SAPIENZA
Università di Roma
Facoltà di Scienze Matematiche Fisiche e Naturali

DOTTORATO DI RICERCA
IN GENETICA E BIOLOGIA MOLECOLARE
XXXVI Ciclo
(A.A. 2023/2024)

Machine learning methods applied to classify complex diseases using genomic data

Dottoranda:
Magdalena Arnal Segura

Docente guida.
Prof. /Dr Gian Gaetano Tartaglia

Coordinatore
Prof. Fulvio Cruciani

Index

Glossary

AD: *Alzheimer's disease* is a progressive neurodegenerative disorder characterized by cognitive decline, memory loss, and changes in behavior, primarily affecting older individuals, and is the most common cause of dementia. The disease is associated with the accumulation of abnormal protein deposits, including beta-amyloid plaques and tau tangles, in the brain.

ADNI: The *Alzheimer's Disease Neuroimaging Initiative* is a multi-site research study that collects and shares data, including genetics and cognitive tests, aiming to develop and validate biomarkers for the early detection and tracking of Alzheimer's disease (AD).

AI: *Artificial Intelligence* refers to the development of computer systems that can perform tasks by learning from data or experience.

AUC-ROC: *Area Under the Receiver Operating Characteristic Curve* is a metric used to evaluate the performance of a binary classification model. It represents the area under the curve created by plotting the true positive rate against the false positive rate across different classification thresholds.

CNN: *Convolutional neural networks* are a class of deep learning models specifically designed for processing structured grid data, such as images. They use convolutional layers to automatically and adaptatively learn hierarchical representations, making them highly effective for tasks like image recognition and computer vision.

CV: Cross-validation is a statistical technique used to assess the performance and generalizability of a machine learning model by partitioning the dataset into subsets, training the model on some subsets, and evaluating it on the remaining ones, helping to mitigate issues of overfitting or underfitting.

Cytoband: A cytoband is a distinct banding pattern on a stained chromosome, created through techniques like G-banding, and is used to identify and describe specific regions of chromosomes in cytogenetics.

dbGaP: *Database of Genotypes and Phenotypes* is a publicly available repository of genetic and phenotypic data from human studies.

DNA: *Deoxyribonucleic acid* carries the genetic instructions used in the development, functioning, and reproduction of all known living organisms and many viruses. DNA consists of two long strands forming a double helix, with each strand made up of nucleotides containing a sugar (deoxyribose), a phosphate group, and one of four nitrogenous bases: adenine (A), thymine (T), cytosine (C), or guanine (G).

DE: *Layer deeplift* is a method used in the field of XAI that aims to provide insights into how input features in a deep learning model contribute to the model's output.

DL: *Deep learning* is a subset of machine learning that involves the training of artificial neural networks with multiple layers (deep neural networks) to automatically learn hierarchical representations of data, enabling the extraction of complex features and patterns for diverse tasks.

EOAD: *Early onset Alzheimer's disease* refers to the occurrence of Alzheimer's disease symptoms in individuals under the age of 65.

eQTL: *Expression quantitative trait loci* refer to genomic variants associated with changes in gene expression levels, indicating a genetic influence on the regulation of gene activity.

ET: *Extremely randomized trees* is an ensemble learning method in machine learning that builds multiple decision trees during training and combines their predictions for increased accuracy and robustness.

Features: They are the input variables or characteristics of the data that the machine learning model uses to make predictions or classifications. In

this work, features are synonymous with predictors in models, and consist in one sex feature and multiple genomic features.

FL: *Federated learning* is a machine learning approach where a model is trained across decentralized edge devices or servers holding local data samples, allowing the model to learn patterns from diverse data sources.

FFN: *Feedforward neural networks* are a type of machine learning model formed of layered networks of interconnected neurons, where information flows unidirectionally, without feedback loops.

GB: *Gradient-boosted decision trees* are an ensemble learning method in machine learning that combines the predictive power of multiple decision trees sequentially, with each tree correcting the errors of the previous one.

GBP: *Guided backpropagation* is a technique used in the field of XAI that modifies the backpropagation algorithm to highlight and interpret the contributions of specific input features in a neural network's output, providing insights into the features influencing a particular prediction.

Genomic variant: A genomic variant refers to a specific alteration or difference in the sequence of DNA within an individual's genome, encompassing single nucleotide changes, insertions, deletions, and structural variants.

Genotyping arrays: Genotyping arrays are platforms used in genetics to analyze and detect variations in DNA sequences at specific genetic markers across an individual's genome, providing information about the genetic variants present in a person.

GLM: *Generalized linear models* are a class of statistical models that extend linear regression to accommodate non-normally distributed response variables.

GWAS: *Genome-wide association studies* is a statistical approach in population genomics that aim to identify associations between specific genetic variants, and traits or diseases on a genome-wide scale.

HLA: *Human leukocyte antigens* are proteins on the surface of human cells that play an important role in the immune system by presenting antigens and regulating immune responses.

HPC: *High-performance computing* refers to the use of advanced computing systems, typically involving parallel processing and large-scale computational resources, to solve complex problems or perform data-intensive tasks at a significantly faster rate than conventional computing environments.

HWE: *Hardy-Weinberg equilibrium* is a genetic principle stating that, under specific conditions including random mating and the absence of evolutionary forces, the frequencies of alleles and genotypes in a population remain constant across generations.

IMSGC: *International Multiple Sclerosis Genetics Consortium* is a global research collaborative consortium that aims to identify and understand the genetic factors that contribute to multiple sclerosis.

LD: *Linkage disequilibrium* refers to the non-random association or correlation between genetic variants at two or more loci on a chromosome, indicating that these variants are inherited together more frequently than expected by chance.

LIG: *Layer integrated gradients* is a technique in explainable artificial intelligence (XAI) that attributes the predictions of a deep neural network to its input features by integrating gradients throughout the network's layers, offering insights into the importance of each feature across different levels of abstraction in the model.

LOAD: *Late onset Alzheimer's disease* refers to the occurrence of Alzheimer's disease symptoms in individuals aged 65 or older, typically manifesting later in life, and is characterized by progressive cognitive decline, memory loss, and changes in behaviour.

LR: *Logistic regression* is a statistical method used in machine learning for binary classification tasks, where it models the probability of an event

occurring as a function of input features, and employs the logistic function to map the output to a probability range between 0 and 1 for predicting categorical outcomes.

MAF: *Minor allele frequency* is a measure in population genetics that represents the frequency at which the less common allele of a genetic variant occurs in a given population.

MCI: *Mild cognitive impairment* is a condition characterized by noticeable cognitive decline that is greater than expected for an individual's age but not severe enough to be classified as dementia.

Missense variant: A missense variant, also called a nonsynonymous variant, is a type of genetic variant in which a single nucleotide change in the DNA sequence results in the substitution of one amino acid for another in the corresponding protein, potentially affecting the protein's function.

MDR: *Multifactor dimensionality reduction* is a statistical method used in genetic epidemiology to detect and model interactions among multiple genetic and environmental factors that contribute to complex traits or diseases, helping identify high-dimensional combinations of variables associated with the outcome of interest.

MHC: *Major histocompatibility complex* is a set of genes that encode cell surface proteins essential for the immune system's recognition of self and non-self entities. This region is also known as the HLA complex in humans.

ML: *Machine learning* is a branch of artificial intelligence that involves the development of algorithms allowing computers to learn patterns and make decisions from data, enabling them to improve performance and adapt to new information over time.

MS: *Multiple sclerosis* is a chronic autoimmune disease of the central nervous system where the immune system mistakenly attacks the protective covering of nerve fibers (myelin), leading to communication disruptions between the brain and the rest of the body. This can result in

a wide range of symptoms, including fatigue, difficulty walking, numbness or tingling, and problems with coordination and balance.

NGS: *Next generation sequencing* is a high-throughput DNA sequencing technology that enables the rapid and parallel sequencing of millions of DNA fragments, allowing for comprehensive analysis of genomic information, including the identification of genetic variants, gene expression levels, and other genomic features.

PC: *Principal components* are the key axes or directions in a dataset that capture the most significant variation. In data analysis, principal component analysis (PCA) is a technique that identifies and orders these components, enabling the reduction of data dimensionality while retaining as much of the original variability as possible.

PD: *Parkinson's disease* is a neurodegenerative disorder that primarily affects movement. It is characterized by the progressive loss of dopamine-producing neurons in the brain, leading to symptoms such as tremors, rigidity, and impaired balance and coordination.

PRS: *Polygenic risk score* is a numerical assessment that summarizes an individual's genetic predisposition to a particular trait or disease based on the cumulative effects of multiple genetic variants across the genome. It is calculated by combining the weighted contributions of various genetic markers associated with the trait or condition of interest.

RF: *Random forest* is an ensemble learning method in machine learning that builds multiple decision trees during training and combines their predictions for improved accuracy and robustness. It introduces randomness by training each tree on a subset of the data and using random feature subsets, reducing overfitting and enhancing predictive performance.

RFE: *Recursive feature elimination* is a feature selection technique in machine learning that recursively removes less important features from the dataset, typically based on the coefficients or feature importance

scores obtained from a model, aiming to improve model performance and interpretability.

RFECV: *Recursive feature elimination with cross-validation* is a feature selection technique in machine learning that combines recursive feature elimination with cross-validation to iteratively identify the most relevant subset of features while assessing the model's performance. RFECV iteratively evaluates feature subsets, ranking them based on cross-validated model performance, and eliminates features until an optimal subset is determined.

RNA: *Ribonucleic acid* is a molecule composed of nucleotide units containing a sugar (ribose), a phosphate group, and one of four nitrogenous bases (adenine, guanine, cytosine, or uracil) and is essential for the flow of genetic information from DNA to protein.

sQTL: *Splicing quantitative trait loci* refer to genomic variants associated with changes in splicing patterns of transcripts, indicating a genetic influence on the regulation of gene activity.

SC: *Schizophrenia* is a severe mental disorder characterized by disturbances in thought processes, perceptions, and emotions.

SM: *Saliency maps* are visual representations highlighting the most influential regions of an input data, such as an image, as identified by a machine learning model. They are commonly used in computer vision to interpret and understand which parts of the input contribute most to the model's predictions.

SNP: *Single nucleotide polymorphism* is a type of genetic variant that occurs at a single position in the DNA sequence, where one nucleotide (A, T, C, or G) is replaced by another in a population. SNP refer to common genetic variants.

SNV: *Single nucleotide variant* is a type of genetic variant that involves the substitution of a single nucleotide (A, T, C, or G) in the DNA sequence. SNV include both common SNP and rarer mutations.

UKB: *UK Biobank* is a large-scale biomedical database and research resource that includes genetic and health-related data from over half a million participants in the United Kingdom.

XAI: *Explainable AI* refers to the development of tools applied to artificial intelligence systems in order to provide transparent and understandable insights into the decision-making processes, allowing humans to understand the rationale behind AI-driven outcomes.

Summary

Complex diseases present challenges in disease prediction due to their multifactorial nature. Unlike single-gene disorders, these diseases result from the interplay of multiple genetic, environmental, and lifestyle factors. In parallel, Machine learning (ML) and deep learning (DL) techniques have gained popularity for predicting phenotypic traits and disease conditions based on different types of clinical data, including genomic data. In this work, sometimes I refer to ML and DL as separate methods, but it is important to note that, while DL is a more specialized and sophisticated branch of ML, it still falls under the broader umbrella of ML techniques. These methods have been proven to be powerful in detecting complex patterns, including epistasis in the data. Alternatively, one of the most common methods used in population genomics to estimate the genomic predisposition to develop a disease is the polygenic risk score (PRS).

In my work I hypothesised that ML methods could be useful for classifying individuals with complex diseases, due to their ability to capture complex patterns and synergisms in the data. Consequently, I explored the prediction of four different complex diseases, multiple sclerosis (MS), Alzheimer's disease (AD), schizophrenia (SC), and Parkinson's disease (PD) using ML models with genomic data.

The primary goal of this research was to investigate the robustness and variability of the ML methods. Different models were tested to classify affected and healthy individuals, and their performance was compared. The main results of this part are summarized below:

- Logistic regression appeared to be the most robust method across folds and diseases. Alternatively, DL methods exhibited high variability across folds. These results may partially be attributed to the limited sample size available in this study, which could have favored simpler methods.

- Regarding the impact of biases present in the data, for diseases with imbalanced sex representation, the models tended to reproduce this imbalance in the predictions of the testing set,

highlighting a common limitation associated with biases in the application of ML methods.

- When comparing the performance of PRS with ML methods, PRS consistently performed at an average level. Therefore, I concluded that, with the available sample size, both methods are comparable in stratifying individuals by disease risk. However, PRS still offers several practical advantages over ML methods.

- After implementing feature selection techniques to exclude non-informative predictors from the models, the performance of ML models did not improve. This underscores the capacity of ML methods to achieve optimal performance, even in the presence of correlated features due to linkage disequilibrium.

Understanding which genomic variants are considered informative for disease discrimination during the training process could provide significant insights into the underlying genetic basis of the diseases and identify potential targets for further investigation. Related to this, the secondary goal of this study was to apply explainability tools to extract the features considered more informative by the models. The main results of this part are discussed below:

- The results confirmed the polygenicity of MS, as evidenced by the prioritized genomic features distributed across different chromosomes.

- The prevalence of HLA gene annotations among the top genomic features on chromosome 6 aligns with their significance in the context of MS.

- The highest-prioritized genomic variants were identified as expression or splicing quantitative trait loci (eQTL or sQTL) located in non-coding regions within or near genes associated with the immune response and MS.

Overall, given that ML are self-learning methods and are increasingly popular for clinical applications, this research provides a deeper understanding of how these methods learn to classify complex diseases.

1. Introduction

1.1 Genomic data

Genomic data refers to the collection of information related to an organism's genome. This data includes the sequence of DNA nucleotides, variations in the DNA sequence, and other structural elements of the genome.

Since the first initiative aimed at sequencing the entire human genome, developed between 1990 and 2004[1], known as the Human Genome Project (HGP), the main challenge for geneticists has been the generation of genomic data. This need for more data, originated from the belief that important information influencing human fate, particularly human disease, is enclosed in DNA, promoted advancements in genomic sequencing technologies.

Next generation sequencing (NGS) is a group of high-throughput sequencing technologies that can sequence millions of DNA fragments simultaneously. NGS technology started with the development of pyrosequencing and was first commercially available in 2005 as the 454/Roche platform[2]. Others followed, and as a result, nowadays there are a diversity of NGS technologies dedicated to sequence the DNA such as Illumina sequencing, Ion Torrent sequencing, Nanopore sequencing and PacBio, among others[3]. These technologies have revolutionized the field of genomics by enabling the rapid and cost-effective sequencing of entire genomes.

In addition to DNA sequencing techniques, genotyping arrays were developed in the mid-1990s using technologies such as Affymetrix's GeneChip and Illumina's BeadArray platforms. Genotyping arrays remain an essential tool in genetic research, offering a balance between cost-effectiveness, accuracy, and throughput[4]. The arrays are targeted genetic experiments allowing the analysis of an individual's genetic variation at specific known positions in their genome. They are designed primarily to assess single nucleotide polymorphisms, although other genetic variants can be interpreted, providing a general overview of an individual's genetic

makeup[5]. The differences between DNA sequencing and genotyping array technologies are outlined in Table 1.

	DNA sequencing (NGS)	Genotyping arrays
Methodology	Determine the precise sequence of nucleotides in a DNA sample, allowing for the identification of known and novel genomic variants.	Analyse predefined genomic variants across the genome using a microarray-based approach.
Cost and Throughput	While initially more expensive, they offer higher throughput and the ability to obtain information beyond predefined genomic variants, making them more cost-effective for certain applications.	Genotyping arrays are often more cost-effective when analysing a specific set of known genomic variants across a large number of samples. They require less storage and are easier to analyse.
Application	Employed in a wide range of applications, including clinical and research fields.	Commonly used in genome-wide association studies, and clinical diagnostics, where genotyping involves identifying known variants associated with specific traits or diseases.

Table 1 outlines the differences between DNA sequencing (NGS) technologies and genotyping arrays.

Thanks to the development and improvement of NGS and genotyping array technologies, the generation of genomic data has been optimized, leading to an exponential increase in the amount of such data. In fact, in recent years, there have been multiple studies generating genomic data from large-scale cohorts[6 7]. This has created a new dilemma of how to store, manage, integrate and analyze this wealth of information[6 7]. Overall, the advancements in genotyping and sequencing technologies have introduced a new era where the challenge shifts from data generation to data processing and analysis[8].

In the context of DNA, it is important to distinguish between germline, and somatic mutations. Germline mutations are the genetic alterations that occur in the reproductive cells. These mutations are passed on to offspring (inherited) during the process of fertilization and usually affect every cell in the offspring's body. Alternatively, somatic mutations are acquired mutations that arise in non-reproductive cells and are not inherited, usually

affecting cells in a localized tissue or area. In this work I focused on germline variability as the inherited risk of developing a disease. Germline information remains consistent across all endogenous cells within the same organism. Consequently, germline information in DNA can be extracted from a variety of cell tissues in the body, including saliva, using non-invasive techniques.

AGCGTCGATGGAGATT	Original sequence
AGCGT------AGATT AGCGTAGATT	Deletion
AGCGTCGATGGAGATT AGCGTCGACCATTGGAGATT	Insertion
AGCGTCGATGGAGATT AGCGTCGCTGGAGATT	Substitution (synonymous or nonsynonymous)

Table 2 lists the types of small-scale mutations.

Single nucleotide substitutions are the most well-characterized type of mutation, commonly referred to as single nucleotide polymorphisms (SNP) or single nucleotide variants (SNV). SNPs specifically refer to common genetic variants, while SNVs refer to all single nucleotide variants, including both common polymorphisms and rarer mutations. SNVs can be synonymous, meaning that the changes in the single nucleotide do not directly alter the amino acid sequence of the protein, or nonsynonymous, also named as missense mutations, which change the amino acid composition of the corresponding protein. The Single Nucleotide Polymorphism Database (dbSNP)[9] includes the annotation of both SNPs and SNVs, small insertions, and deletions listed in Table 2, assigning a unique identifier to the variant consisting in a “rs” followed by a number as detailed in Table 3 in dbSNP format.

Conversely, the integration and annotation of large-scale mutations involving large DNA segments altering the chromosomal structure, remains a challenging task. Primarily, these alterations are not easily

detected with most sequencing technologies or the subsequent analysis pipelines. Additionally, the identifiers assigned to complex mutations are less standardized. Therefore, the identification of these alterations is hindered by the lack of evidence and the absence of harmonized identifiers.

