Approach 4) NINDS1 (Approach 1 and Approach 2) NINDS1 (Approach 1 and Approach 2) NINDS1 (Approach 3 and Approach 4)	[134][135][136][137]
rs2070762 rs4794665 rs6088520	Height	Approach 1 (Autopsy and NINDS1) Approach 2 (Autopsy and NINDS1) Autopsy (Approach 1 and Approach 2) Autopsy (Approach 3 and Approach 4) NINDS1 (Approach 1 and Approach 2) NINDS1 (Approach 1 and Approach 3) NINDS1 (Approach 1 and Approach 4) NINDS1 (Approach 2 and Approach 3) NINDS1 (Approach 2 and Approach 4) NINDS1 (Approach 3 and Approach 4) Tier1 (Approach 3 and Approach 4)	[138][139][140]
rs2240308	Oligodontia-colorectal cancer syndrome	Approach 1 (Autopsy and NINDS1)	[141]

Table 4.18: SNPs in association with phenotypes - 2

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs2298632	QT interval (The time it takes for the electrical system to fire an impulse through the ventricles and then recharge)	Approach 1 (Autopsy and NINDS1)	[142][143]
	High Density Lipoprotein (HDL) cholesterol	Approach 1 (Autopsy and NINDS1) Approach 2 (Autopsy and NINDS1) Autopsy (Approach 1 and Approach 2) NINDS1 (Approach 1 and Approach 2)	[144][145]
	Electrocardiographic traits multivariate	Approach 1 (Autopsy and NINDS1) Approach 2 (Autopsy and NINDS1) Autopsy (Approach 1 and Approach 2)	[146]
rs3892715	Attention Deficit Hyperactivity Disorder (ADHD)	Approach 1 (Autopsy and NINDS1) NINDS1 (Approach 1 and Approach 3) NINDS1 (Approach 1 and Approach 4) NINDS1 (Approach 3 and Approach 4)	[147][148][149]
rs4077636	Lung function FEV1 (The amount of air exhaled may be measured during the first)/Forced vital capacity (FVC)	Approach 1 (Autopsy and NINDS1) Approach 1 (Autopsy and Tier1) Approach 1 (NINDS1 and Tier1) Approach 2 (Autopsy and NINDS1) Approach 2 (Autopsy and Tier1) Approach 2 (NINDS1 and Tier1) Autopsy (Approach 1 and Approach 2) NINDS1 (Approach 1 and Approach 2) NINDS1 (Approach 3 and Approach 4) Tier1 (Approach 1 and Approach 2)	[150]

Table 4.19: SNPs in association with phenotypes - 3

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs7152906 rs1950829	Major depressive disorder Multi Trait Analysis of GWAS (MTAG)	Approach 1 (Autopsy and NINDS1) Approach 2 (Autopsy and NINDS1) Approach 2 (NINDS1 and Tier1) Autopsy (Approach 1 and Approach 2) Autopsy (Approach 1 and Approach 3) Autopsy (Approach 1 and Approach 4) Autopsy (Approach 2 and Approach 3) Autopsy (Approach 2 and Approach 4) Autopsy (Approach 3 and Approach 4) NINDS1 (Approach 1 and Approach 2) Tier1 (Approach 1 and Approach 2)	[151][152][153]
rs1801274	Programmed Death Ligand 1 (PDL-1) on cluster of differentiation 14 (CD14+), cluster of differentiation 14 (CD16+) monocyte	Approach 1 (NINDS1 and Tier1) Approach 2 (NINDS1 and Tier1) NINDS1 (Approach 1 and Approach 2) Tier1 (Approach 1 and Approach 2)	[154][155]
	Inflammatory bowel disease	Approach 1 (NINDS1 and Tier1) NINDS1 (Approach 1 and Approach 2) Tier1 (Approach 1 and Approach 2)	[156][157][158]
rs12490036	Mean corpuscular hemoglobin concentration	Approach 2 (Autopsy and NINDS1) Autopsy (Approach 2 and Approach 4)	[159][160][161][162]
rs12643013	Iron	Approach 2 (Autopsy and NINDS1)	[163][164]
rs6749972	Smoking initiation (ever regular vs never regular) (MTAG)	Approach 2 (Autopsy and NINDS1)	[165][166]
rs1801274	Ankylosing spondylitis	Approach 2 (NINDS1 and Tier1) NINDS1 (Approach 1 and Approach 2) Tier1 (Approach 1 and Approach 2)	[167]
rs1919309	Apolipoprotein A1 levels	Approach 3 (Autopsy and NINDS1) NINDS1 (Approach 3 and Approach 4)	[167][168]

Table 4.20: SNPs in association with phenotypes - 4

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs2284178	Behcet Syndrome	Approach 3 (Autopsy and NINDS1) Approach 4 (Autopsy and NINDS1) Autopsy (Approach 3 and Approach 4) NINDS1 (Approach 3 and Approach 4)	[169]
rs901273 rs10894032 rs4300072 rs7554436 rs7646765	Stroke	Approach 3 (Autopsy and NINDS1) Approach 4 (Autopsy and NINDS1) Autopsy (Approach 3 and Approach 4) NINDS1 (Approach 3 and Approach 4)	[170][171]
rs130423	Glucose	Approach 3 (Autopsy and NINDS2) Approach 4 (Autopsy and NINDS1) Approach 4 (Autopsy and NINDS2) Approach 4 (NINDS1 and NINDS2) Autopsy (Approach 3 and Approach 4) NINDS2 (Approach 3 and Approach 4)	[172][173]
rs12659814	Creatinine	Approach 3 (NINDS1 and NINDS2)	[174][175]
rs252139	Amyotrophic lateral sclerosis	Approach 3 (NINDS1 and NINDS2)	[176]
rs194933	Heart Rate	Approach 4 (Autopsy and NINDS1) NINDS1 (Approach 3 and Approach 4)	[177][178]

Table 4.21: SNPs in association with phenotypes - 5

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs934178	Red cell distribution width	Autopsy (Approach 1 and Approach 2) Autopsy (Approach 2 and Approach 3) Autopsy (Approach 2 and Approach 4) Autopsy (Approach 3 and Approach 4)	[179]
rs1870676	Intelligence	Autopsy (Approach 3 and Approach 4)	[180]
rs6590810	Hair color	NINDS1 (Approach 1 and Approach 2)	[181]
rs1007415 rs385893	Platelet Count	NINDS1 (Approach 3 and Approach 4)	[182]
rs11076194	Heel bone mineral density	NINDS1 (Approach 3 and Approach 4)	[183][109]
rs4771493	Numerical cognitive ability	NINDS1 (Approach 3 and Approach 4)	[184]
rs11248060 rs239748 rs999473 rs2313982	Parkinson's disease (PD)	NINDS1 (Approach 3 and Approach 4) Tier1 (Approach 3 and Approach 4)	Direct link with PD
rs4886755 rs799160	Urate levels	NINDS1 (Approach 3 and Approach 4) Tier1 (Approach 2 and Approach 3)	[185][186]
	Hemoglobin	NINDS1 (Approach 3 and Approach 4)	[187][188]
rs7026582	Lipids	NINDS1 (Approach 3 and Approach 4)	[189]