Format	Description	Example	Protein-Coding	Non-Coding
Protein format	Based on the protein reference sequence	APOE_ p.R176C APOE_ p.R202C	Yes	No
Coding format	Based on the coding DNA reference sequence	APOE_ c.526C>T APOE_ c.604C>T	Yes	No
Genome format	Based on the whole DNA reference sequence	GRCh37: APOE_ g.45412079C>T GRCh38: APOE_ g.44908822C>T	Yes	Yes
dbSNP format	dbSNP Reference SNP (rs or RefSNP) number	rs7412	Yes	Yes

Table 3 illustrates the different variant annotation formats for single nucleotide substitutions located in coding regions. Columns "Protein-coding" and "Non-coding" indicate if the annotation format allows for the representation of variants located in these regions, respectively.

There are different types of genomic identifiers assigned to genomic mutations, and general recommendations for variant annotations exist[10]. However, different formats are used across datasets, sometimes complicating the pre-processing and analysis of genomic data. There are four main variant annotation formats described in Table 3, and each database uses one or several of them. The following conflicts arise when linking variants across datasets:

- In the column "Example" of Table 3, there are several IDs using the same format in protein, coding, and genome format. For example, *p.R176C* and *p.R202C*, in the case of protein format, represent the same variant in different protein isoforms. In the coding format, *c.526C>T* and *c.604C>T* refer to the same variant in different transcripts of the same gene. Finally, in the genome format, variant annotation depends on the reference genome version (GRCh37 or GRCh38).

- Protein format and coding format do not represent variants in non-coding regions.
- dbSNP format only includes previously described variants. Rare variants are sometimes missing in this database.

In this project, variant annotations from dbSNP were used, focusing on small-scale genomic variants, including small insertions, deletions, and SNVs as described in Table 2. This decision is supported by the design of the UK Biobank Axiom Array[11], used as the primary source of genomic data in this study. This array is designed to capture predominantly common SNVs and well-known disease-causing mutations that are usually represented in dbSNP. Additionally, dbSNP annotation facilitates the annotation of genomic variants in non-coding regions.

Some regions in the human genome adhere to special annotation conventions. This is the case of the region on chromosome 6 corresponding to the major histocompatibility complex (MHC), known as the Human leukocyte antigen (HLA) in humans. Due to the high variability and complexity shown in this region, it has its own nomenclature system for different alleles. All alleles start with “HLA”, followed by the name of the gene and an asterisk. The first two digits after the asterisk specify the serologically defined allele group, and together with the third and fourth subsequent digits, they indicate a unique protein sequence[12]. As an example, the allele *HLA-DRB1*15:01* refers to the allele group 15, and the specific protein 1, within the *HLA-DRB1* gene.

1.2 Inheritance patterns and population genomics

In the 19th century, Gregor Mendel experiments with pea plants allowed to explain patterns of inheritance and provided a framework for understanding the transmission of genetic traits[13]. The three fundamental principles formulated by Gregor Mendel include:

- Principle of dominance and uniformity: Alleles can be either dominant or recessive. An organism with at least one dominant allele will show the effect of the dominant allele. The principle of uniformity is represented in Figure 1.

- Principle of segregation: There are two alleles for a given trait, one inherited from each parent, and the alleles segregate during the formation of gametes. Consequently, each gamete carries only one allele for each gene. The principle of segregation is also represented in Figure 1.
- Principle of independence assortment: Different pairs of alleles for different traits are inherited independently of each other.

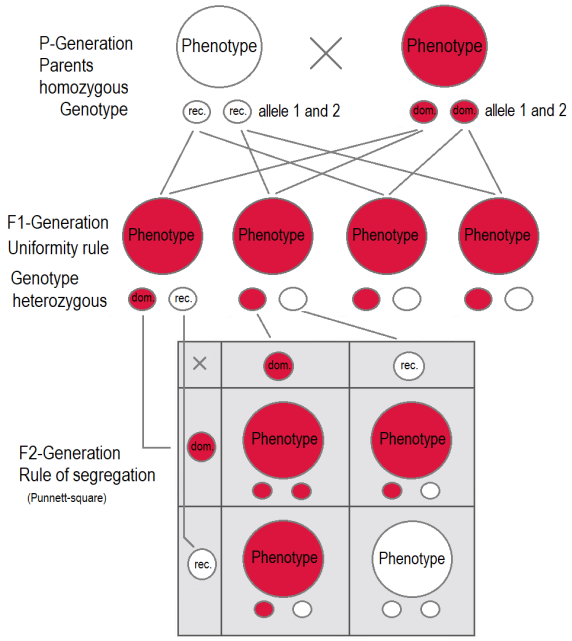

Figure 1 represents two of the Mendel's principles, the principle of uniformity and the principle of segregation. The figure was obtained from reference[14]. 'P' corresponds to the two pure-breeding parental generations involved in a particular cross. 'F1' and 'F2' represent the first and second generation of the 'P' cross.

Simple Mendelian inheritance help clinicians and researchers in the diagnose and risk prediction of some human diseases[15 16]. Family-based genetic studies, for example, involve the study of genetic variants within families to identify the genetic basis of various Mendelian traits and diseases[16]. While Mendel's principles have several exceptions and apply

only to simple Mendelian traits, understanding the Mendelian inheritance provides a foundation for studying more complex patterns of inheritance, such as those involving multiple genes (polygenic traits) and interactions with environmental factors[17].

Germline mutations accumulate over time and evolution and are submitted to evolutionary pressure. This means that mutations in sensitive areas of the genome will not result in a successful embrio, and this footprint has been used in evolutionary genomics to identify the areas of the genome which are more preserved, and therefore likely more sensitive to changes[18 19]. The initiatives sequencing large cohorts with diverse species and human populations boosted the knowledge on this field[18 20]. In this regard, protein-coding regions exhibit stronger evolutionary constraints compared to non-coding regions[20].

However, over the evolution there has also been the accumulation of genomic variants with small to medium effects, that are located in less sensitive areas, and are apparently benign by itself, but under certain circumstances related to the environment or the presence of a certain genomic background, become predisposing or protecting against diseases or conditions[21]. In fact, many genomic variants listed in curated databases as associated with human diseases are located in non-coding regions, without any predicted impact in the protein sequence or structure[22 23]. This trend exemplifies how the protein-coding DNA sequence alone is not always deterministic of what is going to be manifested in the phenotype. In some cases, genomic variants in non-coding regions have a regulatory role affecting the transmission of information to the RNA and Proteins[22]. In addition, gene transcription can be altered by epigenetic factors, which play an important role in regulating various cellular processes, and can be influenced by both, genetic and environmental factors.

Within the non-coding genome, evolutionary constrained regions are usually associated with known regulatory elements and variants linked to complex human diseases[20]. Partially due to the regulation of gene expression, that depends on the interactions across various loci, for most of the diseases, rather than having a single strong genetic determinant, various genomic loci simultaneously show different degrees of association

with a disease or trait[24] [23]. Consequently, using the word causality to speak about a mutation in the DNA sequence coding for an RNA that directly determines the protein structure and subsequent phenotype is a simplification that cannot be extrapolated to all the traits.

One important property of genomic data is the presence of linkage disequilibrium (LD), which is associated with the phenomenon of recombination and affects the way DNA replicates during meiosis, as depicted in Figure 2. During the process of recombination, which occurs during the formation of gametes (sperm and egg cells), segments of DNA from each parent are exchanged. Recombination typically breaks down associations between alleles at different loci, promoting genetic diversity. However, genes or genetic markers that are physically close to each other on a chromosome are more likely to be inherited together without undergoing recombination. Consequently, over generations, the genetic variants in physical proximity tend to co-occur more frequently than expected by chance, leading to an elevated correlation. This complex interdependence among genetic variants complicates the interpretation of individual variant associations[23].

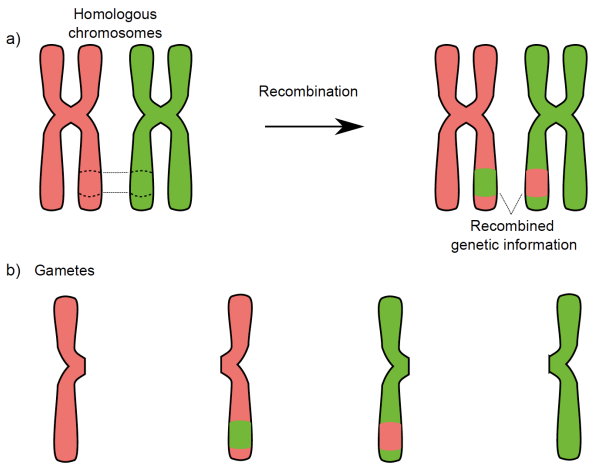

Figure 2 represents the process of recombination in (a) and the subsequent representation in gametes in (b). Linkage disequilibrium originates from the tendency of nearby genomic regions to be inherited together, despite recombination. The figure is based on this reference[25].

Population genomics is the field of study that tries to disentangle the complexity of the human genome above described by analyzing genomic information from many individuals to study patterns of genetic variation within and between populations.

One example is genomic imputation, a widely used method in population genomics to infer missing genotypes. As previously noted, genotyping arrays can only capture a predefined set of genomic variants. Additionally, genotyping arrays and NGS generate missing values due to technical limitations. To address these issues, the combination of genetic principles, such as LD, available reference panels, computational advancements, and statistical algorithms has made genomic imputation a viable and widely used approach in genomics research. It allows researchers to infer and predict genotypes at unobserved variants, contributing to a more complete view of the genomic landscape. Tools such as IMPUTE5[26] and BEAGLE[27] can impute missing variants based on population-specific reference panels.

Nevertheless, there are several regions of the genome where general imputation estimates are inaccurate, and custom imputation approaches are needed. An example is the region on chromosome 6 corresponding to the HLA complex, considered the most variable (polymorphic) region of the human genome[28]. The complex structure in this region often requires the use of specific methods for imputation, such as SNP2HLA[29], HIBAG[30] or HLA*IMP[31].

In recent decades, population genomics has also made advances in the genomic characterization of diseases, largely due to the improvement in genomic technologies, the increase in the available genomic data, and the emergence of genome-wide association studies (GWAS)[23]. GWAS are based on the analysis of genetic variants across the entire genome of many individuals to identify associations between specific genetic markers and particular traits or diseases[23]. This tool has been used to identify genetic variants that contribute to the risk of complex diseases, with small to large effects. The number of associated genomic variants with human conditions reported by GWAS is expected to grow as sample sizes increase in the future.

There are several steps in the GWAS analysis, including the collection of samples from large cohorts, genotyping using arrays or DNA sequencing techniques, quality control of the genomic data, imputation of missing genomic variants, conducting the statistical test for association in the discovery cohort, and finally, seeking an independent replication with a validation cohort. The primary output of GWAS analysis, known as summary statistics, consists in a list of *p-values*, effect sizes and their directions associated to the tested genomic variants with respect to the phenotype of interest.

While the utility of GWAS is undeniable, there are several limitations associated with the use of this technique. First, the presence of LD makes it challenging to identify the exact causal variant. In this regard, GWAS report blocks of correlated SNVs in LD that have a statistically significant association with the trait of interest, rather than single SNVs. Therefore, this can lead to false positives and false negatives when a genomic variant is in LD with a causal variant, and the associated signal is attributed to the non-causal variant in LD. Related to this, the findings from GWAS in one population may not easily generalize to other populations due to differences in LD patterns.

Researchers often use fine-mapping tools to try to address issues derived from LD and find the causal SNVs in GWAS signals. There are several fine-mapping approaches based on Bayesian models, including CAVIAR[32], FINEMAP[33], PAINTOR[34] and SuSIE[35]. However, the set of genomic variants selected by Bayesian models is not always consistent across methods[23].

Another approach is the conditional association analysis. In this method, the genetic variant that shows the most significant association with the trait of interest in the initial GWAS analysis, named as lead variant, is added as a covariate in genotype-phenotype regression models, while the association between the trait and other variants in the region is evaluated. If additional variants in the region still show significant associations after conditioning on the lead variant, this suggests that multiple, distinct genetic effects contribute to the trait of interest[23].

Despite of the available tools, prioritizing the causal SNV over highly correlated SNVs in LD using fine-mapping methods is challenging. In this context, including diverse ethnicities in GWAS can enhance the fine-mapping task. This is because the differences in LD structure among ancestries could aid in limiting the size of LD blocks associated with a certain phenotype. Therefore, initiatives aimed to increase the diversity in GWAS discovery cohorts have the potential to facilitate the identification of new causal SNVs in the future[23].

The multiple testing problem is another limitation affecting GWAS that arises from the vast number of statistical tests conducted simultaneously across the entire genome. This problem is related to the type I error rate, as the more tests conducted, the higher probability of obtaining false positives by chance exists. This problem is particularly present in GWAS of complex diseases, that tend to have many SNVs with small effects contributing to the disease, and some SNVs with an effect size close to the threshold may be overestimated. The gold standard for addressing false discoveries in GWAS is to compare the effect sizes of SNVs between the discovery cohort and an independent replication cohort[23].

Finally, in GWAS, associations between SNVs and the phenotype are commonly tested using linear regression models for continuous phenotypes, or logistic regression models for binary phenotypes. Therefore, while GWAS is effective in uncovering the main effects of genomic variants within LD blocks concerning a particular condition, it is less suited for detecting interactions between genomic variants, commonly referred to as gene-gene interactions or epistasis influencing disease risk. To address this limitation, several statistical methods have been developed to discover non-Mendelian disease transmission involving genomic interactions[36]. Methods such as multifactor dimensionality reduction (MDR)[37] [38], AprioriGWAS[39], fpgrowth[40] [41], several Bayesian methods[42], and machine learning methods[36] offer this possibility.

In population genomics, the most popular statistical approach used to quantify the genomic predisposition of individuals to develop a trait or disease is the polygenic risk score (PRS). The assumption behind this

method is, rather than focusing on the exact causal signal in the DNA associated with the disease, consider the sum of many genetic effects independently associated with the condition. In its core equation, PRS is calculated with a weighted sum of the effects sizes obtained from GWAS summary statistics. Summing all these effects across the genome gives a score indicating the risk propensity to a disease.

There are several methods for computing PRS, one of them being *clumping and thresholding* (C+T). This method is implemented by selecting the SNVs with increasing *p-value* thresholds of association with the trait, and reducing the number of correlated SNVs in LD through clumping[43] [44]. As an example, PRSice2 is a package that enables PRS calculation using the C+T method[45].

PRS has been extensively used in many studies, demonstrating its ability to extract disease risk propensity scores from diverse cohorts and conditions[46] [47] [48] [49]. However, a limitation still exists, as PRS are not designed to detect complex patterns and epistatic events between genomic variants associated with a disease or condition.

GWAS results are updated regularly in databases such as the GWAS Catalog[50] and dbGAP[51], providing the summary statistics required for the PRS calculation. In addition, ClinVar[52] and DisGeNET[53] represent publicly accessible databases that compile data regarding genetic variants and their connections to human traits and diseases, obtained from a diverse range of sources, including published GWAS studies.

Overall, GWAS and PRS serve as valuable resources for gaining insights into the spectrum of genomic variants linked to specific medical conditions. Yet, GWAS and PRS are limited in their ability to account for the synergistic effects caused by various genomic loci and lack specificity due to LD. Consequently, selecting genetic determinants for follow-up in laboratory and clinical studies remains a challenge, and some of the mechanisms in which predisposing and protective genetic alterations contribute to complex diseases are still unknown.

1.3 Machine learning and deep learning methods

Machine learning (ML) and deep learning (DL) are subfields of artificial intelligence (AI) that focus on the development of algorithms and models that enable computers to learn and make predictions or decisions without being explicitly programmed for a particular task. ML is a broader concept that includes various techniques where systems learn from data and improve their performance on a specific task over time, and DL is a subset of ML that focuses on neural networks with multiple layers (deep neural networks). The emergence of big data has facilitated the application of these methods, contributing to their increasing popularity in a wide range of fields, including population genomics[54].

Training is the phase during which the ML model learns, meaning the weights or coefficients of the model are adjusted based on the training data. Testing is the phase when the model is evaluated, making predictions on new unseen data that should be representative of the target population, also known as testing data. Overfitting is a concept in ML and statistics where a model learns the training data too well, capturing biases and random fluctuations rather than just the underlying patterns associated with the trait of interest. This can lead to a model that performs very well on the training data but fails to generalize effectively, resulting in poor performances on testing data. One measure to avoid overfitting is implementing a proper strategy to divide samples during the training and testing steps.

In order to train and test ML models, the usual practice is to divide samples in training, validation, and test sets. The samples in the training set are used to train ML models. Samples in the validation and test sets are used to test the model in the process of selecting the best parameters (hyperparameter selection), and in the final evaluation, respectively. There are several strategies to split samples, and usually several rounds of training, validation, and testing are required to try different combination of parameters (hyperparameter configurations) until reaching the optimum model. K-fold cross-validation (CV) is a robust splitting approach to train models that consists in splitting the sample size in K folds, using K-1 folds for training and validation, and the remaining fold for testing. The training

and testing are repeated as many times as the number of folds, using a different fold for testing in each iteration.

Nested cross-validation (nested CV) is an adaptation of the K-fold CV that consists in setting one outer loop and one inner loop of CV. This strategy has proven to be useful in reducing overfitting and correctly estimate the variance of the models[55]. In this approach, the CV in the inner loop is performed on the training set of the outer loop and is used to select the optimum hyperparameter configuration. Conversely, the CV in the outer loop is used to train the final model with the selected hyperparameter configuration obtained from the inner loop, and to test the model with the remaining test set that has not been used for hyperparameter selection or training the model. Iterating over different folds in the inner and outer loop allows for the use of different samples in training, validation, and testing in each iteration, optimizing the use of all the available samples. At the end, nested CV generate as many final models as number of folds in the outer loop. A representation of the nested CV used in this study is provided in Figure 3.

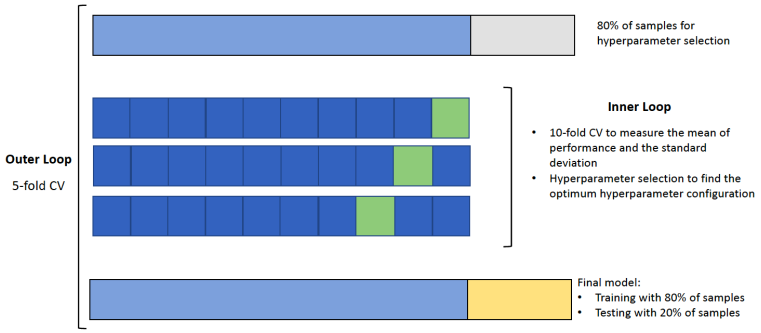

Figure 3 represents the nested CV approach used in this work, which consists in 10-fold CV for the inner loop, and 5-fold CV for the outer loop.

When the final model has been trained and predictions on the test set give optimum results, an additional measure to ensure there is no overfitting and that the model generalizes well is to test the model on an external validation dataset from an independent cohort not used in training or

testing. Assessing the model's performance on external validation datasets provides a better understanding of its generalization capabilities.

There is a diversity of ML methods that differ in the learning strategies and architectures employed to learn from data and make predictions. In the following lines, I describe the different ML methods used in this work:

1.4.1 Logistic Regression (LR)

LR models the probability using the sigmoid function, defined in Figure 4(a). The linear combination x is defined with the formula in Figure 4(b) where b_0 is the bias or intercept term, and $b_1, \ldots, b_n$ are the coefficients associated with the input features. The model is trained by optimizing the coefficients to minimize the negative log-likelihood or the cost function. This optimization process typically employs algorithms such as large-scale bound-constrained optimization (lbfgs)[56] [57], stochastic average gradient (SAG)[58], or stochastic average gradient accelerated (SAGA)[59]. The decision boundary, which separates regions corresponding to class 1 and class 0, is determined by the threshold of the logistic function. This threshold creates an hyperplane in the feature space as shown in Figure 4(c).

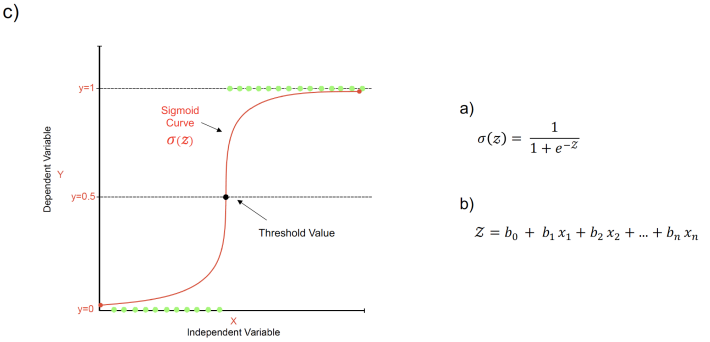

Figure 4 illustrates the formulas constituting the logistic function in parts (a) and (b), along with a representation of the hyperplane in (c).

Tree-based Ensemble Methods

Ensemble methods combine the predictions of multiple individual models to create a more robust and accurate predictive model. Three ensemble ML methods that combine multiple decision trees to make predictions were used in this study:

1.4.2 Gradient-Boosted Decision Trees (GB)

GB is an ensemble learning method that combines the power of decision trees and boosting algorithms[60]. The ensemble technique implies that multiple weak classifiers, in this case shallow decision trees, are combined to create a stronger predictive model. The boosting algorithm involves the sequential training of weak models, giving more weight to misclassified instances in each iteration. Subsequent models focus on correcting errors made by the previous ones. The final prediction is made by aggregating the predictions of all weak learners in the ensemble. The GB process is shown in Figure 5.

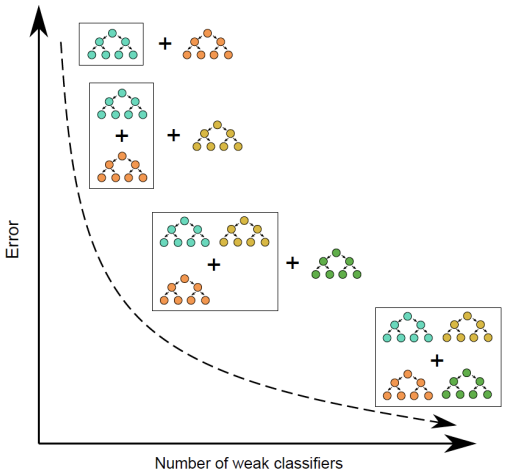

Figure 5 illustrates the working principle of GB, demonstrating the sequential addition of weak classifiers and the gradual reduction of errors.

1.4.3 Random Forest (RF)

In RF[61], multiple independent decision trees are built in parallel with different subsets of the data. The final prediction is then an aggregation of the predictions of all individual trees. For each feature under consideration at a split point in the decision tree, RF selects the optimal split point.

1.4.4 Extremely Randomized Trees (ET)

ET is very similar with RF, with the difference that, for each feature under consideration at a split point in the decision tree, the random, instead of the optimal split, is applied[62]. This introduces additional randomness and reduces the variance of the model, making ET less sensitive to noise in training data compared with RF.

The process of RF and ET is represented in Figure 6. In summary, GB, RF, and ET are all ensemble tree-based methods. However, there are some key differences between these methods:

- GB trains decision trees sequentially, where each subsequent tree is built to correct the mistakes made by the previous trees.
- ET is similar to RF in building multiple decision trees in parallel. However, it differs in the way it selects the splitting points in the decision tree. ET randomly selects splitting points, while RF applies the optimal split.

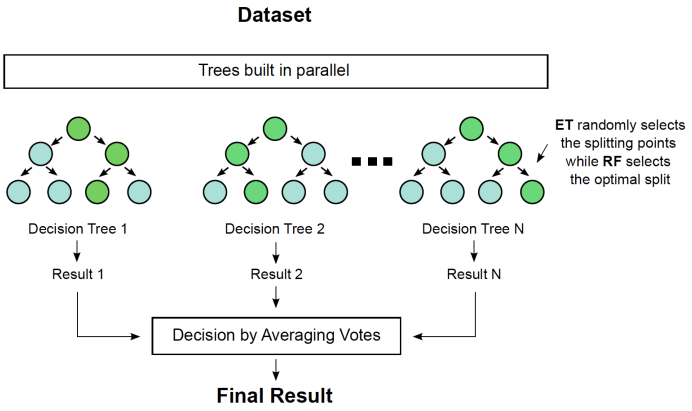

Figure 6 illustrates the working principle of RF and ET.

Deep learning methods (DL)

DL methods process input data through multiple layers of interconnected nodes, applying activation functions and adjusting weights during training to learn and make predictions on various tasks. The term "deep" refers to the depth of the network, which allows it to automatically learn hierarchical representations of data[63].

1.4.5 Feedforward Networks (FFN)

FFN falls under the category of DL methods. The following aspects are key characteristics of this method:

- Regarding the architecture, FFN begins with an input layer where each node represents one of the features of the input data, as represented in Figure 7 with yellow nodes. The following are one or more hidden layers represented in Figure 7 as grey nodes. Each hidden layer consists in a variable number of nodes that are connected to all the nodes in the adjacent layers. The final layer, also known as output layer, produces the network's output and is represented in Figure 7 with a red node. The number of nodes in the output layer depends on the nature of the task. In the case of this study, where I used FFN as binary

classifiers, there is a single node in the output layer providing the raw logits, which are subsequently converted to probabilities with the sigmoid function. These probabilities indicate the strength of association with the binary class.

- The concepts of "width" and "depth" are terminologies that refer to different aspects of the network architecture. Width refers to the number of nodes in a single layer, and depth refers to the number of layers in the network, including the input and output layers. Increasing width and depth in FFN is associated with more capacity to capture complex patterns in the data. However, it also requires more computational resources and is prone to issues like vanishing gradients and overfitting. In my work, I tried different values of width and depth in the hyperparameter selection step until reaching the optimum hyperparameter configuration.

- Nodes in the same hidden layer are independent of each other. Each node in a hidden layer receives inputs of all the nodes in the previous layer and produces outputs to all the nodes in the next layer. The operation that happens inside each node consists in a dot product, also referred to as a weighted sum, and an activation function that transform the input into a non-linear output. There are several activation functions, in my work I used the leaky rectified linear unit (leaky ReLU).

- Forward propagation is the information flow from the input layer through hidden layers, to the output layer. The loss function measures the difference between the predicted outputs and the actual target values. There are various loss functions, and the one I used in this study is the binary cross entropy with logits loss (BCEWithLogitsLoss).

- During backward propagation, the weights are adjusted based on the average outputs of the loss function, and the optimizer algorithm adjusts the weights accordingly. In this work I used the Adam optimizer, which was set to work with different learning rates during hyperparameter selection.

- During training, forward and backward propagation are repeated multiple times, and the goal is to minimize the results of the loss function. An epoch corresponds to one round where the entire dataset goes through forward and backward propagation. If the number of epochs is set too small, the model may not learn the representations in the data and, therefore, will provide poor predictions. Conversely, if the number of epochs is set too large, the model risks overfitting. In my work, I tried different number of epochs in the hyperparameter selection step.

By iterating through the training data and updating the weights, the FFN gradually learns to make better predictions and generalizes to new, unseen data.

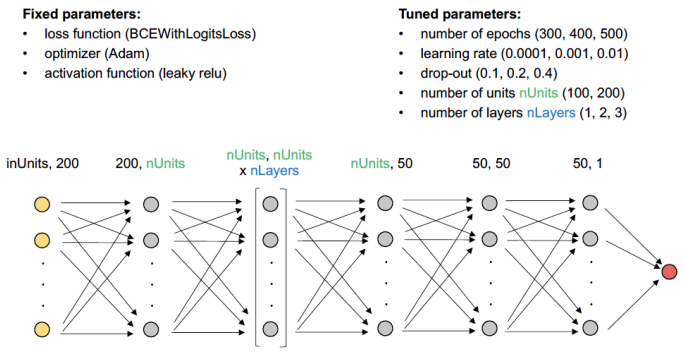

Figure 7 represents the architecture of the FFN employed in this study.

1.4.6 Convolutional Neural Networks (CNN)

CNN also fall under the category of DL methods. There are three types of layers in a CNN, as depicted in Figure 8:

- In convolutional layers, filters (kernels) are employed to create feature maps. The convolution process involves applying a kernel to values within a sliding window, represented as blue squares in Figure 8, steps 1) and 3). Kernels begin randomly and learn during training by tuning their weights in the convolution step. Each channel in the convolutional layer correspond to a feature map generated by a single kernel, highlighting a specific pattern or feature present in the data. In the case of images, for example, kernels work as filters that extract features, such as objects, from different regions. During convolution, deeper into the model, the ideal scenario is to increase the number of channels (feature maps) and decrease the number of features. This makes the representations increasingly abstract, and each layer has a higher receptive field (see more of the image) as we go deeper into the model.

- After convolution, a downsampling step (in this work, “pooling”) is usually added to decrease the number of features in each channel, averaging the values in a window, depicted as green squares in Figure 8, steps 2) and 4). This is because convolution increases the dimensionality very quick, and in this context, pooling serves to control the dimensionality of the data and detect relevant features in a spatial area. Therefore, in pooling layers, the goal is to reduce dimensionality and increase the size of the receptive field.

- CNN layers are designed to generate feature maps rather than making predictions. In this study, CNN models were created adding two steps of convolution and pooling before the FFN, responsible for generating the final output

and making predictions. As a result, the final layers of the CNN adopt the architecture previously described for the FFN, illustrated in Figure 7. The representations obtained from the convolution and pooling steps are flattened and serve as input to the FFN input layers.

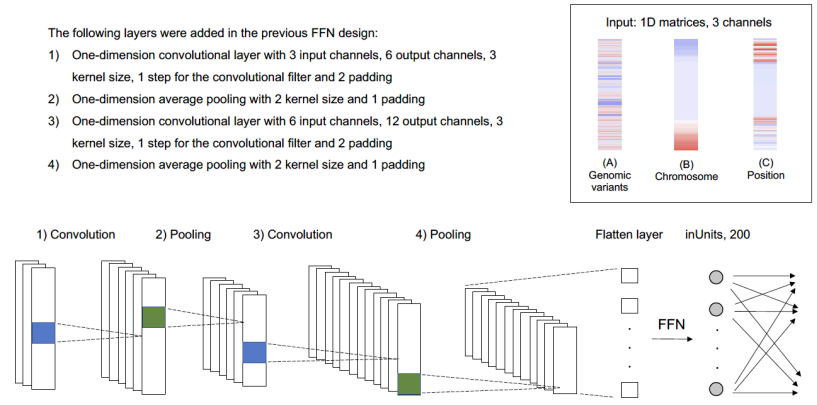

Figure 8 illustrates the architecture of the one-dimensional CNN used in this study.

In summary, the main difference between FFN and CNN is that, in the case of FFN, there is a one-to-one mapping between a feature and an input unit. However, this approach is not always robust. For instance, in image processing, focusing on individual pixels may not be optimal. Instead, it is more effective to use a CNN to increase the receptive field and group pixels that together form a specific component, such as a nose, an ear, or an eye in a facial image. This grouping allows the network to map these composite features into a single unit, enhancing its ability to recognize meaningful patterns in the data.

As input variables, ML methods can accept a list of genomic variants without any prior assumptions about the genetic contribution to the traits. The methods themselves calculate the importance of these variants during the learning step. In this regard, the genomic patterns of correctly classified individuals learned by different ML methods in the training step

can provide valuable information about the effects of genomic variants and foster the discovery of distinct genomic profiles associated with diseases or conditions. This is because one important aspect of ML is their ability to detect interactions and complex patterns in the data[64 36 65]. Also, in clinical diagnostic processes, it is important to understand which features contribute to the model's output and the associated biological pathways. Therefore, when using ML models to address sensitive problems, the use of black-box models with a lack of transparency can be problematic, and is important to identify the rules that lead the model to achieve correct or optimum performance.

A possible solution to this limitation is the Explainable AI (XAI), which is a field in data science that aims to improve the understanding of AI models, including ML models, by using different interpretability algorithms[66 67 68]. The use of XAI methods in ML may reveal genomic variants involved in complex patterns and epistatic events. In this respect, the application of XAI in FFN and RF, has been used to detect interactions between gene loci, as well as between gene loci and environmental factors related to disease status[69 70]. For this work, I selected ML methods that allow the application of XAI techniques.

One of the main limitations of self-learning methods such as ML is that, given that these methods learn from data, they tend to reproduce biases present in the training data in the test dataset. This poses a problem if ML models make decisions entirely based on biases rather than relevant features. For instance, it has been reported that AI systems tend to reproduce racial and gender imbalances when trained with real-world data[71]. In this respect, it is important to detect these trends and apply appropriate balancing strategies in the training data to ensure an equal representation of different groups when possible.

Another limitation of self-learning methods, especially DL, is the requirement for large datasets in order to train the models to be robust. The storage of health-care data, such as genomic data, is protected under specific privacy clauses, and the distribution of such data is not always possible. A potential solution to this limitation could be federated learning (FL), which is a ML approach that enables training models across multiple

decentralized devices without exchanging raw data[72]. In traditional ML, data is typically collected and centralized in a single location for model training. However, FL takes a different approach by allowing model training to occur locally on individual devices. This is done by sending local model updates to a central server, which aggregates the updates and then sends back the updated model to the devices. This process is repeated multiple times until the model converges. FL is a relatively new technique, and its application in building models with genomic data is just beginning to gain traction[72 73 74].

There are several challenges associated with the application of FL. First of all, FL typically require high traffic between the devices and the central server, which can be a bottleneck in some applications[75]. Also, FL is not immune to privacy attacks. For example, an attacker could try to infer information about the data on a device by analyzing the local model updates[76]. Additionally, FL can be challenging to implement when the data on the devices is heterogeneous[77]. Although I did not use FL in my work, I describe this method here as it will be relevant in the future perspectives of ML methods developed in the final discussion.

1.4 Description of the diseases under study

Complex diseases are a diverse group of conditions that often pose challenges in accurate diagnosis due to their phenotypic heterogeneity. Affected individuals may exhibit a wide range of symptoms and varying disease severity, leading to frequent misdiagnoses. Therefore, the identification of robust biomarkers becomes crucial in facilitating early detection and appropriate clinical treatment from the onset of initial symptoms in these diseases. In this regard, genomic data has emerged as a promising tool in advancing methods for disease detection and treatment of complex diseases.

In the context of genetics, complex diseases or conditions arise from the combined effects of multiple genomic variants and genes, are influenced significantly by both the physical and the social environment, and display non-Mendelian inheritance patterns. As a result, the task of finding genomic factors that contribute to the predisposition or protection against

these diseases is not trivial. It often requires the use of large cohorts with comprehensive genomic information at high density to obtain statistically significant results.

The complex diseases used in this study belong to the ICD-10 categories "F" (Mental, Behavioral and Neurodevelopmental disorders) and "G" (Diseases of the nervous system) with over 900 cases identified in UK Biobank (UKB).

1.5.1 Multiple sclerosis (MS)

Multiple sclerosis (MS) is a chronic inflammatory and neurodegenerative disease of the central nervous system. It is also considered an autoimmune condition where the immune system attacks the myelin sheath, which is the layer that surrounds and protects the nerve cells. This disease can cause a wide range of symptoms such as fatigue, limited vision and mobility problems, among others. Each person with the condition is affected differently. MS affects 2.8 million people worldwide and is more common in Caucasian populations. The mean age of onset is between 20 to 30 years, approximately affecting three females for every one male, with a sex bias also in clinical course[78].

The aetiology of the disease is multifactorial, involving many genes, predominantly immune system genes. In fact, MS is considered a highly polygenic disease, and genomic variants in the HLA region, located on chromosome 6, have the strongest signal in GWAS studies[79 80]. Additionally, environmental factors such as vitamin D deficiency have also been associated with the disease[81]. In recent years, several studies have provided evidence of Epstein-Barr virus (EBV) infection predisposing to MS[82 83 84 85]. EBV may cause MS through the reprogramming of latently infected B lymphocytes and the chronic presentation of viral antigens, which trigger autoreactivity through molecular mimicry of the *Epstein-Barr nuclear antigen 1* viral protein and the *GlialCAM* human endogenous protein[86 87]. Despite all the

advancements made in research, there is no cure for the disease, and current treatments are directed towards improving recovery from attacks[88].

The International Multiple Sclerosis Genetics Consortium (IMSGC) is a research collaboration composed of members from academic institutions and research centers worldwide dedicated to studying the genetic factors that contribute to MS susceptibility and progression. IMSGC has identified many genetic regions contributing to MS susceptibility applying GWAS on large cohorts [89] [90] [80]. IMSGC also performed the largest GWAS meta-analysis on MS to date[91], analyzing data from 47,429 people with MS and 68,374 control subjects, and they established a reference map of the genetic architecture of MS that includes 200 autosomal susceptibility variants outside the HLA region, one chromosome X variant, and 32 variants within the extended HLA region. These studies provided evidence for a polygenic component to the genetics of MS, and the presence of a cumulative effect of multiple genetic variants scattered across the genome, each contributing only a modest individual effect. However, while GWAS studies have identified many genomic loci associated with MS, the functional relevance of some of these loci remains to be fully elucidated.

PRS has been used in MS demonstrating its utility in understanding MS susceptibility, severity, and prediction. A recent study conducted PRS analysis on MS to assess the associations of the genomic background with both disease status and severity in cohorts of European descent. The study found that individuals within the top 10% of PRS were at greater than five-fold increased risk of developing MS in UKB[92]. Also, the inclusion of PRS in clinical risk models increased the risk discrimination by 13% to 26% over models based only on conventional risk factors. Conversely, another study demonstrated that the PRS developed for MS using an European population performed poorly in predicting MS risk within the South Asian-ancestry population,

highlighting the importance of developing population-specific PRS[93].