Table 4.22: SNPs in association with phenotypes - 6

SNPs ID	Phenotypes	Datasets / Approaches	Literature linking phenotype to PD
rs706779	Vitiligo	NINDS1 (Approach 3 and Approach 4)	[190]
	Type 1 diabetes	NINDS1 (Approach 3 and Approach 4)	[191][192]
rs799160	Triglyceride levels	Tier1 (Approach 2 and Approach 3)	[193]
rs9303277	Primary biliary cholangitis	Tier1 (Approach 2 and Approach 3)	[194][195]
rs9409664	Inflammation	Tier1 (Approach 2 and Approach 3) Tier1 (Approach 2 and Approach 4) Tier1 (Approach 3 and Approach 4)	[196][197]
rs10117	Even-plus syndrome	Tier1 (Approach 3 and Approach 4)	[198]

Table 4.23: SNPs in association with phenotypes - 7

Chapter 5

Conclusion

This chapter summarizes the major findings in relation to the objectives and research questions and analyzes their importance and contribution. Additionally, I discuss the study's limitations and suggest areas for additional investigation.

The major findings:

1. We identified four SNP IDs (11248060, rs239748, rs999473, and rs231398) that have a direct link with PD. Additional research demonstrates that 50 identified SNP IDs (rs13006682, rs1037100, rs1367445, rs2827784, rs4409785, rs11727767, rs2551043, rs7039377, rs4984406, rs1420956, rs2070762, rs4794665, rs6088520, rs2240308, rs2298632, rs3892715, rs4077636, rs7152906, rs1950829, rs1801274, rs12490036, rs12643013, rs6749972, rs1801274, rs1919309, rs2284178, rs901273, rs10894032, rs4300072, rs7554436, rs7646765, rs130423, rs12659814, rs252139, rs194933, rs934178, rs1870676, rs6590810, rs1007415, rs385893, rs11076194,

rs4771493, rs4886755, rs799160, rs7026582, rs706779, rs799160, rs9303277, rs9409664, and rs10117) are indirectly associated with PD. These indirectly associated SNPs can be further investigated as potential biomarkers for PD.

2. Combining datasets (e.g., Approaches 3 and 4) increases the number of SNPs found in more than one dataset and has comparable performance than training on a single data set (approach 0).

3. RF seems to be a suitable classifier to use with GWAS data once the number of features (SNPs) is reduced by a feature selection step.

4. Our methodology can be applied to a wide range of diseases, including the most prevalent ones like breast cancer, lung cancer, colorectal cancer, and others.

Limitations and future directions:

1. This study was very resource intensive and running the whole process took several days for some of the datasets.

2. Due to different genotyping platforms, the set of SNPs genotyped on each dataset was quite different from some of the other datasets. Because of this, many SNPs were discarded and not considered as potential biomarkers. Repeating this same study on SNPs detected by whole-genome sequencing might resolve this issue.

3. Having access to other datasets obtained from different populations would allow to see if our identified SNPs can be replicated.

Bibliography

[1] O. Tysnes and A. Storstein, "Epidemiology of Parkinson's disease," *Journal of Neural Transmission*, vol. 124, no. 8, pp. 901–905, 2017.

[2] E. R. Dorsey, A. Elbaz, E. Nichols, *et al.*, "Global, regional, and national burden of Parkinson's disease, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016," *The Lancet Neurology*, vol. 17, no. 11, pp. 939–953, 2018.

[3] J. Jankovic, "Parkinson's disease: Clinical features and diagnosis," *Journal of Neurology, Neurosurgery, & Psychiatry*, vol. 79, no. 4, pp. 368–376, 2008.

[4] J. Contreras-Vidal and G. E. Stelmach, "Effects of Parkinsonism on motor control," *Life Sciences*, vol. 58, no. 3, pp. 165–176, 1995.

[5] A. A. of Neurological Surgeons, *American Association of Neurological Surgeons*, `https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Parkinsons-Disease`, Accessed: October 2022.

[6] J. Opara, W. Brola, M. Leonardi, and B. Błaszczyk, “Quality of life in Parkinsons disease,” *Journal of Medicine and Life*, vol. 5, no. 4, p. 375, 2012.

[7] S. J. Johnson, M. D. Diener, A. Kaltenboeck, H. G. Birnbaum, and A. D. Siderowf, “An economic model of Parkinson’s disease: Implications for slowing progression in the United States,” *Movement Disorders*, vol. 28, no. 3, pp. 319–326, 2013.

[8] S. L. Kowal, T. M. Dall, R. Chakrabarti, M. V. Storm, and A. Jain, “The current and projected economic burden of Parkinson’s disease in the United States,” *Movement Disorders*, vol. 28, no. 3, pp. 311–318, 2013.

[9] J.-X. Yang and L. Chen, “Economic burden analysis of Parkinson’s disease patients in China,” *Parkinson’s Disease*, vol. 2017, 2017.

[10] C. Tremblay, P. D. Martel, and J. Frasnelli, “Trigeminal system in Parkinson’s disease: A potential avenue to detect Parkinson-specific olfactory dysfunction,” *Parkinsonism & Related Disorders*, vol. 44, pp. 85–90, 2017.

[11] T. A. Zesiewicz, K. L. Sullivan, and R. A. Hauser, “Nonmotor symptoms of Parkinson’s disease,” *Expert Review of Neurotherapeutics*, vol. 6, no. 12, pp. 1811–1822, 2006.

[12] H. Braak, K. Del Tredici, U. Rüb, R. A. De Vos, E. N. J. Steur, and E. Braak, “Staging of brain pathology related to sporadic Parkinson’s disease,” *Neurobiology of Aging*, vol. 24, no. 2, pp. 197–211, 2003.

[13] R. B. Postuma, D. Berg, M. Stern, *et al.*, "MDS clinical diagnostic criteria for Parkinson's disease," *Movement Disorders*, vol. 30, no. 12, pp. 1591–1601, 2015.

[14] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy, "Decision support framework for Parkinson's disease based on novel handwriting markers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 23, no. 3, pp. 508–516, 2014.

[15] C. R. Pereira, D. R. Pereira, G. H. Rosa, *et al.*, "Handwritten dynamics assessment through convolutional neural networks: An application to Parkinson's disease identification," *Artificial Intelligence in Medicine*, vol. 87, pp. 67–77, 2018.

[16] F. Wahid, R. K. Begg, C. J. Hass, S. Halgamuge, and D. C. Ackland, "Classification of Parkinson's disease gait using spatial-temporal gait features," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 6, pp. 1794–1802, 2015.

[17] T. D. Pham and H. Yan, "Tensor decomposition of gait dynamics in Parkinson's disease," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 8, pp. 1820–1827, 2017.

[18] A. Cherubini, M. Morelli, R. Nisticó, *et al.*, "Magnetic resonance support vector machine discriminates between Parkinson's disease and progressive supranuclear palsy," *Movement Disorders*, vol. 29, no. 2, pp. 266–269, 2014.