In another study, authors aimed to identify genetic loci linked to the progression of disability in individuals with MS by applying ML methods[94]. To achieve this, the authors used RF and gradient boosting machine (GBM) models, alongside a mixed-effect ML platform. This hybrid approach merged the strengths of RF and GBM, incorporating generalized mixed-effects regression trees. The primary goal was to effectively identify individuals with MS who were prone to experiencing a deterioration in their condition in the future, and obtain the genomic profiles captured by the ML methods that enhanced the identification of cases. The investigation focused into 208 well-established loci related with disease progression and extracted genetic decision rules from the ensemble models. Finally, the study pointed to seven genetic loci that displayed an association with an elevated risk of MS disability worsening.

Regarding the application of DL methods in MS risk prediction, a recent study[95] used an artificial neural network (ANN) model, which is a subcategory of DL methods that includes FFNs. The authors aimed to predict MS risk using genetic data from 401 MS patients and 390 healthy subjects. The locally interpretable model-agnostic explanation (LIME) was used to explain model predictions.

The epistatic events among genomic variants associated with MS have been studied in several works. Researchers in a published work[96] used a penalized regression incorporating elastic net with a stability selection method by iterative subsampling to detect potential interactions of loci associated with MS risk. This approach identified new association loci for MS predisposition. Alternatively, researchers in a recent publication[97] used an approach called association rule mining (ARM) applied to individuals with MS and controls to discover genomic patterns amongst the known MS risk variants. They aimed to uncover

patterns of gene-gene and gene-environment interactions associated with MS risk. Researchers concluded that certain combinations of MS risk variants are linked to an increased risk of developing the disease.

1.5.2 Alzheimer's disease (AD)

Alzheimer's disease (AD) is a neurodegenerative condition and the most common form of dementia. It is characterized by symptoms such as memory loss, language deficits, disorientation, mood changes, and in advanced stages provokes the loss of body functions and death[98]. AD is found in about 1 in 8 people aged 65 to 74, reaching almost half of people over 85 years old [98]. From the first official report of AD until 1977, the diagnosis of the disease was reserved for individuals between the ages of 45 and 65 who developed symptoms of dementia[99]. Nowadays, these cases roughly represent 5% of the total diagnosed AD and are named early onset Alzheimer's disease (EOAD)[100]. With the general increase in life expectancy, the disease has become more prevalent in the population above 65 years old, representing around 95% of the total AD cases, termed late onset Alzheimer's disease (LOAD)[101]. Although there are differences in the age at onset, progression time and genetic background, a similar pathological process is observed in both forms of the disease[102].

At the pathophysiological level, AD is defined by the accumulation of anomalous folded *Amyloid beta* protein outside neurons and the abnormal aggregation of the *Tau* protein inside cells[103]. These two events lead to the loss of neurons and synapses in the cerebral cortex and certain subcortical regions, promoting the cognitive impairments perceived in AD patients[103]. The altered biological pathways causing AD are not yet fully understood, and the disease still has no cure.

In the case of LOAD, heritability is estimated to be around 58% to 79%[104]. LOAD appears to have a polygenic nature, with genetic risk being predominantly influenced by *APOE*, acting on top of a

highly polygenic background[105] [106]. In this regard, GWAS focused on AD have identified more than 30 different susceptibility loci that are associated with the disease in European populations[107] [108] [109] [110] [111]. Genes in these loci play roles in *Amyloid* and *Tau* pathways, lipid-related processes, immune response, and microglial function.

PRS have been applied to AD, stratifying by the major *APOE* risk alleles, showing to be significant predictors of age-specific risk for the disease [112] [113] [114]. In this context, PRS could be used to detect asymptomatic individuals with the greatest probability of developing AD in the near future. It is worth to note that despite the elevated occurrence of AD in individuals of African and Hispanic ancestry relative to those of European or Asian ancestry, the majority of GWAS studies have been conducted within European populations, leading to potential biases in PRS[49].

ML classifiers have been previously used to classify AD using genotyping data. Authors in a published work[115] examined the application of six different ML methods, including RF, to predict the risk of LOAD using genomic data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The research systematically compared various ML models and found that the best-performing models achieved around 0.72 Area Under the Receiver Operating Characteristic Curve (AUC-ROC), indicating their potential for predicting LOAD risk. Another study contrasted different ML methods and proposed enhancing prediction by adding markers from misclassified samples[116].

There are also published works using DL methods with genomic data to classify individuals with AD. A recent study, for example, tried to address the problem of the limited population representativity by creating a DL-based framework designed to enhance the accuracy of genetic risk prediction by incorporating data from diverse populations[117]. The method, named as DisPred, employs two key components: a disentangling autoencoder approach to separate the influence of ancestry from the

representation of specific phenotypes, and an ensemble modeling approach, to combine predictions derived from the disentangled latent representation and the original data. Authors evaluated DisPred's performance by predicting the risk of AD in a multi-ethnic cohort consisting of both AD cases and controls. The results outperformed the ones obtained with existing models, especially in minority populations and for individuals with mixed ancestry.

An example of XAI tools applied to DL models in AD is in a published method called "c-Diadem[118], a DL classifier that uses pathway constraints in a multimodal neural network to identify potential genetic markers for AD. This tool incorporates genetic data and KEGG pathway constraints to predict the presence of AD, mild cognitive impairment (MCI), as well as cognitively normal (CN) individuals. The c-Diadem model reached an accuracy of 0.69 and an AUC-ROC of 0.70 in the test dataset. The SHapley Additive exPlanations (SHAP) scores were used to identify specific genes and genetic variants that had a significant impact on the model's predictions. Also, authors in another work[119] proposed DeepGAMI, an interpretable DL model designed to improve genotype based phenotype prediction. DeepGAMI was trained using genotype and gene expression data in the context of brain disorders. Additionally, it used integrated gradients for model interpretability.

Another example of XAI applied to DL is the *explainable variational autoencoder* (E-VAE) classifier model as proposed in a published work[120]. Authors in this work applied E-VAE on genomic data from the health and retirement study (HRS) to classify AD and related dementias (ADRD) and controls without dementia, achieving a predictive accuracy of 0.71 in the HRS dataset. They also tested the model in an independent cohort (ROSMAP), reaching an accuracy of 0.62. In addition, they provided insights into the biological mechanisms of ADRD through interpretable latent features extracted from the models using a linear decoder approach.

Regarding the detection of epistatic events in AD, researchers in ADNI consortia developed a computational tool called GenEpi[121], designed to identify gene-gene interactions associated with complex traits using a ML approach. GenEpi employs a two-stage modeling workflow to identify both within-gene and cross-gene epistasis. The tool adopts two-element combinatorial encoding when producing features and constructs prediction models using L1-regularized regression with stability selection. The study emphasizes the significance of uncovering epistasis for understanding the complex pathogenesis of AD. Also, VariantSpark, a ML approach to GWAS, and BitEpi, designed for uncovering epistatic events[122] were used by authors in a recently published study[123]. The goal was to uncover AD-associated genetic variants and interactions in two separate cohorts, ADNI and UKB. By considering significant epistatic interactions in their analysis, they were able to explain 10.41% more of the variation in AD compared to the LR method, that does not specifically account for interactions.

1.5.3 Schizophrenia (SC)

Schizophrenia (SC) is a mental disorder in which affected patients experience hallucinations, delusions, extremely disordered thinking and behaviour that impairs daily functioning and, in some cases, can be disabling. A review of studies published between 1980 and 2000 found that the lifetime prevalence of SC and related disorders is about 5.5 per 1,000 of individuals, but there was a significant variability across geographical regions[124]. The pathophysiological mechanisms behind SC are yet not fully understood. Neuroimaging studies have shown that the brain is fundamentally affected in the illness, with widespread structural gray and white matter involvement, functionally abnormal cortical and subcortical information processing, and neurometabolic dysregulation present in patients[125]. Studies have identified several candidate genes that may be associated with an increased risk of SC[126]. Additionally, genetic predisposition to SC

has been associated with an increased use of cannabis[127]. However, more research is needed to fully understand the genetic and biological mechanisms underlying SC.

Several studies applied PRS to distinct SC symptoms and treatment responses[128] [129] [47]. A study revealed that individuals possessing elevated PRS for SC displayed limited response enhancements under antipsychotic drug treatment. Consequently, the genetic predisposition indicated by the PRS could serve as a prognostic biomarker for treatment[48]. Also, other studies have shown that the PRS for SC may be linked to heterogeneity in cognitive performance[130] [131]. Cardiovascular disease is a major cause of excess mortality in people with SC. In this regard, a high PRS for SC is associated with cardiac impairments[132]. Overall, these studies suggest that PRS may be useful in predicting the prognostic of SC, including treatment response, cognitive, and cardiac impairments.

Several works applied ML for the genetic prediction of SC[133]. The researchers in a published study[134] used a support vector machines (SVM) approach to classify individuals with SC from controls in a large cohort. They compared the accuracy of the SVM-based approach with the traditional PRS method. The study aimed to determine if SVM are effective for identifying nonlinear genetic effects, such as interactions between genes. The findings revealed that PRS achieved better classification accuracy than both linear and nonlinear SVM. Additionally, researchers noticed that nonlinear SVM were more accurate than linear SVM when dealing with a high number of genetic variants. Despite the better performance of PRS, authors proposed that nonlinear SVM could be a useful tool for making predictions based on genetic interactions.

In a recent study[135], authors introduced a SVM ensemble for classifying individuals with SC and healthy controls, using both functional magnetic resonance imaging (fMRI) and genomic data. The method was evaluated with 40 subjects (20 patients and 20

controls) using a validated leave-one-out approach. The best classification accuracy was obtained with the model combining fMRI and SNP information reaching 0.87. Even though the cohort used in this study was small and results should be interpreted cautiously, the authors concluded that combining genetic and fMRI data yields higher accuracy than using each data type separately. Other studies reached similar conclusions when combining neuroimaging and genomic data in DL models to classify individuals with SC[136 137].

The authors of a published work[138] developed GenNet, a DL framework for predicting phenotypes from genetic variants. They applied neural network structures that are interpretable, incorporating biological knowledge from public databases, resulting in networks with connections that mimic molecular interactions. GenNet suggested potential associations of novel genes with SC, and pointed to biological pathways that could be implicated in SC.

Authors in another study[139] presented a stepwise DL technique with multi-precision data (SLEM), an approach for investigating the role of SNP combinations in the development of SC by focusing on intermediate molecular and cellular functions. SLEM initially constructs core networks using limited but accurate multilevel assay data. Subsequently, it refines the weights of intermediate interactions using a larger but less precise dataset from public GWAS data. This method is aimed to offer insights into the epistatic genetic factors contributing to SC.

1.5.4 Parkinson's disease (PD)

Parkinson's disease (PD) is characterized by both motor and non-motor symptoms. Motor symptoms include tremors, bradykinesia (slowness of movement), rigidity, and postural instability. Non-motor symptoms include neuropsychiatric features, speech disorders, and sleep disturbances[140]. PD is twice more frequent in males with respect to females[141], and is diagnosed based on

clinical criteria, as there are no molecular test for the diagnosis of the disease[140]. The specific presentation of rest tremor, bradykinesia, rigidity, and loss of postural reflexes are used to differentiate PD from related parkinsonian disorders. However, given the lack of specific molecular biomarkers for PD, and the high similarity with other parkinsonian disorders, there is a risk of misdiagnosis[140]. At the pathophysiological level, PD is associated with the loss of dopaminergic neurons in the substantia nigra of the brain[142], but the exact molecular mechanisms triggering PD are not fully understood.

Several GWAS have been conducted for PD[143 144 145], and at least 90 independent risk variants have been identified that explain around 16% to 36% of the heritable risk of PD[146]. Mutations in the *SNCA* gene, which encodes *Alpha-synuclein*, are the most common genetic risk factor of PD[147].

PRS applied to PD have shown poor ability to predict the development of PD in healthy individuals[146 148 149 150]. These studies concluded that, with the current available data, meaningful PRS-based prognosis of PD at an individual level is not feasible yet.

Authors in a recent work[151] developed a two-stage quality-based sampling using RF for the selection and prioritization of SNPs obtained from GWAS in PD. The proposed method separated first SNPs into informative and irrelevant groups based on the GWAS *p-values*. When building the RF model, the SNP subspace for each tree was composed only of SNPs from the informative sub-groups. The proposed model identified 25 SNPs with a potential association with PD.

In a recent study[152], authors combined GWAS with ML techniques to enhance the understanding and prediction of PD. They initially employed correlation and GWAS analyses to identify the top demographic and genetic factors associated with the disease. Subsequently, the authors applied ANN, LR, RF and SVM

methods for predicting PD risk, using XAI methods on these models to assess the predictive power of individual genomic input features. Following this approach, the authors identified new loci potentially associated with PD.

In order to detect genomic interactions associated with PD, the previously described ML approach applied to AD, GenEpi, was also applied on a PD dataset consisting of 5,540 cases and 5,862 controls[153]. GenEpi identified significant SNP-SNP interactions with effects on PD risk at five independent genomic loci, including seven PD-associated genes (*GAK*, *TMEM175*, *SNCA*, *PLEKHM1*, *CRHR1*, *MAPT* and *NSF*).

Overall, the application of ML methods to identify individuals with complex diseases based on genomic data, capturing genomic patterns including interactions associated with these diseases, has gained popularity in recent years. However, a limitation is that the published works on this topic are relatively recent, and there are not yet many studies in this field. Additionally, some of the studies referenced in the previous lines used relatively small sample sizes in the analysis (fewer than 1,000 cases), which makes it challenging to draw generalizable conclusions.

2. Objectives

Machine Learning (ML) methods have demonstrated to be powerful tools in detecting complex patterns, including interactions and non-linear relationships in the data. In this work, I hypothesised that these methods could offer advantages in tasks such as detecting genomic patterns associated with complex diseases. This is because traditional tools, like GWAS and PRS, are typically designed to capture linear additive associations and may miss synergistic effects in the data. In addition, as ML models learn from the data, I was interested in evaluating the effect of different properties of genomic data on model performance.

ML methods were applied for the purpose of classifying individuals with multiple sclerosis (MS), Alzheimer's disease (AD), schizophrenia (SC), and Parkinson's disease (PD) in comparison to non-affected controls sourced from the UK Biobank (UKB) using data from genotyping arrays.

The primary objective of this study was to assess the variability and robustness of ML techniques in predicting complex diseases using genomic data. This is relevant because ML methods are sensitive to biases in the data, and their results may vary depending on the data used during training, the design of the model, the strategy used for training, and the hyperparameter space, among other reasons. Also, genomic variants inherently exhibit correlation due to linkage disequilibrium (LD), which may impact model performance. In addition, I compared the performance of ML models to the PRS. In summary, the primary goal of this study include:

- Evaluating models and investigating performance differences among ML methods and diseases.
- Assessing the influence of potential biases in the model predictions.
- Comparing the performance of ML methods with PRS.
- Implementation of feature selection techniques.

The secondary goal of this study is to apply explainability (XAI) tools to the ML models to extract information about the prioritized features that contributed the most in the classification task, pointing to predisposing or

protective genomic variants in the diseases under study. The secondary goal of this study include:

- Analysing the consistency of feature rankings across ML methods.
- Identifying the most informative genomic variants based on rankings generated by ML methods.
- Reporting the synergies among the prioritized genomic variants.

After exploring these aspects, I aim to provide insights into the considerations to take into account when using ML methods with genomic data for disease classification.

3. Results

3.1 Performance of the models

The first section of results involves the evaluation of the performance of ML methods, exploring their ability to classify cases and controls, as well as the variability observed across different folds, methods, and diseases.

Table 4 (a) and (b) show the evaluation metrics for models constructed with MS and AD, respectively. The mean and standard deviation of different evaluation metrics across the five folds in the outer loop of the nested CV are provided. The mean performance scores for both diseases typically ranged around 0.6 and 0.7, with few exceptions. Notably, FFN and CNN methods exhibited the least stable performance, as evidenced by the highest standard deviation across folds. In the case of AD, GB method performed similarly to CNN and FFN with low mean performances, while GB demonstrated relatively good performance in MS.

a) Multiple sclerosis

	accuracy mean	accuracy sd	specificity mean	specificity sd	sensitivity mean	sensitivity sd	AUC-ROC mean	AUC-ROC sd
GB	0.628	0.007	0.635	0.005	0.622	0.017	0.670	0.009
ET	0.625	0.006	0.660	0.014	0.590	0.022	0.660	0.006
RF	0.612	0.008	0.657	0.011	0.567	0.022	0.655	0.011
LR	0.635	0.005	0.635	0.008	0.634	0.010	0.690	0.009
FFN	0.629	0.014	0.599	0.059	0.660	0.075	0.674	0.012
CNN	0.619	0.011	0.652	0.058	0.587	0.067	0.654	0.018

b) Alzheimer's disease

	accuracy mean	accuracy sd	specificity mean	specificity sd	sensitivity mean	sensitivity sd	AUC-ROC mean	AUC-ROC sd
GB	0.637	0.021	0.651	0.022	0.623	0.034	0.671	0.021
ET	0.675	0.013	0.723	0.006	0.627	0.031	0.708	0.011
RF	0.681	0.011	0.723	0.007	0.639	0.026	0.709	0.014
LR	0.674	0.010	0.693	0.005	0.655	0.024	0.715	0.012
FFN	0.645	0.018	0.693	0.068	0.598	0.072	0.694	0.026
CNN	0.629	0.024	0.665	0.043	0.594	0.034	0.667	0.024

Table 4 comprises two independent tables showing the mean and standard deviation of evaluation metric values across the five folds in the outer loop of the nested CV. The evaluation metrics represented in the table include balanced accuracy, specificity, sensitivity, and AUC-ROC. For each column, the color scale ranges from darker to lighter, indicating better to worse performance, respectively. (a) presents results corresponding to MS, while (b) presents results corresponding to AD.

In both diseases, LR method exhibited low values of standard deviation and displayed consistent results across various evaluation metrics. ET and RF closely approached LR's performance in the context of AD. In general, sensitivity stood out as the evaluation metric with the least favourable results in terms of mean and standard deviation, indicating that models face more challenges in classifying the positive class than the negative class, as evidenced by the comparison with specificity.