[19] H. Choi, S. Ha, H. J. Im, S. H. Paek, and D. S. Lee, "Refining diagnosis of Parkinson's disease with deep learning-based interpretation of dopamine transporter imaging," *NeuroImage: Clinical*, vol. 16, pp. 586–594, 2017.

[20] F. Segovia, J. M. Gorriz, J. Ramirez, F. J. Martinez-Murcia, and D. Castillo-Barnes, "Assisted diagnosis of Parkinsonism based on the striatal morphology," *International Journal of Neural Systems*, vol. 29, no. 09, p. 1 950 011, 2019.

[21] B. E. Sakar, M. E. Isenkul, C. O. Sakar, *et al.*, "Collection and analysis of a Parkinson speech dataset with multiple types of sound recordings," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 4, pp. 828–834, 2013.

[22] C. Ma, J. Ouyang, H.-L. Chen, and X.-H. Zhao, "An efficient diagnosis system for Parkinson's disease using kernel-based extreme learning machine with subtractive clustering features weighting approach," *Computational and Mathematical Methods in Medicine*, vol. 2014, 2014.

[23] P. A. LeWitt, J. Li, M. Lu, *et al.*, "3-hydroxykynurenine and other Parkinson's disease biomarkers discovered by metabolomic analysis," *Movement Disorders*, vol. 28, no. 12, pp. 1653–1660, 2013.

[24] F. Maass, B. Michalke, D. Willkommen, *et al.*, "Elemental fingerprint: Reassessment of a cerebrospinal fluid biomarker for Parkinson's disease," *Neurobiology of Disease*, vol. 134, p. 104 677, 2020.

[25] S. Nuvoli, A. Spanu, M. L. Fravolini, *et al.*, "[123I] Metaiodobenzylguanidine (MIBG) cardiac scintigraphy and automated classification techniques in Parkinsonian disorders," *Molecular Imaging and Biology*, vol. 22, no. 3, pp. 703–710, 2020.

[26] C. Váradi, K. Nehéz, O. Hornyák, B. Viskolcz, and J. Bones, "Serum n-glycosylation in Parkinson's disease: A novel approach for potential alterations," *Molecules*, vol. 24, no. 12, p. 2220, 2019.

[27] A. Nunes, G. Silva, C. Duque, *et al.*, "Retinal texture biomarkers may help to discriminate between Alzheimer, Parkinson, and healthy controls," *PlOS One*, vol. 14, no. 6, e0218826, 2019.

[28] Z. Wang, X. Zhu, E. Adeli, *et al.*, "Multi-modal classification of neurodegenerative disease by progressive graph-based transductive learning," *Medical Image Analysis*, vol. 39, pp. 218–230, 2017.

[29] M. Plus, *What are single nucleotide polymorphisms (SNPs)?* `https://medlineplus.gov/genetics/understanding/genomicresearch/snp/`, 2022.

[30] K. Strimbu and J. A. Tavel, "What are biomarkers?" *Current Opinion in HIV and AIDS*, vol. 5, no. 6, p. 463, 2010.

[31] D. M. M. C. R. M. USA, *Mayo-perlegen LEAPS (linked efforts to accelerate Parkinson's solutions) collaboration*, `https://www.ncbi.nlm.nih.gov/`

projects/gap/cgi-bin/study.cgi?study_id=phs000048.v1.p1, Accessed: September 2022.

[32] D. A. Singleton, *National Institute of Neurological Disorders and Stroke (NINDS) Genome-wide genotyping in Parkinson's disease*, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000089.v4.p2, Accessed: September 2022.

[33] U. Johns Hopkins University Center for Inherited Disease Research (CIDR). Baltimore MD, *CIDR: Genome Wide Association Study in Familial Parkinson's disease (PD)*, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000126.v2.p1&phv=45238&phd=1032&pha=2865&pht=321&phvf=&phdf=&phaf=&phtf=&dssp=1&consent=&temp=1, 2016.

[34] U. Johns Hopkins University Center for Inherited Disease Research (CIDR). Baltimore MD, *Autopsy-confirmed Parkinson disease GWAS consortium (APDGC)*, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000394.v1.p1, 2012.

[35] H. Schwender, "Imputing missing genotypes with weighted k nearest neighbors," *Journal of Toxicology and Environmental Health, Part A*, vol. 75, no. 8-10, pp. 438–446, 2012.

[36] M. Afshar and H. Usefi, "Dimensionality reduction using singular vectors," *Scientific Reports*, vol. 11, no. 1, pp. 1–13, 2021.

[37] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," *Neural Computation*, vol. 26, no. 1, pp. 185–207, 2014.

[38] H. Climente-González, C.-A. Azencott, S. Kaski, and M. Yamada, "Block HSIC Lasso: Model-free biomarker detection for ultra-high dimensional data," *Bioinformatics*, vol. 35, no. 14, pp. i427–i435, 2019.

[39] N. Education, *Deoxyribonucleic acid / DNA*, https://www.nature.com/scitable/definition/dna-107/, Accessed: September 2022.

[40] N. Education, *Genome*, https://www.nature.com/scitable/definition/genome-43/, Accessed: September 2022.

[41] N. Education, *Genes*, https://www.nature.com/scitable/definition/gene-29/, Accessed: September 2022.

[42] N. Education, *Chromosomes*, https://www.nature.com/scitable/definition/chromosome-6/, Accessed: September 2022.

[43] N. Education, *Alleles*, https://www.nature.com/scitable/definition/allele-48/, Accessed: September 2022.

[44] N. Education, *Genotype*, https://www.nature.com/scitable/definition/genotype-234/, Accessed: September 2022.

[45] N. Education, *Phenotypes*, https://www.nature.com/scitable/definition/phenotype-35/, Accessed: September 2022.

[46] N. H. G. R. Institute, *Locus*, `https://www.genome.gov/genetics-glossary/Locus`, 2022.

[47] H. L. Nicholls, C. R. John, D. S. Watson, P. B. Munroe, M. R. Barnes, and C. P. Cabrera, "Reaching the end-game for GWAS: Machine learning approaches for the prioritization of complex disease loci," *Frontiers in Genetics*, vol. 11, p. 350, 2020.

[48] S. Leem, I. Huh, and T. Park, "Enhanced permutation tests via multiple pruning," *Frontiers in Genetics*, vol. 11, p. 509, 2020.

[49] Y. S. Cho, M. J. Go, Y. J. Kim, *et al.*, "A large-scale genome-wide association study of asian populations uncovers genetic factors influencing eight quantitative traits," *Nature Genetics*, vol. 41, no. 5, pp. 527–534, 2009.

[50] M. A. Nalls, N. Pankratz, C. M. Lill, *et al.*, "Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease," *Nature Genetics*, vol. 46, no. 9, pp. 989–993, 2014.

[51] J. Mei, C. Desrosiers, and J. Frasnelli, "Machine learning for the diagnosis of Parkinson's disease: A review of literature," *Frontiers in Aging Neuroscience*, vol. 13, p. 633 752, 2021.

[52] A. J. Espay, P. Bonato, F. B. Nahab, *et al.*, "Technology in Parkinson's disease: Challenges and opportunities," *Movement Disorders*, vol. 31, no. 9, pp. 1272–1282, 2016.

[53] S. Aich, J. Youn, S. Chakraborty, *et al.*, "A supervised machine learning approach to detect the on/off state in Parkinson's disease using wearable based gait signals," *Diagnostics*, vol. 10, no. 6, p. 421, 2020.

[54] Y. M. Aye, S. Liew, S. X. Neo, *et al.*, "Patient-centric care for Parkinson's disease: From hospital to the community," *Frontiers in Neurology*, vol. 11, p. 502, 2020.

[55] E. Rastegari and H. Ali, "A bag-of-words feature engineering approach for assessing health conditions using accelerometer data," *Smart Health*, vol. 16, p. 100 116, 2020.

[56] M. Juutinen, C. Wang, J. Zhu, *et al.*, "Parkinson's disease detection from 20-step walking tests using inertial sensors of a smartphone: Machine learning approach based on an observational case-control study," *PlOS One*, vol. 15, no. 7, e0236258, 2020.

[57] C. Fernandes, L. Fonseca, F. Ferreira, *et al.*, "Artificial neural networks classification of patients with Parkinsonism based on gait," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, 2018, pp. 2024–2030.

[58] F. Cuzzolin, M. Sapienza, P. Esser, *et al.*, "Metric learning for Parkinsonian identification from IMU gait measurements," *Gait & Posture*, vol. 54, pp. 127–132, 2017.

[59] H. Abujrida, E. Agu, and K. Pahlavan, "Machine learning-based motor assessment of Parkinson's disease using postural sway, gait and lifestyle features on crowdsourced smartphone data," *Biomedical Physics & Engineering Express*, vol. 6, no. 3, p. 035 005, 2020.

[60] H. Zhang, K. Deng, H. Li, R. L. Albin, and Y. Guan, "Deep learning identifies digital biomarkers for self-reported Parkinson's disease," *Patterns*, vol. 1, no. 3, p. 100 042, 2020.

[61] L. Kurvits, F. Lättekivi, E. Reimann, *et al.*, "Transcriptomic profiles in Parkinson's disease," *Experimental Biology and Medicine*, vol. 246, no. 5, pp. 584–595, 2021.

[62] C. Su, J. Tong, and F. Wang, "Mining genetic and transcriptomic data using machine learning approaches in Parkinson's disease," *npj Parkinson's Disease*, vol. 6, no. 1, pp. 1–10, 2020.

[63] D. Pietrucci, A. Teofani, V. Unida, *et al.*, "Can gut microbiota be a good predictor for Parkinson's disease? a machine learning approach," *Brain Sciences*, vol. 10, no. 4, p. 242, 2020.

[64] G. Tallapureddy and D. Radha, "Analysis of ensemble of machine learning algorithms for detection of Parkinson's disease," in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, 2022, pp. 354–361.

[65] K. Marek, D. Jennings, S. Lasch, *et al.*, "The Parkinson Progression Marker Initiative (ppmi)," *Progress in Neurobiology*, vol. 95, no. 4, pp. 629–635, 2011.

[66] D. Ahmadi Rastegar, N. Ho, G. M. Halliday, and N. Dzamko, "Parkinson's progression prediction using machine learning and serum cytokines," *NPJ Parkinson's Disease*, vol. 5, no. 1, pp. 1–8, 2019.

[67] M. J. F. F. for Parkinson's Research, *Michael J. Fox Foundation for Parkinson's Research*, `https://www.michaeljfox.org/what-we-fund?lrrk2-cohort-consortium=`, Accessed: October 2022.

[68] D. S. Dhami, A. Soni, D. Page, and S. Natarajan, "Identifying Parkinson's patients: A functional gradient boosting approach," in *Conference on Artificial Intelligence in Medicine in Europe*, Springer, 2017, pp. 332–337.

[69] B. F. O. Coelho, A. B. R. Massaranduba, C. A. dos Santos Souza, G. G. Viana, I. Brys, and R. P. Ramos, "Parkinson's disease effective biomarkers based on hjorth features improved by machine learning," *Expert Systems with Applications*, p. 118 772, 2022.

[70] A. Li and C. Li, "Detecting Parkinson's disease through gait measures using machine learning," *Diagnostics*, vol. 12, no. 10, p. 2404, 2022.

[71] A. L. Goldberger, L. A. Amaral, L. Glass, *et al.*, "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, e215–e220, 2000.

[72] S. Frenkel-Toledo, N. Giladi, L. Gruendlinger, R. Baltadjieva, and J. Hausdorff, “Treadmill walking as an “external cue” to improve gait rhythm and stability in Parkinson’s disease,” in *Journal of the American Geriatrics Society*, Blackwell Publishing Inc 350 Main St, Malden, Ma 02148 Usa, vol. 52, 2004, S217–S217.

[73] M. R. Corces, A. Shcherbina, S. Kundu, *et al.*, “Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer’s and Parkinson’s diseases,” *Nature Genetics*, vol. 52, no. 11, pp. 1158–1168, 2020.

[74] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O’Sullivan, “A review of feature selection methods for machine learning-based disease risk prediction,” *Frontiers in Bioinformatics*, vol. 2, p. 927 312, 2022.

[75] N. L. of Medicine, *National Library of Medicine*, `https://www.ncbi.nlm.nih.gov/gap/`, Accessed: September 2022.

[76] W. Nichols, V. Elsaesser, N. Pankratz, *et al.*, “LRRK2 mutation analysis in Parkinson’s disease families with evidence of linkage to PARK8,” *Neurology*, vol. 69, no. 18, pp. 1737–1744, 2007.

[77] N. Pankratz, M. W. Pauciulo, V. E. Elsaesser, *et al.*, “Mutations in LRRK2 other than G2019S are rare in a north american–based sample of familial Parkinson’s disease,” *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 21, no. 12, pp. 2257–2260, 2006.

[78] W. C. Nichols, N. Pankratz, D. Hernandez, *et al.*, “Genetic screening for a single common LRRK2 mutation in familial Parkinson’s disease,” *The Lancet*, vol. 365, no. 9457, pp. 410–412, 2005.

[79] J. Wilk, J. Tobin, O. Suchowersky, *et al.*, “Herbicide exposure modifies GSTP1 haplotype association to Parkinson onset age: The GenePD Study,” *Neurology*, vol. 67, no. 12, pp. 2206–2210, 2006.

[80] M. Sun, J. C. Latourelle, G. F. Wooten, *et al.*, “Influence of heterozygosity for parkin mutation on onset age in familial Parkinson disease: The GenePD Study,” *Archives of Neurology*, vol. 63, no. 6, pp. 826–832, 2006.

[81] S. Karamohamed, J. Latourelle, B. Racette, *et al.*, “BDNF genetic variants are associated with onset age of familial Parkinson disease: GenePD Study,” *Neurology*, vol. 65, no. 11, pp. 1823–1825, 2005.

[82] S. Karamohamed, L. Golbe, M. Mark, *et al.*, “Absence of previously reported variants in the SCNA (G88C and G209A), NR4A2 (T291D and T245G) and the DJ-1 (T497C) genes in familial Parkinson’s disease from the GenePD Study,” *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 20, no. 9, pp. 1188–1191, 2005.