Table 5 (a) and (b) show the evaluation metrics for models constructed using SC and PD, respectively. The mean performance scores for both diseases typically range from 0.5 to 0.6. In both diseases, CNN exhibited a large difference between the prediction of the positive and negative class, with the highest specificity values, hovering around 0.6, corresponding to the prediction of the negative class, and the lowest sensitivity values, around 0.4, corresponding to the prediction of the positive class. Conversely, in the case of PD, the FFN method displayed an opposite trend, with sensitivity having the highest value and specificity the lowest value across ML methods. In the SC models, performance appears to be almost random, with evaluation metrics close to 0.5, making it challenging to draw conclusions. In models constructed in PD, the LR method demonstrated a balance between reduced variability and relatively good performance, making it appear to be the most effective method.

a) Schizophrenia

	accuracy mean	accuracy sd	specificity mean	specificity sd	sensitivity mean	sensitivity sd	AUC-ROC mean	AUC-ROC sd
GB	0.530	0.012	0.526	0.011	0.534	0.028	0.543	0.018
ET	0.528	0.023	0.522	0.005	0.535	0.049	0.538	0.022
RF	0.516	0.014	0.523	0.008	0.509	0.036	0.525	0.013
LR	0.531	0.023	0.537	0.008	0.525	0.048	0.542	0.023
FFN	0.532	0.017	0.536	0.052	0.527	0.055	0.544	0.020
CNN	0.519	0.022	0.608	0.111	0.430	0.140	0.520	0.023

b) Parkinson's disease

	accuracy mean	accuracy sd	specificity mean	specificity sd	sensitivity mean	sensitivity sd	AUC-ROC mean	AUC-ROC sd
GB	0.569	0.006	0.552	0.012	0.585	0.019	0.591	0.012
ET	0.570	0.015	0.564	0.009	0.575	0.036	0.592	0.013
RF	0.571	0.010	0.554	0.008	0.588	0.025	0.593	0.009
LR	0.579	0.010	0.576	0.004	0.583	0.018	0.610	0.012
FFN	0.570	0.008	0.523	0.056	0.618	0.064	0.601	0.011
CNN	0.568	0.017	0.640	0.086	0.496	0.112	0.597	0.017

Table 5 has the same structure as Table 4, with (a) representing results corresponding to SC and (b) representing results corresponding to PD.

These findings emphasize the variability in the performance of ML methods when tested across various diseases. In particular, DL models (FFN and CNN) exhibited significant instability, showing substantial differences in specificity and sensitivity, along with a high standard deviation across folds. In contrast, LR appeared to be the method with the most consistent performance across the positive and negative class and different diseases.

For the diseases with greater performance, MS and AD, an external validation dataset was employed as an extra test set to evaluate the model's generalization performance. In the case of MS, I obtained access to two different cohorts from the International Multiple Sclerosis Genetics Consortium (IMSGC). I analysed these two cohorts, namely IMSGC MS and IMSGC MSRD, independently since they represented different populations, United Kingdom (UK) and United States (US), respectively. The data from these datasets was structured in family trios and therefore, I could only use one MS case for each affected family to evaluate models. Consequently, only sensitivity is depicted in Figure 9.

Similarly to the UKB cohort, FFN and CNN methods exhibited notable variability across folds in the US cohort, and CNN method in the UK cohort. The sensitivity in IMSGC test sets did not show lower values compared to the UKB test sets, supporting the model's ability to generalize from UKB data to other MS datasets. Interestingly, in the IMSGC MS cohort, which is formed with individuals from the UK, the sensitivity was better than that in UKB for all methods except for CNN (as shown in Figure 9(a)), and a better sensitivity was observed with GB in the IMSGC MSRD cohort as well (as shown in Figure 9(b)). This could be attributed to the more precise and specific selection of diagnosed MS cases in a dedicated study, such as the one conducted by IMSGC[154], in contrast to the strategy employed in UKB, where individuals were selected using general clinical records. Additionally, the greater similarity between UKB and IMSGC MS, both formed by subjects from the UK, may explain the significant differences found in all the ML methods except for CNN in this IMSGC cohort.

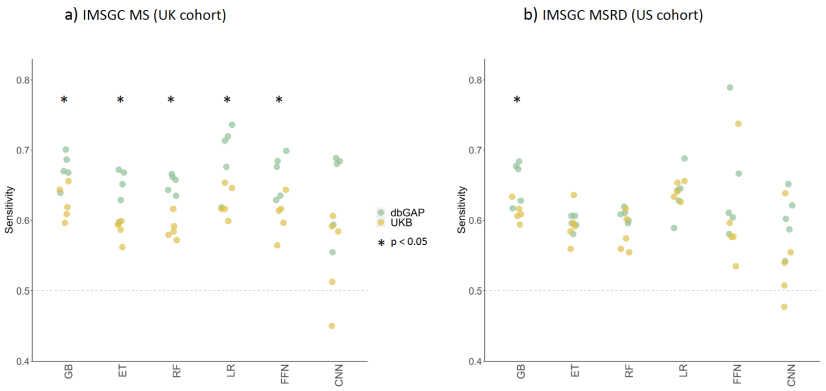

Figure 9 shows the sensitivity values of the five folds in the outer loop resulting from training the models with the UKB cohort, and testing the models in the dbGAP (in green) and UKB (in yellow) cohorts. In (a), the results are shown for the IMSGC MS cohort, and in (b), for the IMSGC MSRD cohort. The significance of the differences between IMSGC and UKB cohorts was assessed using a Wilcoxon rank-sum test.

In the case of AD, the ADNI dataset was employed as the validation set, comprising a cohort of individuals from the US. This dataset included cases and controls, and the balanced accuracy metric was employed to evaluate the models. No discernible differences in accuracy emerged when comparing the UKB test set and the ADNI dataset, as illustrated in Figure 10. These results underscore the model's ability to generalize across diverse populations, from the UKB (representing the UK population) to ADNI (representing the US population) and dismisses concerns about potential overfitting.

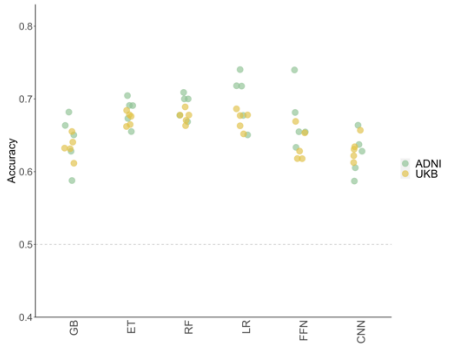

Figure 10 shows the values of balanced accuracy in the five folds of the outer loop, resulting from training the models with the UKB cohort, and testing the models in the ADNI (in green) and UKB (in yellow) cohorts.

3.2 Influence of potential biases in the model predictions

In the second section of the results, I evaluated the impact of other variables apart from genomic features, such as age at first diagnosis and sex, on the performance of the models. The goal was to check if model predictions were biased with respect to any of these two variables.

Figure 11 illustrates the differences in the age at the first diagnosis between true positives (TP), defined as samples correctly classified as positives by all ML methods, and false negatives (FN), defined as samples classified as false negatives by at least one ML method. As noted in the Methods section 5.1, 3% of the total AD cases had the category of early onset Alzheimer's disease (EOAD) in UKB. In this regard, in Figure 11(b) no significant differences in the age at the first diagnosis were observed between true positives and false negatives in AD. Therefore, I dismissed the possibility that EOAD significantly contributed to the number of FN, resulting in a negative impact on the performance of the AD models.

In the case of MS and PD (Figure 11 (a) and (d), respectively), no significant differences were observed in the age at the first diagnosis for FN or TP. In SC, an earlier age at the first diagnosis was reported in true positive females compared with false negative females (Figure 11(c)). It is

worth mentioning that the age of individuals at the first disease report may have potential biases in UKB, as clinical records are incomplete for some participants. Consequently, the statistically significant results found in SC females indicate a trend that should be corroborated with other dedicated studies.

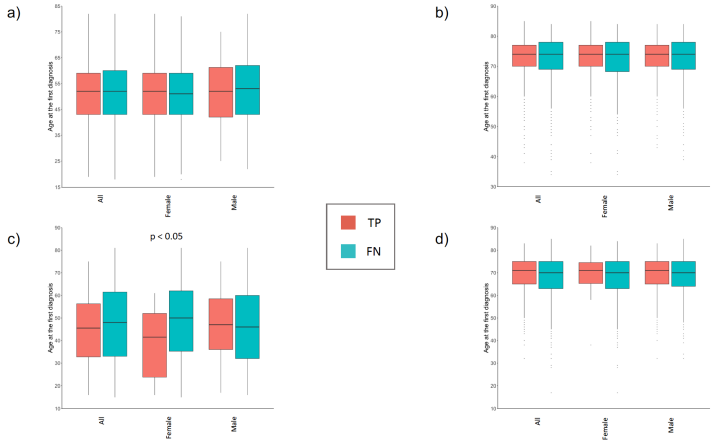

Figure 11 illustrates the differences in the age of individuals at the first report of the disease in the UKB clinical records. TP corresponds to individuals with the disease who were correctly classified as having the disease by all ML methods, while FN corresponds to individuals with the disease who were incorrectly classified by at least one method. The significance of the differences between TP and FN was assessed using a t-test. The plots in (a), (b), (c), and (d) represent the results for MS, AD, SC, and PD, respectively.

In addition to the genomic features, the binary sex feature was used in the models. This feature is especially relevant in diseases with sex imbalance, such as MS, reported to be three times more frequent in females with respect males[78 155], and PD, which is twice more frequent in males with respect to females[141]. A similar imbalance of females to males is present in the UKB cohort for both diseases, as indicated in Table 15 of the Methods section 5.1.

Figure 12 and Figure 13 display the percentage of samples that were correctly classified by 0, 1, 2, 3, 4, 5, or 6 methods, including GB, ET, RF, LR, FFN and CNN methods, considering all samples and stratifying by the

variable sex. In the figures, I will stress the numbers of the yellow bars, indicating the percentage of samples that were correctly classified as true positives or true negatives by the six ML methods. These samples likely contain the most representative predisposing or protective features of the disease, as they were correctly classified by all methods.

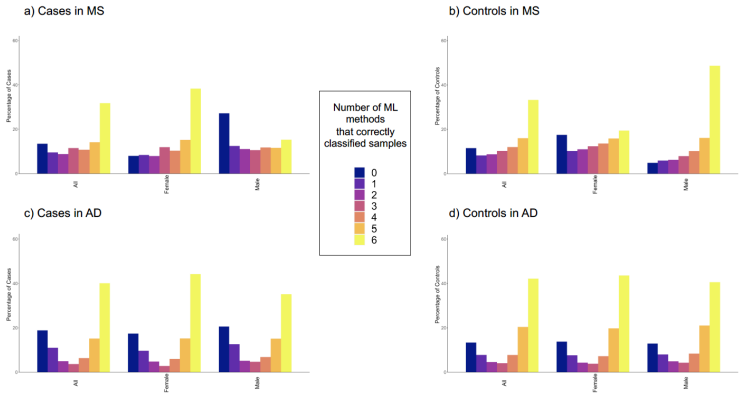

Figure 12 shows the percentage of cases and controls that were correctly classified by 0 to 6 ML methods in all samples, females, and males. The plots in (a) and (b) represent the cases and controls in MS, respectively. The plots in (c) and (d) represent the cases and controls in AD, respectively.

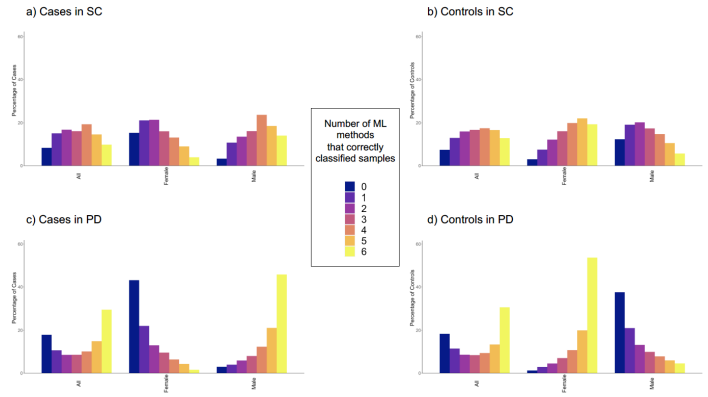

Figure 13 follows the same structure as Figure 12. The plots in (a) and (b) represent the cases and controls in SC, respectively. The plots in (c) and (d) represent de cases and controls in PD, respectively.

For MS, a sex bias was observed, with a higher percentage of cases consistently predicted by the six methods in females compared to males (Figure 12(a), yellow bars) and a higher percentage of controls consistently predicted by the six methods in males compared with females (Figure 12(b), yellow bars). In PD, the opposite sex-dependent trends are observed with males correctly classified as having the disease showing more agreement across methods compared to females (Figure 13(c), yellow bars), and females correctly classified as controls showing more agreement across methods compared to males (Figure 13(d), yellow bars). For SC there was not a clear enrichment in the percentage of samples correctly classified by the six methods represented in Figure 13(a) and Figure 13(b), and the influence of the sex variable, even if not as evident, was inferred to follow a pattern similar to PD. Contrarily, for AD, the six methods consistently correctly classified around 40% of samples as cases (Figure 12(c)) or controls (Figure 12(d)) without any significant difference between females and males. These results suggest that, even if there were differences in the evaluation metrics when using different ML methods in AD (see Table 4(b)), around 40% of the individuals were consistently classified as true positives or true negatives by all methods.

Notably, the diseases showing the highest difference between females and males in Figure 12 (a) and (b), and Figure 13 (c) and (d), MS and PD, were also the ones with the highest sex bias among cases in the UKB cohort. This may indicate that the bias in the classification is caused by the overrepresentation of one sex with respect to the other in the cases of the training set. To investigate this further, different models were built for females and males independently, using the same sample size, and the results of the five folds in the outer loop of the nested CV were compared. DL methods were excluded from these comparisons because they exhibited high variability across folds, making it difficult to draw any conclusions. The goal was to check if females or males showed better predictions when using independent models for each sex, thus removing the variability introduced by the sex feature. The comparison of the specificity and sensitivity is shown in Figure 14 and in Table 6. In Figure 14, the green dots represent the results of specificity and sensitivity in the original model, which considered both females and males. The

performance of models constructed independently for females and males are indicated by yellow and blue dots, respectively.

The first observation is the considerable variability across folds in the independent models for each sex, likely caused by the reduction in the number of samples in training and testing, making generalization more challenging and unstable. Partly due to the high variability across folds, the performance of models independently built for each sex did not surpass that of models constructed with both sexes in any case. Instead, as depicted in Figure 14, the performance in the original models was higher compared to the female and male models in some instances, and the variability across folds was visibly lower in most cases.

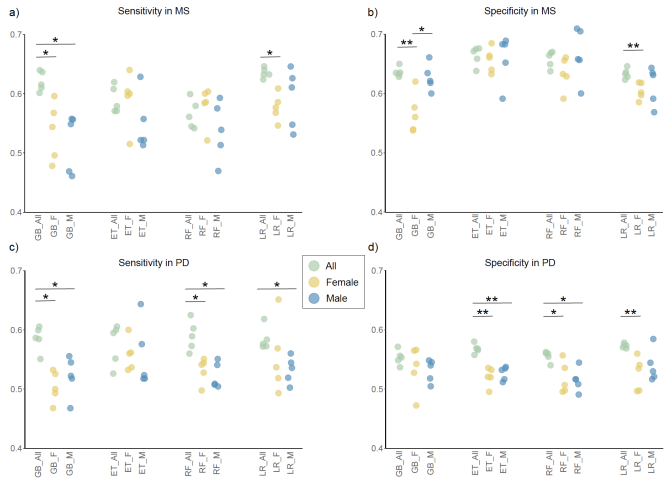

Figure 14 depicts dot plots illustrating the differences in sensitivity and specificity between the original models and models built exclusively for females and males. Figures (a) and (b) represent the values of sensitivity and specificity in MS, respectively. Figures (c) and (d) represent the values of sensitivity and specificity in PD, respectively. The significance of the differences across original models and independent models built for each sex was assessed with a Wilcoxon rank-sum test.

The specificity of GB male models was higher than GB female models in MS (Figure 14(b)). For the rest of the comparisons, no significant differences between females and males were observed, although MS

female models generally showed less standard deviation across folds compared with MS males, as observed in the values of standard deviation of balanced accuracy coloured darker represented in Table 6(a). Despite these differences, there is not sufficient evidence to support the idea that one sex is better predicted than the other, or that the performance improved in the independent models built for each sex compared to the original ones.

a)

	accuracy mean	accuracy sd	sensitivity mean	sensitivity sd	specificity mean	specificity sd
GB female	0.552	0.028	0.536	0.049	0.567	0.034
GB male	0.573	0.027	0.519	0.049	0.627	0.023
ET female	0.624	0.015	0.591	0.046	0.656	0.021
ET male	0.604	0.035	0.549	0.048	0.660	0.041
RF female	0.606	0.013	0.579	0.033	0.634	0.028
RF male	0.602	0.034	0.538	0.049	0.666	0.044
LR female	0.591	0.016	0.577	0.023	0.604	0.014
LR male	0.603	0.036	0.592	0.050	0.614	0.032

b)

	accuracy mean	accuracy sd	sensitivity mean	sensitivity sd	specificity mean	specificity sd
GB female	0.519	0.020	0.504	0.026	0.535	0.038
GB male	0.526	0.021	0.522	0.034	0.531	0.019
ET female	0.539	0.017	0.558	0.027	0.521	0.016
ET male	0.541	0.028	0.556	0.055	0.527	0.012
RF female	0.525	0.009	0.532	0.021	0.518	0.027
RF male	0.519	0.006	0.523	0.022	0.515	0.019
LR female	0.540	0.027	0.554	0.061	0.526	0.028
LR male	0.536	0.020	0.533	0.022	0.539	0.027

Table 6 comprises two separate tables showing the mean and standard deviation of evaluation metric values across the five folds in the outer loop of the nested CV for the models independently built for each sex. (a) shows the results for MS, and (b) shows the results for PD. For each column, the color scale ranges from darker to lighter, indicating better to worse performance, respectively.

In addition, to assess the importance of the sex feature in the classification among all diseases, I compared models using only the sex feature with the original models including both, sex and genomic features. The results in Figure 15 (a), (c) and (d) demonstrate that the sensitivity in MS, SC and PD, depicted in green, is higher in the models using only the sex feature (dots in the figure) compared with the original models (triangles in the figure), but at the expense of having lower specificity in the case of MS

and PD. In SC and PD, it appears that the sex feature plays a primary role in the classification, while the genomic variants exhibit less predictive power. This can be observed from the nearly overlapping estimates represented with triangles and dots in Figure 15 (c) and (d) for SC and PD, respectively. In the UKB cohort, the male-to-female ratio is approximately 1.40 for SC, which is lower than the approximately 1.69 ratio for PD (refer to Table 15 in the Methods section 5.1). Even so, the poor performance of SC in models with genomic features, indicating low predictiveness of genomic variants for this disease, could reinforce the predominant use of the sex feature for the SC classification. Contrarily in AD, both sensitivity and specificity were higher in the original models that included genomic features compared to the models only based on the binary class sex. This is illustrated by triangles being higher than dots in Figure 15(b), demonstrating a moderated influence of the sex feature and highlighting the predictiveness of the genomic features in AD.

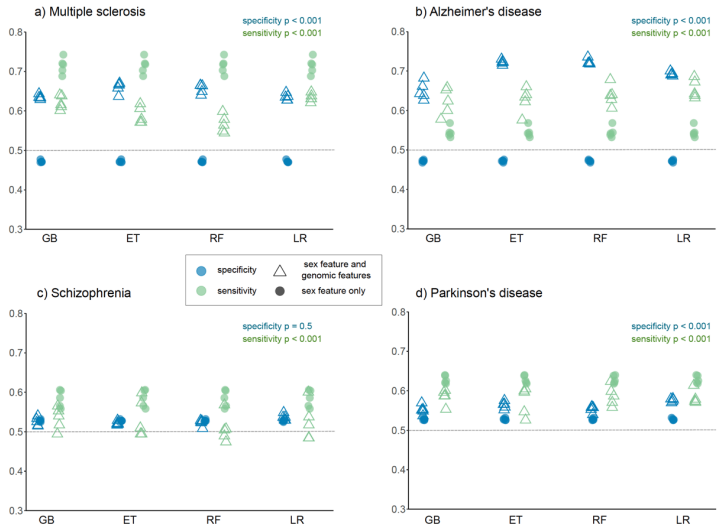

Figure 15 illustrates the comparison of sensitivity and specificity between the original models (triangles in the plot) and models constructed solely with the sex feature (dots in the plot). Diseases MS, AD, SC, and PD are individually represented in (a), (b), (c), and (d), respectively.

These results support the notion that in diseases exhibiting a sex bias, where one sex is more prevalent compared to the other, the sex feature is an informative predictor that, in some cases where genomic features have low predictiveness, it may significantly contribute to the predictions.

3.3 Comparison of machine learning methods with polygenic risk score

In the third section of the results, I compared the results of ML methods with polygenic risk score (PRS), which is the most common tool used in population genomics to predict disease risk based on genomic information. PRS estimates an individual's genetic liability to a disease by aggregating the effects of many common genetic variants associated with the condition. The calculation of PRS is based on linear regression models and follows a different approach compared to ML methods. PRS works under the assumption that the weights for each allele, obtained from GWAS summary statistics, are static, independent, and not modified by other genetic or environmental factors. In contrast, in ML methods, the effect of each allele is estimated during training, where these methods learn patterns in the data to make predictions. Therefore, with ML methods, fewer assumptions about the nature of the genetic effects being modelled are made.

To facilitate the comparison, I used the same samples in the PRS for adjusting the model and testing the results as those selected for the final ML models, corresponding to the outer loop of the nested CV with 5 folds. Therefore, for each fold, the same individuals were compared in the PRS and the ML models. It is important to note that there is no specific maximum number of genomic variants that can be used in PRS models. This is different from ML methods, where dimensionality issues arise when there are a large number of features relative to the number of samples[156 8]. Consequently, I conducted the experiment twice. First, I used all the genomic variants that had successfully passed the quality filters present in both the UKB array and GWAS summary statistics, resulting in PRS ALL models. Second, I limited the analysis to the disease-related variants that were employed in the ML models, creating PRS RED models. The aim

was to evaluate how the performance of PRS models changed when using the entire set of genomic variants present in the genotyping arrays, as opposed to the reduced set of disease-related genomic variants employed in the ML models. However, the approach used for the calculation of PRS is designed to give better results when including many SNVs, even if these are not associated with the trait at a statistically significant *p-value*[43]. Therefore, in this work PRS RED models were used to compare with ML models but do not adhere to the best practices for the PRS calculation.

The genomic variants used in the PRS models are visualized in Figure 16, along with the *p-values* that represent the statistical significance of the association between each SNV and the disease. These *p-values* were obtained from the GWAS summary statistics used in the PRS calculation. Genomic variants included in the ML and PRS RED models are highlighted in green. The Manhattan plots reveal peaks corresponding to hotspots of genomic variants with notably low *p-values* on chromosome 6 for MS and chromosome 19 for AD. In this regard, it is well-documented in the literature the association of the loci coding for the HLA genes on chromosome 6 with MS[157], as well as the association of the *APOE* region on chromosome 19 with AD[158]. *P-values* obtained from GWAS in SC and PD were lower than in the other diseases, with SC displaying the lowest *p-values*. The low *p-values*, which indicate a weak association between the SNVs and the disease, reinforces the challenges previously encountered when attempting to classify SC and PD with ML methods.

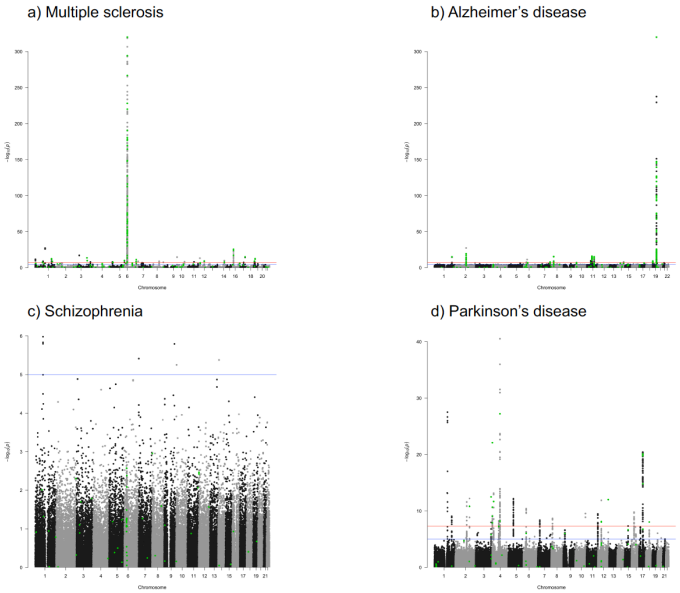

Figure 16 displays the Manhattan plots corresponding to the p-values from the GWAS summary statistics used in the PRS calculation. In green, the SNVs used in the PRS RED models, also used in the ML models, are highlighted. The blue and red horizontal lines indicate the thresholds of $1e^{-5}$ and $5e^{-8}$, respectively, related to the p-values estimating the association with the disease. Results for MS, AD, SC, and PD are presented in (a), (b), (c) and (d), respectively.

Results of the best PRS models for each fold selected after the *p-value thresholding* are shown in Table 7 for MS and AD, and in Table 8 for SC and PD. Among all the disease models, PRS models applied to AD had the highest "PRS.R2" scores, indicating the greatest variance explained by the genomic variants, and the lowest *p-values* (Table 7(b)), followed by MS (Table 7(a)), PD (Table 8(b)), and SC (Table 8(a)), respectively. This decreasing trend in performance across the different diseases matches the results previously exposed in the Results section 3.1 with ML methods. Regarding the variance explained by the covariates including the sex feature, represented in the column "Null.R2", MS followed by SC and PD showed the highest influence of the covariates in the prediction, in some cases exceeding the amount of variance explained by the genomic variants indicated in column "PRS.R2". As explored in the previous

section, these results could be partially explained by the sex imbalance present in these diseases.

a) Multiple sclerosis

Method	Fold	Full.R2	PRS.R2	Null.R2	P	Threshold	Num_SNP
RED	Fold1	0.119	0.054	0.068	9.33E-29	0.010	48
RED	Fold2	0.115	0.053	0.066	4.67E-28	0.010	48
RED	Fold3	0.128	0.058	0.075	1.61E-30	0.010	48
RED	Fold4	0.110	0.050	0.064	1.22E-26	0.445	70
RED	Fold5	0.096	0.052	0.046	9.99E-28	0.010	48
ALL	Fold1	0.146	0.083	0.068	2.36E-42	1.000	6007
ALL	Fold2	0.147	0.087	0.066	5.15E-44	1.000	6007
ALL	Fold3	0.162	0.095	0.075	9.96E-48	1.000	6007
ALL	Fold4	0.142	0.084	0.064	5.03E-43	0.372	2847
ALL	Fold5	0.128	0.086	0.046	8.33E-44	1.000	6007

b) Alzheimer's disease

Method	Fold	Full.R2	PRS.R2	Null.R2	P	Threshold	Num_SNP
RED	Fold1	0.144	0.136	0.009	1.65E-81	0.001	39
RED	Fold2	0.143	0.136	0.008	1.07E-81	0.001	39
RED	Fold3	0.143	0.136	0.008	1.14E-81	0.001	38
RED	Fold4	0.150	0.144	0.007	6.46E-86	0.001	39
RED	Fold5	0.144	0.137	0.008	5.21E-82	0.001	39
ALL	Fold1	0.148	0.139	0.009	6.90E-84	0	77
ALL	Fold2	0.147	0.140	0.008	2.34E-84	0	77
ALL	Fold3	0.149	0.142	0.008	1.24E-84	0	77
ALL	Fold4	0.165	0.160	0.007	6.30E-94	0	77
ALL	Fold5	0.150	0.143	0.008	1.67E-85	0	77

Table 7 consists of two separate tables showing the statistics obtained for each PRS model and fold using the same samples as in ML methods in (a) for MS and (b) for AD. Columns "Full.R2", "PRS.R2", and "Null.R2" represent the observed phenotypic variance explained by the full model including SNVs and covariates, only by the SNVs, and only by the covariates, respectively. Column "P" refers to the empirical p-value of the best model fit calculated with the comparison of randomly shuffling the phenotype and repeating the analysis 10,000 times. The "Threshold" column indicates the p-value threshold in which the genomic variants were selected for the inclusion in the best model, and column "Num_SNP" represents the number of genomic variants included with this threshold.

For MS (Table 7(a)), PRS ALL models explained a greater amount of genomic variability when compared to the PRS RED models, as indicated by the higher values of "PRS.R2" and lower *p-values*. Therefore, in the case of MS, there appears to be a benefit in considering all the variants present in the array for the PRS calculation. In this regard, four out of five folds of PRS ALL models had a threshold of one in MS, meaning that all the SNVs present in the UKB array (target data) and the GWAS summary

statistics (base data) were selected as informative (Table 7(a)). In fact, PRS ALL models built for MS were the ones with the largest number of SNVs in the final models by far after applying C+T, as indicated in the column “Num_SNP”, highlighting the polygenic nature of this disease.

a) Schizophrenia

Method	Fold	Full.R2	PRS.R2	Null.R2	P	Threshold	Num_SNP
RED	Fold1	0.061	0.009	0.053	1.36E-03	0.004	4
RED	Fold2	0.054	0.003	0.051	7.18E-02	0.004	4
RED	Fold3	0.049	0.007	0.042	5.05E-03	0.004	4
RED	Fold4	0.039	0.002	0.037	9.66E-02	0.004	4
RED	Fold5	0.049	0.003	0.046	8.00E-02	0.027	11
ALL	Fold1	0.056	0.004	0.053	3.81E-02	0	29
ALL	Fold2	0.054	0.003	0.051	5.69E-02	0.007	651
ALL	Fold3	0.051	0.009	0.042	1.15E-03	0.002	219
ALL	Fold4	0.045	0.008	0.037	2.23E-03	0.001	174
ALL	Fold5	0.051	0.005	0.046	1.80E-02	0.001	174

b) Parkinson’s disease

Method	Fold	Full.R2	PRS.R2	Null.R2	P	Threshold	Num_SNP
RED	Fold1	0.058	0.023	0.036	5.82E-20	0.466	37
RED	Fold2	0.058	0.026	0.034	4.18E-22	0.235	34
RED	Fold3	0.061	0.024	0.039	1.33E-20	0.262	35
RED	Fold4	0.064	0.029	0.036	7.85E-25	0.232	33
RED	Fold5	0.058	0.026	0.032	1.14E-22	0.232	33
ALL	Fold1	0.067	0.032	0.036	9.79E-27	0.000	45
ALL	Fold2	0.066	0.033	0.034	4.97E-28	0.000	148
ALL	Fold3	0.077	0.040	0.039	2.25E-33	0.000	148
ALL	Fold4	0.073	0.039	0.036	3.71E-32	0.001	506
ALL	Fold5	0.063	0.032	0.032	1.21E-26	0.000	148

Table 8 follows the same structure and variables as Table 7. Comprises two separate tables showing the statistics obtained for each PRS model and fold in (a) for SC and (b) for PD.

Unlike ML, PRS do not return probabilities of the disease, or the class associated with each subject. Instead, PRS scores are continuous values whose theoretical range is variable and increases with the number of SNVs included in the model. Therefore, PRS scores from different models containing different numbers of SNVs cannot be directly compared. In other words, PRS is a tool for disease risk stratification, and the ML models employed in this work are classifiers. To convert PRS scores into predicted classes, the regular practice is to set a cut-off using percentiles, for instance with the upper 99^{th} and lower 50^{th} to identify individuals at high or low risk for the disease as predicted positives and predicted negatives.

Figure 17 and Figure 18 show the quantile plots with the values and confidence intervals of the odds ratio (OR) for each range of PRS percentiles for PRS RED and for PRS ALL, respectively. A good indicator of PRS performance is observing increasing values of OR with higher percentiles, indicating that higher PRS scores are correlated with a higher prevalence of the disease. MS, AD and PD show this increasing trend in Figure 17 and Figure 18 . In SC, the increment of OR with the percentiles is almost inexistent, especially in PRS ALL, where OR values are quite constant. As expected, these results correlate with the *p-values* of the model fit previously reported in Table 7 and Table 8. For the subsequent analysis, the percentile range of PRS (99,100] will be considered to determine individuals at the highest risk of developing the disease, and therefore, predicted positives.

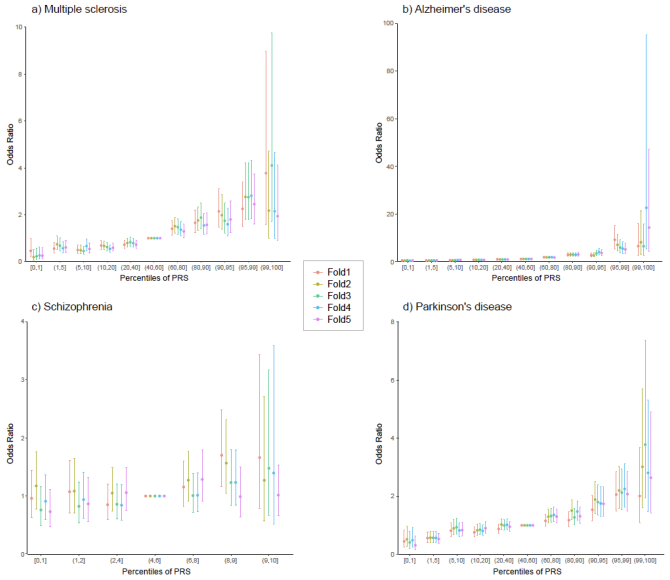

Figure 17 shows the quantile plots of PRS RED models for MS, AD, SC, and PD in (a), (b), (c) and (d), respectively.

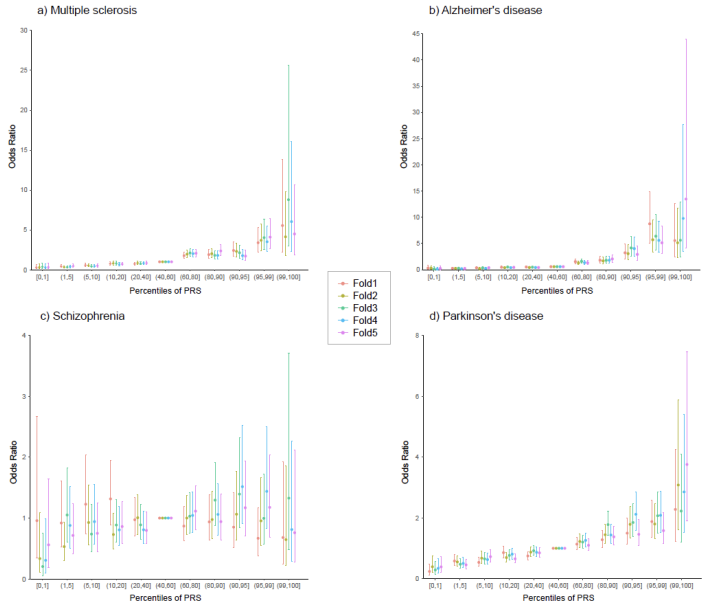

Figure 18 shows the quantile plots of PRS ALL models for MS, AD, SC, and PD in (a), (b), (c) and (d), respectively.

In my work I applied percentiles to the probabilities obtained with ML methods to compare their performance with PRS. For all the methods, I calculated the relative risk (RR) and OR by considering individuals in the upper 99th percentile as those predicted to have the disease. Details on the formulas used for RR and OR calculations are provided in the Methods section 5.4. For each fold, the 99th percentile was equivalent to 165, 166, 163, and 167 individuals in MS, AD, SC and PD, respectively. The total number of individuals with the disease in each fold was, 404, 498, 197 and 625, in MS, AD, SC and PD, respectively. Values of RR and OR higher than one indicated a higher proportion of individuals with the disease in the upper 99th percentile. Specifically, a RR of 2 indicates that individuals in the upper 99th percentile are twice as likely to develop the disease compared to the rest of individuals in lower percentiles.

a) Multiple sclerosis

	RR mean 99th percentile	RR sd 99th percentile	OR mean 99th percentile	OR sd 99th percentile	RR mean ML classification	RR sd ML classification	OR mean ML classification	OR sd ML classification
GB	3.598	1.575	3.901	1.822	2.790	0.162	2.868	0.170
ET	3.886	1.653	4.247	2.004	2.719	0.119	2.795	0.125
RF	3.341	1.687	3.610	1.915	2.456	0.159	2.517	0.168
LR	5.509	1.390	6.232	1.782	2.935	0.119	3.021	0.126
FFN	4.107	0.630	4.455	0.752	2.908	0.426	2.988	0.444
CNN	2.827	1.060	2.984	1.219	2.641	0.269	2.712	0.284
PRS RED	3.013	0.853	3.187	0.958	---	---	---	---
PRS ALL	4.002	0.729	4.333	0.864	---	---	---	---

b) Alzheimer's disease

	RR mean 99th percentile	RR sd 99th percentile	OR mean 99th percentile	OR sd 99th percentile	RR mean ML classification	RR sd ML classification	OR mean ML classification	OR sd ML classification
GB	4.313	0.823	4.795	1.037	3.013	0.564	3.128	0.609
ET	6.643	1.315	8.031	1.914	4.183	0.426	4.408	0.462
RF	6.837	0.719	8.273	1.044	4.391	0.415	4.636	0.453
LR	6.882	0.651	8.336	0.969	4.094	0.368	4.300	0.398
FFN	5.906	1.696	7.011	2.528	3.350	0.611	3.502	0.673
CNN	4.788	1.798	5.473	2.487	2.862	0.608	2.971	0.661
PRS RED	5.987	0.711	7.044	0.992	---	---	---	---
PRS ALL	5.683	1.098	6.644	1.523	---	---	---	---

Table 9 comprises two tables showing the mean and standard deviation of the relative risk (RR) and odds ratio (OR) across the samples used in the five folds. The formulas used in the calculation of RR and OR are provided in the Methods section 5.4. RR and OR were calculated considering as positives the samples ranked within the top 99th percentile with the best scores or probabilities. In the case of ML methods, the cutoff of probability 0.5, which is the default in these methods, was also considered to define positives and calculate the RR and OR and is represented in additional columns. Results for MS and AD are presented in tables (a) and (b), respectively. For each column, the color scale ranges from darker to lighter, indicating better to worse performance, respectively.

a) Schizophrenia

	RR mean 99th percentile	RR sd 99th percentile	OR mean 99th percentile	OR sd 99th percentile	RR mean ML classification	RR sd ML classification	OR mean ML classification	OR sd ML classification
GB	1.834	1.141	1.866	1.187	1.276	0.130	1.280	0.132
ET	1.206	0.452	1.211	0.458	1.267	0.237	1.271	0.240
RF	1.319	0.998	1.334	1.018	1.140	0.133	1.142	0.135
LR	2.139	0.933	2.178	0.979	1.298	0.254	1.302	0.258
FFN	1.109	0.663	1.115	0.675	1.298	0.165	1.302	0.167
CNN	0.898	0.903	0.905	0.913	1.180	0.215	1.182	0.218
PRS RED	1.264	0.848	1.275	0.864	---	---	---	---
PRS ALL	0.905	0.662	0.909	0.675	---	---	---	---

b) Parkinson's disease

	RR mean 99th percentile	RR sd 99th percentile	OR mean 99th percentile	OR sd 99th percentile	RR mean ML classification	RR sd ML classification	OR mean ML classification	OR sd ML classification
GB	1.758	0.801	1.833	0.883	1.707	0.078	1.744	0.083
ET	1.887	0.300	1.958	0.335	1.725	0.196	1.763	0.208
RF	1.568	0.388	1.608	0.423	1.741	0.134	1.779	0.142
LR	2.059	0.630	2.161	0.714	1.857	0.159	1.903	0.170
FFN	2.552	0.516	2.725	0.602	1.750	0.122	1.788	0.129
CNN	1.962	0.684	2.054	0.745	1.745	0.208	1.785	0.220
PRS RED	2.254	0.328	2.372	0.376	---	---	---	---
PRS ALL	2.389	0.700	2.542	0.835	---	---	---	---

Table 10 consists of two tables with the same structure and variables as Table 9, for SC in (a) and PD in (b).