[83] H.-C. Fung, S. Scholz, M. Matarin, *et al.*, “Genome-wide genotyping in Parkinson’s disease and neurologically normal controls: First stage analysis and public release of data,” *The Lancet Neurology*, vol. 5, no. 11, pp. 911–916, 2006.

[84] J. Simon-Sanchez, S. Scholz, H.-C. Fung, *et al.*, "Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line induced alterations in normal individuals," *Human Molecular Genetics*, vol. 16, no. 1, pp. 1–14, 2007.

[85] J. Simon-Sanchez, C. Schulte, J. M. Bras, *et al.*, "Genome-wide association study reveals genetic risk underlying Parkinson's disease," *Nature Genetics*, vol. 41, no. 12, pp. 1308–1312, 2009.

[86] J. M. Sanz, S. Falzoni, M. L. Morieri, A. Passaro, G. Zuliani, and F. Di Virgilio, "Association of hypomorphic P2X7 receptor genotype with age," *Frontiers in Molecular Neuroscience*, vol. 13, p. 8, 2020.

[87] D. M. Maraganore, M. De Andrade, T. G. Lesnick, *et al.*, "High-resolution whole-genome association study of Parkinson's disease," *The American Journal of Human Genetics*, vol. 77, no. 5, pp. 685–693, 2005.

[88] E. Evangelou, D. M. Maraganore, and J. P. Ioannidis, "Meta-analysis in genome-wide association datasets: Strategies and application in Parkinson's disease," *PLOS One*, vol. 2, no. 2, e196, 2007.

[89] T. G. Lesnick, S. Papapetropoulos, D. C. Mash, *et al.*, "A genomic pathway approach to a complex disease: Axon guidance and Parkinson's disease," *PLOS Genetics*, vol. 3, no. 6, e98, 2007.

[90] N. L. of Medicine, *National Centre for Biotechnology Information*, `https://www.ncbi.nlm.nih.gov`, Accessed: September 2022.

[91] OmicScript, *PED*, `http://www.arrayserver.com/wiki/index.php?title=SNP_Data%3A_PED_file_%2B_Map_file`, 2017.

[92] S. T. D. Team, *SRA Toolkit*, `https://github.com/ncbi/sra-tools`, 2020.

[93] Cog-genomics, *BED*, `https://www.cog-genomics.org/plink/1.9/formats#bed`, Accessed: September 2022.

[94] Cog-genomics, *BIM*, `https://www.cog-genomics.org/plink/1.9/formats#bim`, Accessed: September 2022.

[95] G. Analysis, *File format reference*, `https://plink.readthedocs.io/en/latest/plink_fmt/`, Accessed: September 2022.

[96] Cog-genomics, *FAM*, `https://www.cog-genomics.org/plink/1.9/formats#fam`, Accessed: September 2022.

[97] Cog-genomics, *MAP*, `https://www.cog-genomics.org/plink/1.9/formats#map`, Accessed: September 2022.

[98] S. Purcell, B. Neale, K. Todd-Brown, *et al.*, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, 2007.

[99] H. S. A. o. G. Schwender and U. o. D. Gene Expression Data. Dissertation Department of Statistics, *KNNcatimpute: Missing value imputation with knn*, `https://www.rdocumentation.org/packages/scrime/versions/1.3.5/topics/knncatimpute`, 2007.

[100] A. R. Quinlan and I. M. Hall, "BEDTools: A flexible suite of utilities for comparing genomic features," *Bioinformatics*, vol. 26, no. 6, pp. 841–842, 2010.

[101] H. S. I. C. L. (Lasso), *Pyhsiclasso*, `https://github.com/riken-aip/pyHSICLasso`, 2020.

[102] ScienceDirect, *Linkage Disequilibrium*, `https://www.sciencedirect.com/topics/neuroscience/linkage-disequilibrium`, 2022.

[103] M. Arnold, J. Raffler, A. Pfeufer, K. Suhre, and G. Kastenmüller, "SNiPA: An interactive, genetic variant-centered annotation browser," *Bioinformatics*, vol. 31, no. 8, pp. 1334–1336, 2015.

[104] Ensemble, *Computes and returns LD values between the given variant and all other variants in a window centered around the given variant*, `https://rest.ensembl.org/documentation/info/ld_id_get`, 2022.

[105] Ensemble, *Ensemble Biomart*, `http://useast.ensembl.org/biomart/martview/fbd18081f77c6d6aa53ba6f63544988b`, 2022.

[106] H. Mochizuki, Y. Ebihara, Y. Ugawa, *et al.*, “PR prolongation and cardiac 123I-MIBG uptake reduction in Parkinson’s disease,” *European Neurology*, vol. 74, no. 1-2, pp. 107–111, 2015.

[107] C. A. Figueroa and C. J. Rosen, “Parkinson’s disease and osteoporosis: Basic and clinical implications,” *Expert Review of Endocrinology & Metabolism*, vol. 15, no. 3, pp. 185–193, 2020.

[108] L. M. Raglione, S. Sorbi, and B. Nacmias, “Osteoporosis and Parkinson’s disease,” *Clinical Cases in Mineral and Bone Metabolism*, vol. 8, no. 3, p. 16, 2011.

[109] H. Gao, X. Wei, J. Liao, *et al.*, “Lower bone mineral density in patients with Parkinson’s disease: A cross-sectional study from Chinese Mainland,” *Frontiers in Aging Neuroscience*, vol. 7, p. 203, 2015.

[110] F. Van Den Bos, A. D. Speelman, M. Samson, M. Munneke, B. R. Bloem, and H. J. Verhaar, “Parkinson’s disease and osteoporosis,” *Age and Ageing*, vol. 42, no. 2, pp. 156–162, 2013.

[111] X. Qiu, Y. Xiao, J. Wu, L. Gan, Y. Huang, and J. Wang, “C-reactive protein and risk of Parkinson’s disease: A systematic review and meta-analysis,” *Frontiers in Neurology*, vol. 10, p. 384, 2019.

[112] P. Lyra, J. Botelho, V. Machado, *et al.*, "Self-reported periodontitis and C-reactive protein in Parkinson's disease: A cross-sectional study of two American cohorts," *npj Parkinson's Disease*, vol. 8, no. 1, pp. 1–6, 2022.

[113] T. Kogure, T. Tatsumi, Y. Kaneko, and K. Okamoto, "Rheumatoid arthritis accompanied by Parkinson's disease," *JCR: Journal of Clinical Rheumatology*, vol. 14, no. 3, pp. 192–193, 2008.

[114] D. Li, X. Hong, and T. Chen, "Association between rheumatoid arthritis and risk of Parkinson's disease: A meta-analysis and systematic review," *Frontiers in Neurology*, vol. 13, 2022.

[115] C. Li, R. Ou, and H. Shang, "Rheumatoid arthritis decreases risk for Parkinson's disease: A mendelian randomization study," *npj Parkinson's Disease*, vol. 7, no. 1, pp. 1–5, 2021.

[116] J. Bacelis, M. Compagno, S. George, *et al.*, "Decreased risk of Parkinson's disease after rheumatoid arthritis diagnosis: A nested case-control study with matched cases and controls," *Journal of Parkinson's Disease*, vol. 11, no. 2, pp. 821–832, 2021.