The mean and standard deviation of RR and OR across the five folds are provided in Table 9 for MS and AD, and in Table 10 for SC and PD. For ML methods, results are similar as the ones obtained for the general evaluation metrics presented in Table 4 and Table 5, with LR doing relatively well across diseases, and RF, ET and LR showing the best performance in AD.

In agreement with the results previously discussed in Table 7(a) for MS, RR and OR in PRS ALL were better than in PRS RED, and PRS ALL was among the top three best methods for this disease after LR and FFN as shown in Table 9(a). Consequently, LR and FFN proved to be more effective at stratifying the risk of MS, even when using a reduced number of features, which, in the case of PRS RED, did not offer as much support. Contrarily in AD, PRS RED had greater RR and OR with less standard deviation across folds compared with PRS ALL, and both PRS models had average performance when compared with the other ML methods, as shown in Table 9(b).

For SC, PRS RED performed better than PRS ALL and was within the average performance as well, but PRS ALL showed less variability across

folds (see Table 10(a)). Nevertheless, as previously noted, it is difficult to extract any conclusions from SC models due to their poor performance. In fact, the mean values of RR and OR obtained with CNN and PRS ALL models were below one, indicating a lower proportion of cases in the 99th percentile with respect to the other samples with lower scores. Also, even though the other methods had values of RR and OR slightly higher than one, they exhibited high standard deviation. Consequently, for SC, the use of PRS did not lead to an improvement in the results obtained with ML methods.

In PD, PRS RED had slightly lower values of RR and OR compared with PRS ALL but demonstrated less standard deviation across folds (see Table 10(b)), and both PRS approaches were among the top three methods with the highest RR and OR mean only after FFN.

As previously exposed, PRS do not provide probabilities but instead offer risk scores associated with the disease, which are used to identify individuals at high risk and low risk. Although PRS may effectively identify individuals at high and low risk, they are not designed to work as binary classifiers. Instead, ML methods were employed as classifiers in the previous sections of this work, applying a cut-off of probability 0.5, which is the default setting used to classify samples as positives (greater than or equal to 0.5) or negatives (lower than 0.5). The values of RR and OR considering the default cut-off used in the ML classification are provided in Table 9 and Table 10 as well. In MS and AD, results of RR and OR based on the ML classification are lower with respect to considering the 99th percentile of samples with the highest probability as predicted positives. This fact suggests that the use of a cut-off of probability 0.5 in ML methods reduces the proportion of true positives over the false positives with respect to using the 99th percentile.

In MS, FFN demonstrated the lowest standard deviation across folds when considering the top 99th percentile of samples with the highest probabilities as predicted positives. Using the top 99th percentile, the lowest probability of MS in the FFN ML classification was 0.89. Conversely, the same method exhibited the highest standard deviation when using the probability cut-off of 0.5. These results suggest that FFN displayed greater

robustness when setting a higher cut-off for classifying samples with MS. For SC and PD, no differences were observed when applying a cut-off of probability 0.5 or using the 99th percentile.

I checked if the individuals predicted as positives over the 99th percentile or predicted as negatives under the 50th percentile by the PRS RED and PRS ALL models were also consistently classified as positives and negatives across the ML methods. In Figure 19, Figure 20, Figure 21, and Figure 22, samples classified by PRS models in MS, AD, SC, and PD, respectively, are represented comparing their agreement with the six ML methods.

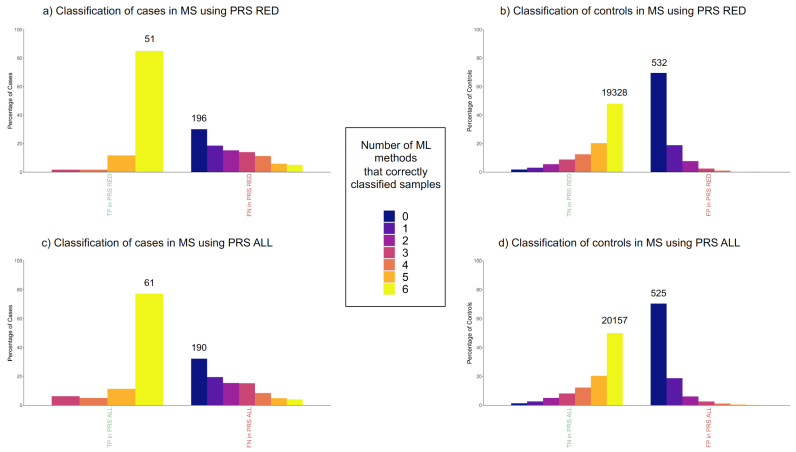

Figure 19 shows the percentage of MS and controls that were correctly classified by 0 to 6 ML methods in comparison with samples correctly classified (TP or TN labeled in green) or incorrectly classified (FN and FP labeled in red) by PRS models. The total number of samples is indicated above the bars for the groups with the highest percentage in each comparison. Plots (a) and (b) show the classification of MS and controls in PRS RED, respectively. Plots (c) and (d) show the classification of MS and controls in PRS ALL, respectively.

In MS (Figure 19), approximately 70% to 80% of cases predicted as positives by PRS were also classified as positives by the six ML methods, as indicated by the yellow bars in Figure 19 (a) and (c). Conversely, around 70% of the samples classified as false positives (FP) in PRS were also

misclassified by the six ML methods, as indicated by the dark blue bars in Figure 19 (b) and (d).

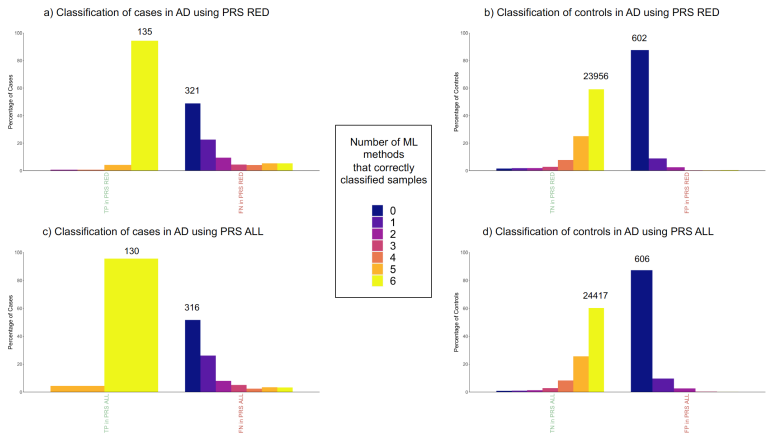

Figure 20 follows the same structure as Figure 19, but for AD.

In AD, the percentage of true positives (TP) in PRS with full agreement across ML methods exceeded the 90%, as indicated in the yellow bars of Figure 20 (a) and (c), while the percentage of true negatives (TN) in PRS with full agreement across ML methods was approximately 60% represented in the yellow bars of Figure 20 (b) and (d). Around 90% of the samples classified as FP in PRS were also mislabelled by the six ML methods (Figure 20 (b) and (d), dark blue bars).

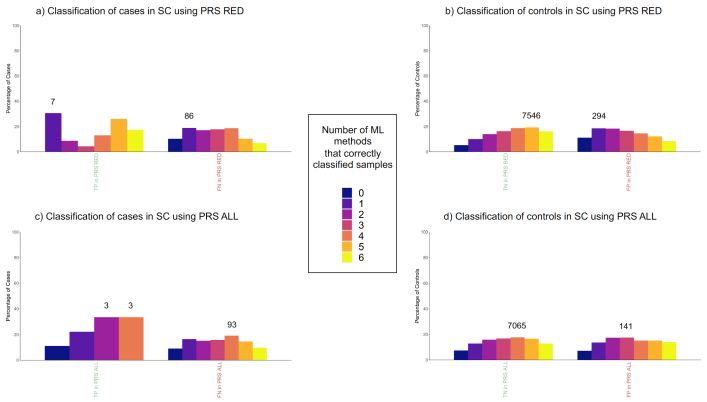

Figure 21 follows the same structure as Figure 19, but for SC.

In contrast, no specific agreement across methods was observed in SC, as indicated by the homogeneous percentages depicted in bars, possibly due to the poor results obtained in the classification for this disease (Figure 21). In PD, around 50% of TP and 40% of TN had full agreement across methods (Figure 22, yellow bars). In summary, the results obtained in MS and AD suggest that PRS and ML models demonstrate consistent classification results, with not only similarities in the values of RR and OR, but also in the classification of specific individuals.

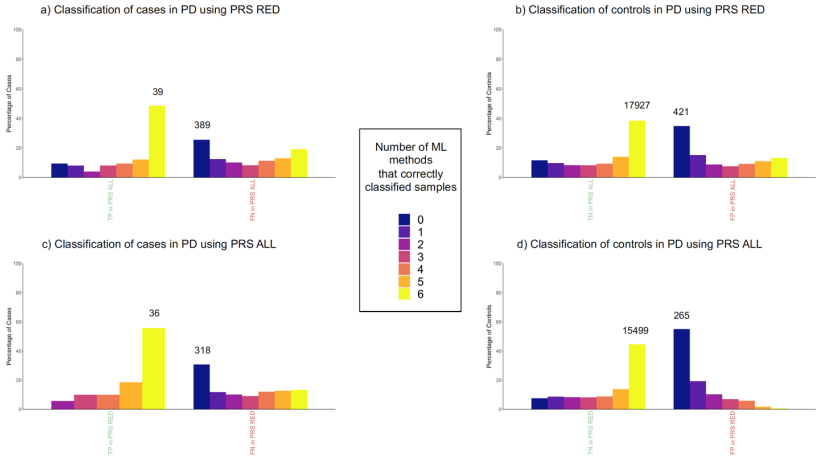

Figure 22 follows the same structure as Figure 19, but for PD.

Overall, the results presented in this section suggest that, with the evaluation based on percentiles, the performance of PRS is similar to that of the ML models. Yet, PRS is still the preferred method to be used in population genomics to stratify individuals with the genetic risk for a disease. In this regard, PRS offer several advantages compared with ML methods. The strengths and weaknesses of these methods will be further elaborated upon in the discussion.

3.4 Implementation of feature selection techniques

In this study I employed curated databases of disease-related variants to select the predictors for the ML models. However, these databases contain genomic variants from diverse studies conducted in various human populations, some of which may not be informative in the UKB cohort. Furthermore, certain genomic variants used as features in models are highly correlated due to LD, with a potential negative impact in the performance of models. To address this, I used feature selection techniques such as recursive feature elimination (RFE) and recursive feature elimination with cross-validation (RFECV) aiming to identify a subset of features with the potential to enhance model performance.

Because of the considerable variability observed across folds in DL methods, which could potentially compromise the robustness of the comparisons, the feature selection techniques were exclusively applied to the other ML methods. In addition, given the predominant influence of the sex feature in PD and SC models, along with the suboptimal performance observed in these diseases in prior results, the analysis in this section will focus only on MS and AD.

In Figure 23 I show that in MS and AD there are no significant differences in sensitivity or specificity when comparing models after applying RFECV and RFE with the original models. In the case of the GB and LR methods applied to AD, there seems to be a slight improvement in specificity using RFE and RFECV compared with the original analysis (see Figure 23 (d)). Although the significance lied between a *p-value* of 0.05 and 0.1 according to a *Wilcoxon signed-rank test*.

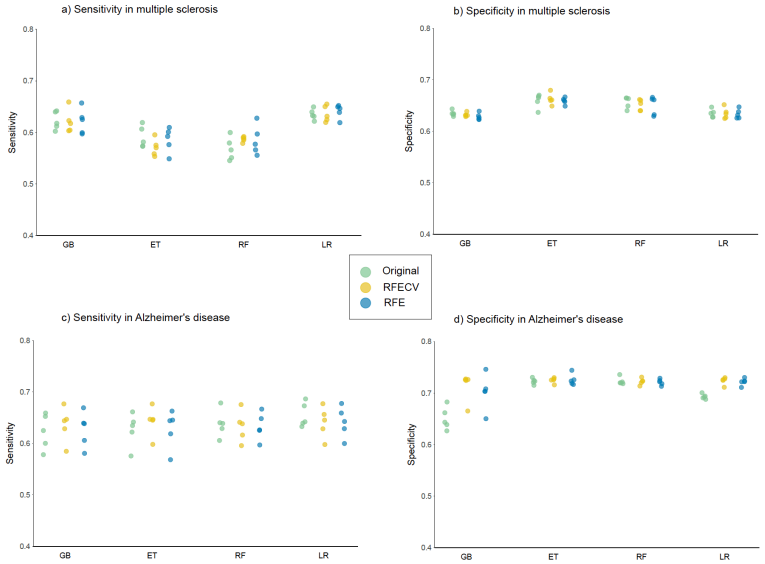

Figure 23 shows dot plots with the values of sensitivity and specificity in the original models, and the models after feature selection with RFECV and RFE. Sensitivity and specificity in MS are represented in plots (a) and (b), respectively. Sensitivity and specificity in AD are represented in plots (c) and (d), respectively.

In Table 11(a) I show that, in the case of MS, 74% to 88% of samples were classified with the same class using the features in the original analysis, RFE and RFECV. Notably, the value is particularly high for AD (see Table 11(b)) where, except for the GB method, around 93% of samples were classified with the same class using different sets of features. These results align with the similarity in performance observed in Figure 23, indicating that the original models, RFE and RFECV lead to the same prediction for the majority of samples.

a) Multiple sclerosis

	% samples with the same prediction in original, RFE and RFECV	Method	Features fold1	Features fold2	Features fold3	Features fold4	Features fold5	n Features 1 time	n Features 2 times	n Features 3 times	n Features 4 times	n Features 5 times
GB	78.31	RFE	150	250	250	250	200	54	33	40	65	120
		RFECV	314	138	252	321	341	19	29	56	124	125
ET	75.76	RFE	200	200	150	50	200	42	40	52	88	34
		RFECV	86	316	274	175	277	33	29	87	109	68
RF	74.10	RFE	100	250	100	100	200	62	66	38	28	66
		RFECV	225	103	216	79	251	30	49	85	44	63
LR	88.35	RFE	200	200	20	250	50	89	127	76	21	13
		RFECV	205	140	25	231	39	102	128	50	18	12

b) Alzheimer's disease

	% samples with the same prediction in original, RFE and RFECV	Method	Features fold1	Features fold2	Features fold3	Features fold4	Features fold5	n Features 1 time	n Features 2 times	n Features 3 times	n Features 4 times	n Features 5 times
GB	74.68	RFE	5	150	5	5	5	140	9	1	1	1
		RFECV	1	29	1	1	1	28	0	0	0	1
ET	93.02	RFE	150	5	100	100	5	34	35	75	4	3
		RFECV	32	1	1	2	1	30	1	0	0	1
RF	94.38	RFE	150	5	100	100	100	32	16	22	75	5
		RFECV	139	3	130	124	99	7	11	29	91	3
LR	92.23	RFE	20	5	20	50	20	29	21	10	1	2
		RFECV	7	1	7	51	3	44	4	4	0	1

Table 11 comprises two tables showing, for each ML method, from left to right, the percentage of samples that were classified with the same class in the original models and models after feature selection, the distinct methods used for feature selection, the number of features in folds from one to five after feature selection, and the number of features selected from one to five times in different folds. In (a), the table corresponds to MS. In (b), the table corresponds to AD, and the values in red correspond to the SNV rs429358, selected across all folds and methods in AD.

The number of features selected with RFE and RFECV in each fold is presented in Table 11 under the columns "features foldx". With few exceptions, in MS (Table 11(a)), the number of features selected by RFECV and RFE in each fold exceeded that in AD (Table 11(b)). Interestingly, for AD, RFECV selected only one SNV in eight folds (Table 11(b) highlighted in red). As noted in the column labelled "n features 5 times", 12 to 125 features were consistently selected across the five folds with the feature selection methods in MS, while AD had only 1 to 5 features selected.

In AD, the variant rs429358 (Table 11(b) highlighted in red) was the one consistently chosen across all folds using various feature selection techniques and ML methods, and it was the only feature selected in the eight different folds following RFECV selection. rs429358 is a SNV with the minor and major alleles being (C) and (T) respectively. This variant is located on chromosome 19 in the *Apolipoprotein E* (*APOE*) gene, and the allele (C) is one of the most extensively studied factors associated with AD

risk and dementia[159] showing an additive risk pattern. The fact that RFECV proposed models with the variant rs429358 alone in the case of AD, without any noticeable impact on model performance, suggests that the majority of predictions were entirely influenced by this variant in the original models.

To support this assumption, in Table 12 the allele frequency (AF) and the percentage of individuals with the rs429358 (C) allele present in the heterozygous or homozygous form are represented in “AF (C)”, “% (C;T)” and “% (C;C)” columns, respectively. The differences between controls, individuals with AD, individuals that were correctly classified as true positives across all ML methods (yellow bars in Figure 12(c)), and as true positives across ML and PRS methods (yellow bars in Figure 20 (a) and (c)) were explored. In controls, the AF of rs429358 (C) was 0.147 which is similar to the expected in the European population (1000 Genomes Europe C=0.155, as obtained from dbSNP[9]). In addition, the percentage of individuals with (C;C) alleles was very low (2%). Comparatively in individuals with AD, the presence of rs429358 (C) was more than the double than in controls.