[117] B. Bellei, A. Pitisci, M. Ottaviani, *et al.*, "Vitiligo: A possible model of degenerative diseases," *PlOS One*, vol. 8, no. 3, e59782, 2013.

[118] M. Nitkowska, R. Tomasiuk, M. Czyżyk, and A. Friedman, “Prolactin and sex hormones levels in males with Parkinson’s disease,” *Acta Neurologica Scandinavica*, vol. 131, no. 6, pp. 411–416, 2015.

[119] C. Kusters, K. Paul, A. D. Folle, *et al.*, “SHBG, and possibly testosterone, are associated with the risk for Parkinson’s disease among women: A Mendelian randomization approach,” in *Movement Disorders*, Wiley 111 River St, Hoboken 07030-5774, Nj USA, vol. 36, 2021, S313–S313.

[120] I. Odajiu, E. I. Davidescu, C. Mitu, and B. O. Popescu, “Patients with Parkinson’s disease and myasthenia gravis—a report of three new cases and review of the literature,” *Medicina*, vol. 56, no. 1, p. 5, 2019.

[121] A. Witoelar, I. E. Jansen, Y. Wang, *et al.*, “Genome-wide pleiotropy between Parkinson’s disease and autoimmune diseases,” *JAMA Neurology*, vol. 74, no. 7, pp. 780–792, 2017.

[122] X. Zhang, A. M. Lucas, Y. Veturi, *et al.*, “Large-scale genomic analyses reveal insights into pleiotropy across circulatory system diseases and nervous system disorders,” *Nature Communications*, vol. 13, no. 1, pp. 1–12, 2022.

[123] V. Shaygannejad, M. Shirmardi, L. Dehghani, and H. Maghzi, “Co-occurrence of multiple sclerosis and Parkinson’s disease,” *Advanced Biomedical Research*, vol. 5, 2016.

[124] T. D. Wingert and J. M. Hershman, "Sinemet® and thyroid function in Parkinson's disease," *Neurology*, vol. 29, no. 7, pp. 1073–1073, 1979.

[125] S.-F. Chen, Y.-C. Yang, C.-Y. Hsu, and Y.-C. Shen, "Risk of Parkinson's disease in patients with hypothyroidism: A nationwide population-based cohort study," *Parkinsonism & Related Disorders*, vol. 74, pp. 28–32, 2020.

[126] J. Garcia-Moreno and J. Chacon, "Hypothyroidism concealed by Parkinson's disease," *Revista de Neurologia*, vol. 35, no. 8, pp. 741–742, 2002.

[127] Y. Y. Cho, B. Kim, D. W. Shin, *et al.*, "Graves' disease and the risk of Parkinson's disease: A Korean population-based study," *Brain Communications*, vol. 4, no. 1, fcac014, 2022.

[128] M. P. Jensen, B. M. Jacobs, R. Dobson, *et al.*, "Lower lymphocyte count is associated with increased risk of Parkinson's disease," *Annals of Neurology*, vol. 89, no. 4, pp. 803–812, 2021.

[129] G.-T. Liu, C.-S. Hwang, C.-H. Hsieh, *et al.*, "Eosinophil-derived neurotoxin is elevated in patients with amyotrophic lateral sclerosis," *Mediators of Inflammation*, vol. 2013, 2013.

[130] K. Rugbjerg, S. Friis, B. Ritz, E. Schernhammer, L. Korbo, and J. Olsen, "Autoimmune disease and risk for Parkinson's disease: A population-based case-control study," *Neurology*, vol. 73, no. 18, pp. 1462–1468, 2009.

[131] F. Garretti, D. Agalliu, C. S. Lindestam Arlehamn, A. Sette, and D. Sulzer, "Autoimmunity in Parkinson's disease: The role of α-synuclein-specific T cells," *Frontiers in Immunology*, vol. 10, p. 303, 2019.

[132] N. Charoenngam, T. Rittiphairoj, B. Ponvilawan, and K. Prasongdee, "Thyroid Dysfunction and risk of Parkinson's disease: A Systematic Review and Meta-Analysis," *Frontiers in Endocrinology*, vol. 13, 2022.

[133] U. Bonuccelli, C. D'Avino, N. Caraccio, *et al.*, "Thyroid function and autoimmunity in Parkinson's disease: A study of 101 patients," *Parkinsonism & Related Disorders*, vol. 5, no. 1-2, pp. 49–53, 1999.

[134] N. Palacios, X. Gao, M. L. McCullough, *et al.*, "Obesity, diabetes, and risk of Parkinson's disease," *Movement Disorders*, vol. 26, no. 12, pp. 2253–2259, 2011.

[135] J. Chen, Z. Guan, L. Wang, G. Song, B. Ma, and Y. Wang, "Meta-analysis: Overweight, obesity, and Parkinson's disease," *International Journal of Endocrinology*, vol. 2014, 2014.

[136] Y.-L. Wang, Y.-T. Wang, J.-F. Li, Y.-Z. Zhang, H.-L. Yin, and B. Han, "Body mass index and risk of Parkinson's disease: A dose-response meta-analysis of prospective studies," *PlOS One*, vol. 10, no. 6, e0131778, 2015.

[137] K.-Y. Park, G. E. Nam, K. Han, H.-K. Park, and H.-S. Hwang, “Waist circumference and risk of Parkinson’s disease,” *npj Parkinson’s Disease*, vol. 8, no. 1, pp. 1–8, 2022.

[138] P. Ragonese, M. D’Amelio, G. Callari, F. Aiello, L. Morgante, and G. Savettieri, “Height as a potential indicator of early life events predicting Parkinson’s disease: A case-control study,” *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 22, no. 15, pp. 2263–2267, 2007.

[139] U. M. Fietzek, F. E. Schroeteler, L. Hahn, K. Ziegler, and A. O. Ceballos-Baumann, “Body height loss characterizes camptocormia in Parkinson’s disease,” *Journal of Neural Transmission*, vol. 125, no. 10, pp. 1473–1480, 2018.

[140] L. Saari, E. A. Backman, P. Wahlsten, M. Gardberg, and V. Kaasinen, “Height and nigral neuron density in Parkinson’s disease,” *BMC Neurology*, vol. 22, no. 1, pp. 1–6, 2022.

[141] H. Fang, Y. Du, S. Pan, M. Zhong, and J. Tang, “Patients with Parkinson’s disease predict a lower incidence of colorectal cancer,” *BMC Geriatrics*, vol. 21, no. 1, pp. 1–8, 2021.

[142] H. Oka, S. Mochio, H. Sato, and K. Katayama, “Prolongation of QTc interval in patients with Parkinson’s disease,” *European Neurology*, vol. 37, no. 3, pp. 186–189, 1997.

[143] F. Ishizaki, T. Harada, H. Yoshinaga, T. Nakayama, Y. Yamamura, and S. Nakamura, “Prolonged QTc intervals in Parkinson’s disease–relation to sudden death and autonomic dysfunction,” *No to shinkei= Brain and nerve*, vol. 48, no. 5, pp. 443–448, 1996.

[144] M. C. Bakeberg, A. M. Gorecki, J. E. Kenna, *et al.*, “Elevated HDL levels linked to poorer cognitive ability in females with Parkinson’s disease,” *Frontiers in Aging Neuroscience*, vol. 13, p. 287, 2021.

[145] X. Huang, N. W. Sterling, G. Du, *et al.*, “Brain cholesterol metabolism and Parkinson’s disease,” *Movement Disorders*, vol. 34, no. 3, pp. 386–395, 2019.

[146] O. Akbilgic, R. Kamaleswaran, A. Mohammed, *et al.*, “Electrocardiographic changes predate Parkinson’s disease onset,” *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[147] A. A. Baumeister, “Is Attention-Deficit/Hyperactivity Disorder a Risk syndrome for Parkinson’s disease?” *Harvard Review of Psychiatry*, vol. 29, no. 2, pp. 142–158, 2021.

[148] H.-C. Fan, Y.-K. Chang, J.-D. Tsai, *et al.*, “The association between Parkinson’s disease and Attention-Deficit Hyperactivity Disorder,” *Cell Transplantation*, vol. 29, p. 0 963 689 720 947 416, 2020.

[149] S. Becker, M. J. Sharma, and B. L. Callahan, "ADHD and Neurodegenerative disease risk: A critical examination of the evidence," *Frontiers in Aging Neuroscience*, vol. 13, 2021.

[150] D. A. Kaminsky, D. G. Grosset, D. M. Kegler-Ebo, *et al.*, "Natural history of lung function over one year in patients with Parkinson's disease," *Respiratory Medicine*, vol. 182, p. 106 396, 2021.

[151] L. Marsh, "Depression and Parkinson's disease: Current knowledge," *Current Neurology and Neuroscience Reports*, vol. 13, no. 12, pp. 1–9, 2013.

[152] F. M. Nilsson, L. V. Kessing, T. M. Sørensen, P. K. Andersen, and T. G. Bolwig, "Major depressive disorder in Parkinson's disease: A register-based study," *Acta Psychiatrica Scandinavica*, vol. 106, no. 3, pp. 202–211, 2002.

[153] A. M. Hemmerle, J. P. Herman, and K. B. Seroogy, "Stress, depression and Parkinson's disease," *Experimental Neurology*, vol. 233, no. 1, pp. 79–86, 2012.

[154] R. S. Wijeyekoon, D. Kronenberg-Versteeg, K. M. Scott, *et al.*, "Monocyte function in Parkinson's disease and the impact of autologous serum on phagocytosis," *Frontiers in Neurology*, vol. 9, p. 870, 2018.

[155] K. Ando, K. Hamada, M. Shida, *et al.*, "A high number of PD-L1+ CD14+ monocytes in peripheral blood is correlated with shorter survival in patients receiving immune checkpoint inhibitors," *Cancer Immunology, Immunotherapy*, vol. 70, no. 2, pp. 337–348, 2021.

[156] H.-S. Lee, E. Lobbestael, S. Vermeire, J. Sabino, and I. Cleynen, "Inflammatory bowel disease and Parkinson's disease: Common pathophysiological links," *Gut*, vol. 70, no. 2, pp. 408–417, 2021.

[157] T. Brudek, "Inflammatory bowel diseases and Parkinson's disease," *Journal of Parkinson's Disease*, vol. 9, no. s2, S331–S344, 2019.

[158] M. K. Herrick and M. G. Tansey, "Is LRRK2 the missing link between inflammatory bowel disease and Parkinson's disease?" *npj Parkinson's Disease*, vol. 7, no. 1, pp. 1–7, 2021.

[159] Q. Deng, X. Zhou, J. Chen, *et al.*, "Lower hemoglobin levels in patients with Parkinson's disease are associated with disease severity and iron metabolism," *Brain Research*, vol. 1655, pp. 145–151, 2017.

[160] R. D. Abbott, G. W. Ross, C. M. Tanner, *et al.*, "Late-life hemoglobin and the incidence of Parkinson's disease," *Neurobiology of Aging*, vol. 33, no. 5, pp. 914–920, 2012.

[161] J. A. Santiago and J. A. Potashkin, "Blood transcriptomic meta-analysis identifies dysregulation of hemoglobin and iron metabolism in Parkinson's disease," *Frontiers in Aging Neuroscience*, vol. 9, p. 73, 2017.

[162] J. Freed and L. Chakrabarti, "Defining a role for hemoglobin in Parkinson's disease," *npj Parkinson's Disease*, vol. 2, no. 1, pp. 1–4, 2016.

[163] S. L. Rhodes and B. Ritz, "Genetics of iron regulation and the possible role of iron in Parkinson's disease," *Neurobiology of Disease*, vol. 32, no. 2, pp. 183–195, 2008.

[164] L. Shi, C. Huang, Q. Luo, *et al.*, "The association of iron and the pathologies of Parkinson's diseases in MPTP/MPP+-induced neuronal degeneration in non-human primates and in cell culture," *Frontiers in Aging Neuroscience*, vol. 11, p. 215, 2019.

[165] C. Wang, C. Zhou, T. Guo, P. Huang, X. Xu, and M. Zhang, "Association between cigarette smoking and Parkinson's disease: A neuroimaging study," *Therapeutic Advances in Neurological Disorders*, vol. 15, p. 17 562 864 221 092 566, 2022.

[166] B. Ritz, P.-C. Lee, C. F. Lassen, and O. A. Arah, "Parkinson's disease and smoking revisited: Ease of quitting is an early sign of the disease," *Neurology*, vol. 83, no. 16, pp. 1396–1402, 2014.

[167] F.-C. Yeh, H.-C. Chen, Y.-C. Chou, *et al.*, "Positive association of Parkinson's disease with ankylosing spondylitis: A nationwide population-based study," *Journal of Translational Medicine*, vol. 18, no. 1, pp. 1–8, 2020.

[168] C. R. Swanson, K. Li, T. L. Unger, *et al.*, "Lower plasma apolipoprotein A1 levels are found in Parkinson's disease and associate with apolipoprotein A1 genotype," *Movement Disorders*, vol. 30, no. 6, pp. 805–812, 2015.

[169] H. Y. Park, J. H. Lee, S. Y. Lee, *et al.*, "Risk for Parkinson's disease in patients with Behçet's disease: A nationwide population-based dynamic cohort study in Korea," *Journal of Parkinson's Disease*, vol. 9, no. 3, pp. 583–589, 2019.

[170] Y.-P. Huang, L.-S. Chen, M.-F. Yen, *et al.*, "Parkinson's disease is related to an increased risk of ischemic stroke—a population-based propensity score-matched follow-up study," *PLOS One*, vol. 8, no. 9, e68314, 2013.

[171] R. Caslake, K. S. Taylor, and C. E. Counsell, "Parkinson's disease misdiagnosed as stroke," *Case Reports*, vol. 2009, bcr0720080558, 2009.

[172] A. Marques, F. Dutheil, E. Durand, *et al.*, "Glucose dysregulation in Parkinson's disease: Too much glucose or not enough insulin?" *Parkinsonism & Related Disorders*, vol. 55, pp. 122–127, 2018.

[173] R. Sandyk, "The relationship between diabetes mellitus and Parkinson's disease," *International Journal of Neuroscience*, vol. 69, no. 1-4, pp. 125–130, 1993.

[174] L.-L. Zhong, Y.-Q. Song, X.-Y. Tian, H. Cao, and K.-J. Ju, "Level of uric acid and uric acid/creatinine ratios in correlation with stage of Parkinson's disease," *Medicine*, vol. 97, no. 26, 2018.

[175] J.-J. Mo, L.-Y. Liu, W.-B. Peng, J. Rao, Z. Liu, and L.-L. Cui, "The effectiveness of creatine treatment for Parkinson's disease: An updated meta-analysis of randomized controlled trials," *BMC Neurology*, vol. 17, no. 1, pp. 1–9, 2017.

[176] D. A. Bosco, M. J. LaVoie, G. A. Petsko, and D. Ringe, “Proteostasis and movement disorders: Parkinson’s disease and amyotrophic lateral sclerosis,” *Cold Spring Harbor Perspectives in Biology*, vol. 3, no. 10, a007500, 2011.

[177] V. Arnao, A. Cinturino, S. Mastrilli, *et al.*, “Impaired circadian heart rate variability in Parkinson’s disease: A time-domain analysis in ambulatory setting,” *BMC Neurology*, vol. 20, no. 1, pp. 1–5, 2020.

[178] A. Alonso, X. Huang, T. H. Mosley, G. Heiss, and H. Chen, “Heart rate variability and the risk of Parkinson’s disease: A therosclerosis risk in communities study,” *Annals of Neurology*, vol. 77, no. 5, pp. 877–883, 2015.

[179] G. Kenangil, B. Ari, F. Kaya, M. Demir, and F. Domac, “Red cell distribution width levels in Parkinson’s disease patients,” *Acta Neurologica Belgica*, vol. 120, no. 5, pp. 1147–1150, 2020.

[180] C. Fardell, K. Torén, L. Schiöler, H. Nissbrandt, and M. Åberg, “High IQ in early adulthood is associated with Parkinson’s disease,” *Journal of Parkinson’s Disease*, vol. 10, no. 4, pp. 1649–1656, 2020.

[181] X. Gao, K. C. Simon, J. Han, M. A. Schwarzschild, and A. Ascherio, “Genetic determinants of hair color and Parkinson’s disease risk,” *Annals of Neurology*, vol. 65, no. 1, pp. 76–82, 2009.

[182] A. Koçer, A. Yaman, E. Niftaliyev, H. Dürüyen, M. Eryılmaz, and E. Koçer, “Assessment of platelet indices in patients with neurodegenerative diseases:

Mean platelet volume was increased in patients with Parkinson's disease," *Current Gerontology and Geriatrics Research*, vol. 2013, 2013.

[183] J. Caplliure-Llopis, D. Escriv, E. Navarro-Illana, M. Benlloch, J. E. de la Rubia Orti, and C. Barrios, "Bone Quality in Patients with Parkinson's Disease Determined by Quantitative Ultrasound (QUS) of the Calcaneus: Influence of Sex Differences," *International Journal of Environmental Research and Public Health*, vol. 19, no. 5, p. 2804, 2022.

[184] G. S. Watson and J. B. Leverenz, "Profile of cognitive impairment in Parkinson's disease," *Brain Pathology*, vol. 20, no. 3, pp. 640–645, 2010.

[185] S. Cipriani, X. Chen, and M. A. Schwarzschild, "Urate: A novel biomarker of Parkinson's disease risk, diagnosis and prognosis," *Biomarkers in Medicine*, vol. 4, no. 5, pp. 701–712, 2010.

[186] M. Wen, B. Zhou, Y.-H. Chen, *et al.*, "Serum uric acid levels in patients with Parkinson's disease: A meta-analysis," *PlOS One*, vol. 12, no. 3, e0173731, 2017.

[187] C. T. Hong, Y. H. Huang, H. Y. Liu, H.-Y. Chiou, L. Chan, and L.-N. Chien, "Newly diagnosed anemia increases risk of Parkinson's disease: A population-based cohort study," *Scientific Reports*, vol. 6, no. 1, pp. 1–7, 2016.

[188] J. H. Kim, J. K. Oh, J. H. Wee, C. Y. Min, D. M. Yoo, and H. G. Choi, "The association between anemia and Parkinson's disease: A nested case-control

study using a national health screening cohort," *Brain Sciences*, vol. 11, no. 5, p. 623, 2021.

[189] M. Fais, A. Dore, M. Galioto, G. Galleri, C. Crosio, and C. Iaccarino, "Parkinson's disease-related genes and lipid alteration," *International Journal of Molecular Sciences*, vol. 22, no. 14, p. 7630, 2021.

[190] A.-H. Ravn, J. P. Thyssen, and A. Egeberg, "Skin disorders in Parkinson's disease: Potential biomarkers and risk factors," *Clinical, Cosmetic and Investigational Dermatology*, vol. 10, p. 87, 2017.

[191] A. Hassan, R. S. Kandel, R. Mishra, J. Gautam, A. Alaref, and N. Jahan, "Diabetes mellitus and Parkinson's disease: Shared pathophysiological links and possible therapeutic implications," *Cureus*, vol. 12, no. 8, 2020.

[192] D. Sergi, J. Renaud, N. Simola, and M.-G. Martinoli, "Diabetes, a contemporary risk for Parkinson's disease: Epidemiological and cellular evidences," *Frontiers in Aging Neuroscience*, vol. 11, p. 302, 2019.

[193] X. Huang, S. Y.-E. Ng, N. S.-Y. Chia, *et al.*, "Higher serum triglyceride levels are associated with Parkinson's disease mild cognitive impairment," *Movement Disorders: Official Journal of the Movement Disorder Society*, vol. 33, no. 12, pp. 1970–1971, 2018.

[194] M. R. H. Tehrani and M. Poursadeghfard, “Parkinson’s disease accompanied by primary biliary cirrhosis: A case report,” *Gastroenterology Nursing*, vol. 43, no. 2, pp. 196–198, 2020.

[195] V. A. Mosher, M. G. Swain, J. X. Pang, *et al.*, “Primary biliary cholangitis alters functional connections of the brain’s deep gray matter,” *Clinical and Translational Gastroenterology*, vol. 8, no. 7, e107, 2017.

[196] M. Pajares, A. I Rojo, G. Manda, L. Boscá, and A. Cuadrado, “Inflammation in Parkinson’s disease: Mechanisms and therapeutic implications,” *Cells*, vol. 9, no. 7, p. 1687, 2020.

[197] M. G. Tansey, R. L. Wallings, M. C. Houser, M. K. Herrick, C. E. Keating, and V. Joers, “Inflammation and immune dysfunction in Parkinson’s disease,” *Nature Reviews Immunology*, pp. 1–17, 2022.

[198] M. A. Moseng, J. C. Nix, and R. C. Page, “Biophysical consequences of even-plus syndrome mutations for the function of mortalin,” *The Journal of Physical Chemistry B*, vol. 123, no. 16, pp. 3383–3396, 2019.