AD individuals that were classified as true positives across all ML methods were 64% (C;T) and 36% (C;C) in Table 12, and with the exception of three AD subjects, all of them had at least one copy of the rs429358 (C) allele. In contrast, when looking at the AD individuals that were consistently classified as AD across ML and PRS methods, all of them had at least one rs429358 (C) allele, 91% of them had (C;C) alleles, while 9% had (C;T) alleles. Therefore, the AD individuals with the highest risk of developing the disease according to PRS and ML methods seem to match those having the (C;C) alleles, following a predicted additive risk pattern where individuals with (C;C) alleles have more chances of developing the disease than individuals with (C;T) alleles, and individuals with (T;T) alleles are likely to be classified as controls. With these results, I demonstrated that the models constructed for AD predominantly relied on a single SNV, in this case rs429358 (C), and that the high consistency observed in the classification of individuals across different methods for this disease is primarily attributed to this variant.

	AF (C)	% (C;T)	% (C;C)
Controls	0.147	25.34	1.99
AD	0.394	48.96	14.94
TP across ML	0.678	63.73	35.97
TP across ML and PRS	0.955	8.94	91.06

Table 12 shows, in columns from left to right, the allele frequency of the rs429358 (C) minor allele, the percentage of individuals with the heterozygous form of the allele (C;T), and the percentage of individuals with the two copies of the minor allele (C;C). In rows from top to bottom, there are controls, individuals with AD, AD that were correctly classified by the six ML methods, and AD that were correctly classified by the six ML methods and PRS. For each column, the color scale ranges from darker to lighter, indicating higher to lower values, respectively.

In the case of MS, the HLA variant *HLA-A*02:01* was the only genomic variant consistently selected across different folds and methods. However, I discarded the possibility that in MS the models relied only on this variant for making the predictions. This conclusion is supported by the fact that the number of features used in the models after feature selection was never just one in MS; instead, it ranged from 20 to 341 features, suggesting the presence of polygenicity.

As shown in Figure 23, the reduction in the number of features after using feature selection tools did not lead to a significant increase in performance. However, it was observed that reducing the number of features with RFE and RFECV had the effect of decreasing the number of correlated features due to LD. This trend is represented in Figure 24, where the Spearman rank correlation coefficient between the number of selected SNVs and the number of correlated pairs of SNVs with an $|r| > 0.7$ is presented for each ML method and feature selection technique. The correlation coefficients consistently exceeded 0.9, indicating a very strong correlation, with only few exceptions. For instance, in MS, the correlation was r=0.77 for RFECV and r=0.8 for RFE, which still indicated the presence of strong correlation. In the case of GB models applied to AD, a correlation of 0.56 was observed. However, this value is probably caused by the skewness in the number of features selected in AD with GB, with one or five features being selected in four out of five folds with RFECV and RFE, respectively (see Table 11). Overall, RFE and RFECV tools applied to ML models, which selected a reduced number of predictors with fewer correlated pairs as

illustrated in Figure 24, did not demonstrate a substantial decrease or enhancement in performance in Figure 23. These results suggest that the presence of correlation among the genomic variants did not significantly impact the model's performance.

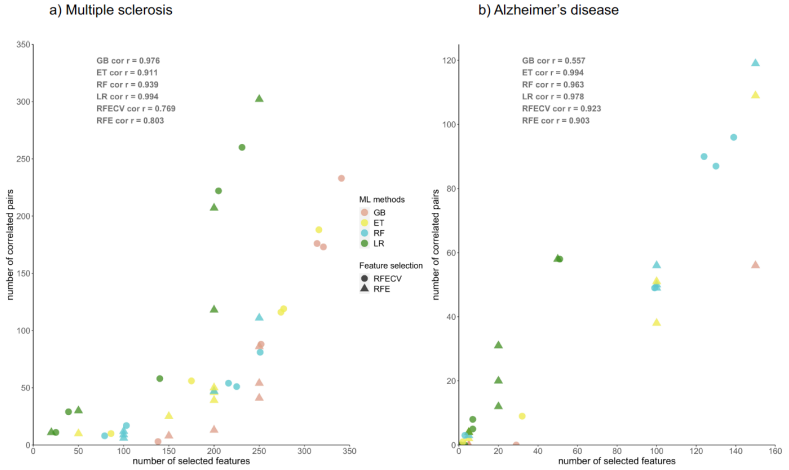

Figure 24 displays the relationship between the number of selected features and the number of correlated pairs of features after the application of feature selection tools. The plots are divided into MS and AD represented in (a) and (b) respectively. The correlation coefficients were calculated using Spearman correlation.

3.5 Variability in feature ranks

In the fifth section of the results, I applied explainability tools to extract the importance of the features assigned by the models. The goal was to compare the ranking of features across methods and check if genomic variants were ranked similarly, demonstrating consistent attribution of importance. To proceed with the use of explainability tools, I focused only on MS. AD was excluded from these analysis because, as demonstrated in the previous section, the classification for this disease heavily relied on a single SNV.

The plots in Figure 25 and Figure 26 are made by ranking the genomic features according to the importance assigned by the ML methods, and

comparing the ranking of genomic variants made by different methods and folds using Pearson correlation coefficients. These plots serve to represent the variability in the importance scores assigned to features and to assess whether models consistently ranked genomic variants in the same manner.

In Figure 25, the pairwise correlation of feature rankings obtained with different DL methods, XAI methods and folds is depicted. Different colours are assigned to the labels for each combination of fold and DL method. A clear distinction is noticeable between FFN and CNN, represented as two separate branches in the dendrogram, showing that the primary differences in the rankings are attributed to the choice of DL method. Higher correlations are observed across the folds in FFN, indicated by green labels and the lower red triangle in Figure 25, in comparison to the CNN folds. Feature ranks exhibited strong positive correlation, nearly reaching one, across the four XAI methods (layer Integrated gradients (LIG), layer deeplift (DE), saliency maps (SM) and guided backpropagation (GBP)), when applied to the same fold and DL method, as indicated by the five labels of the same colour always grouped together in the dendrogram, and the intense red small triangles distributed across the diagonal in the plot.

In Figure 26, the correlation across ML methods, including DL methods, is presented. Given the high similarity observed across different XAI methods applied to the same fold and DL method in Figure 25, only the LIG method was used to represent de ranks of DL methods for comparisons with the other ML methods. In Figure 26, different colours are assigned to labels depending on the ML method. A clear distinction emerges between the tree-based methods (GB, ET, and RF), coloured with green labels, and the other methods, as indicated by the intense red triangle in the lower part of the figure. Therefore, tree-based ML methods appear to rank variants in a similar manner. Conversely, LR and CNN exhibited more variability and less correlation across folds, even showing instances of negative correlation in some folds in comparison to the other tree-based methods and FFN, indicated by the blue squares in Figure 26.

These results evidence that tree-based methods exhibited relative low variability in the way they assigned importance measures to features. In contrast, LR and CNN exhibited unique feature ranking patterns across different folds, although it's worth noting that in the case of LR, this variability did not significantly affect the overall performance, as LR performed relatively well and had low standard deviation in the evaluation metrics across folds (refer to Table 4(a)).

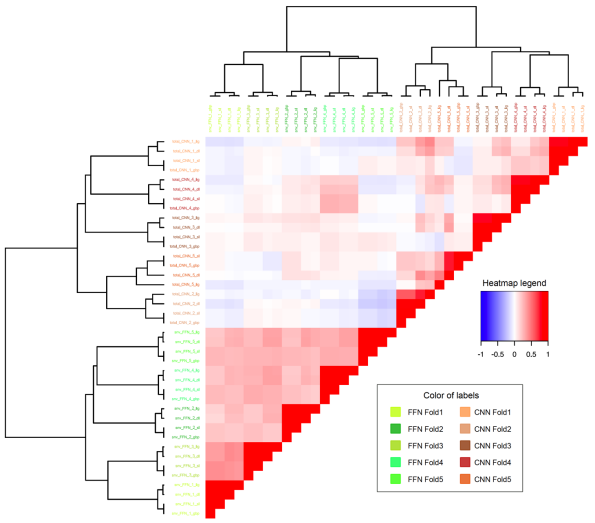

Figure 25 shows the pairwise Pearson correlation coefficients of the ranking of features obtained with layer integrated gradients (LIG), layer deeplift (DE), saliency maps (SM) and guided backpropagation (GBP) applied to CNN depicted with warm colors in labels, and FFN depicted with green colors in labels. The clustering in the dendrogram is made with Euclidean distances and ward-D2.

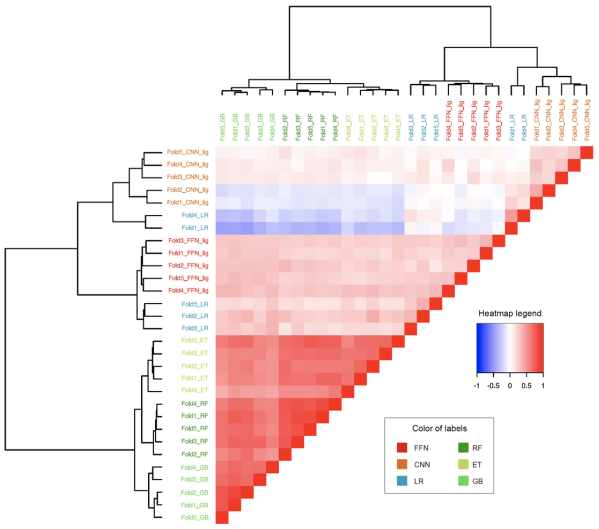

Figure 26 shows the pairwise Pearson correlation coefficients of the ranking of features obtained with different ML methods. LIG is used for DL methods. The clustering in the dendrogram is made with Euclidean distances and ward-D2.

3.6 Prioritized genomic variants in multiple sclerosis

As mentioned in previous sections, one of the advantages of using ML methods is that the attribution of importance to features is flexible and follows different approaches depending on the algorithm or architecture employed in the models during training. For this reason, after applying explainability tools, I was interested in identifying the genomic features that were considered most relevant for classifying MS and controls in the UKB cohort using the different methods.

The genomic variants were ranked with ordinal numbers, with values close to one representing higher importance. In total, there were 136 genomic variants that were among the top 10% with the best rank at least in one ML method, with 50 of them present on chromosome 6. From now on I will refer to them as prioritized variants. The enrichment of more than a third of the prioritized variants on chromosome 6 is consistent with the

previously reported MS hotspot in this chromosome, as obtained from the GWAS summary statistics used in the PRS calculation (refer to Figure 16(a)).

In Figure 27, I represented a circos plot with a heatmap depicting the ranks and respective locations of all the genomic variants used as features in MS. The genomic variants that were prioritized in at least one method are annotated with the name, excluding variants in chromosome 6. Due to the high density of prioritized genomic variants in chromosome 6, I depicted this chromosome independently in Figure 28, along with the names of the highest-ranked variants and the pairwise LD.

The heatmaps in Figure 27 and Figure 28 illustrate the substantial variability of ranks assigned to genomic variants, often displaying diverse colours corresponding to ranks obtained using different ML methods. Using AlphaMissense[160], a tool based on AlphaFold that predicts the impact of SNVs on the protein structure, missense variants were annotated with their predicted effects: ambiguous, likely benign, or likely pathogenic. Notably, all missense variants used as features in the MS models were annotated as likely benign, which aligns with the polygenic nature of MS, wherein the cumulative effect of numerous small genetic effects across the genome predisposes or protects against the disease[161]. Alternatively, most of the SNVs were predicted to have an effect in expression (eQTL) or splicing (sQTL) as annotated using GTEx and highlighted with purple colour in the labels of the SNVs in Figure 27 and Figure 28.

Three prioritized variants were missense SNVs, highlighted with green labels in Figure 27: rs6897932 located in the *IL7R* gene of chromosome 5, rs763361 located in the *CD226* gene of chromosome 18, and rs5771069 located in the *IL17REL* gene of chromosome 22. The *IL7R* gene encodes the interleukin-7 receptor, involved in the development and function of T cells. *CD226*, on the other hand, encodes a glycoprotein also known as *DNAX accessory molecule-1* (*DNAM-1*), which plays a role in the regulation of T cell activation and the immune response. In addition to MS, genetic variants in *IL7R* and *CD226* have been associated with other autoimmune diseases, such as type 1 diabetes and rheumatoid arthritis[162]

[163] [164] [165]. Notably, there was an absence of prioritized missense variants in chromosome 6.

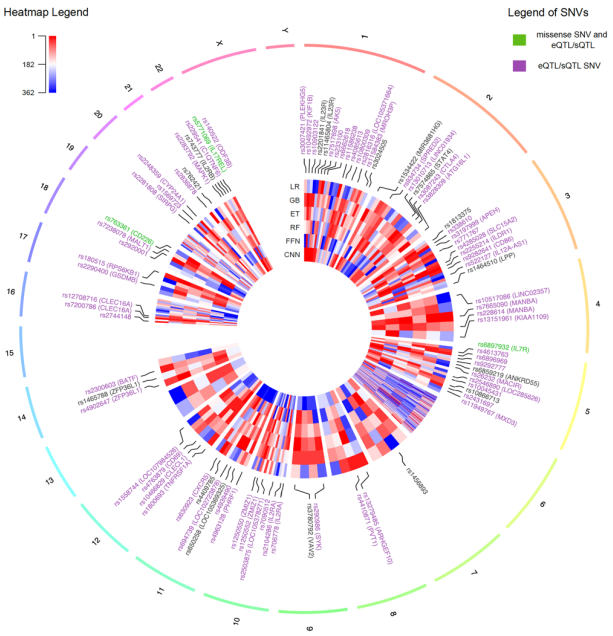

Figure 27: Circos plot representing all the genomic features used in the MS models distributed across the genome. The heatmap indicates the ranks of the features as assigned by each ML method, with values close to one in red indicating higher importance. The variants that were prioritized by at least one method are indicated with their names. The names of the SNVs are colored in purple if they are annotated with an eQTL or sQTL in at least one tissue in GTEx. The labels of missense SNVs with annotated QTLs are colored in green. The labels of chromosome 6 were excluded due to the high density of prioritized genomic variants in this chromosome.

The top ten best-ranked genomic features in chromosome 6 were determined by summing the ranks obtained with the six ML methods and are labelled in Figure 28. *HLA-DRB1*15:01* is in close proximity to other prioritized HLA variants, *HLA-DRB1*03:01* and *HLA-DRB5*Null*, collectively pointing to a well-documented MS-related locus in the cytoband 6p21.32 [166]. However, *HLA-DRB1*15:01* and *HLA-DRB5*Null*

exhibited a strong LD (r^2=0.93), making it challenging to distinguish between these HLA types in terms of their association with the disease. In fact, the two variants located in the 6p21.32 cytoband show similar LD patterns with other genomic features, as shown in Figure 28. *HLA-C*04:01* and rs2524089 are located in the 6p21.33 cytoband. *HLA-A*02:01*, and rs2523393 situated in the *HLA-F* gene, both belong to the 6p22.1 cytoband. Finally, the variants rs17119, rs10806425 and rs17066096 are found in cytobands 6p23, 6q15 and 6q23.3, respectively, with the last two situated in the long arm of the chromosome. Therefore, there is not only one location in chromosome 6 associated with MS, instead, the top ten risk loci are widely distributed.

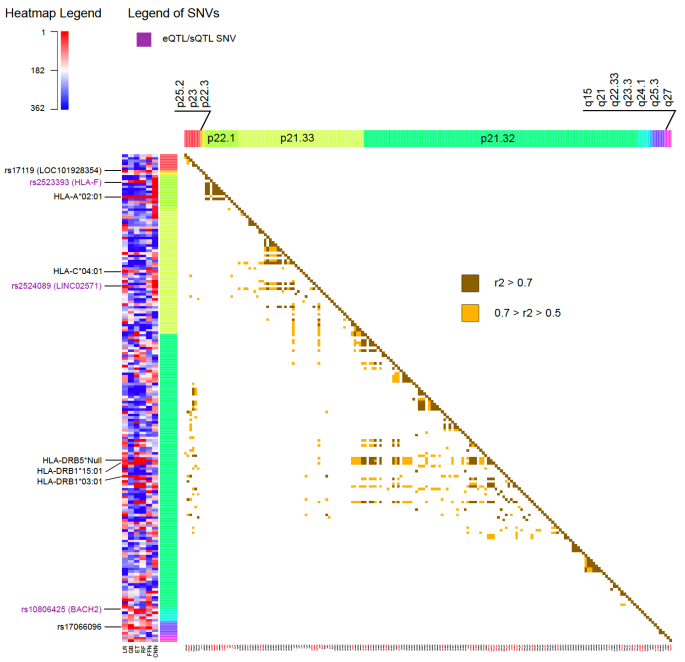

Figure 28: The heatmap on the left represents the ranks of all the features on chromosome 6 as assigned by each ML method, with values close to one in red indicating higher importance. The top ten best-ranked genomic variants on this chromosome are labeled with their corresponding names. Labels in purple indicate the presence of QTLs in at least one tissue in GTEx. The heatmap on the right indicates the presence and strength of LD between pairs of genomic variants.

Among the top genomic features in chromosome 6 there were five HLA types: *HLA-A*02:01*, *HLA-C*04:01*, *HLA-DRB5*Null*, *HLA-DRB1*15:01* and *HLA-DRB1*03:01*. Additionally, the SNV rs2523393, located in *HLA-F*, was identified as an eQTL and sQTL for this gene. Furthermore, the SNV rs2524089, an intron variant in *LINC02571*, was recognized as an eQTL and sQTL for the genes *HLA-B*, *HLA-C*, and *HLA-E*. In this regard, the prevalence of HLA gene annotations among the top genomic features on chromosome 6 highlights their significance in the context of MS.

The top ten best-ranked genomic features across all chromosomes obtained by summing the ranks from the six methods are listed in Table 13. This table mirrors the information presented in Figure 27 and Figure 28, highlighting the variability in rankings across methods, and demonstrating that none of the top ten genomic features consistently earned the status of prioritized variant (highlighted in red) across all methods. Some of the genomic variants annotated in Figure 27 and Figure 28, which did not rank among the top ten in Table 13 and were not previously mentioned in the text, showed evidence from other studies supporting their association with MS. Nevertheless, given the extensive number of prioritized variants in this disease, I only delved into the top ones in greater detail in the following lines.

When considering all chromosomes, the highest-ranked genomic variant was *HLA-A*02:01* on chromosome 6. In the UKB cohort, *HLA-A*02:01* was more frequent in controls compared to individuals with MS, with a Fisher test *p-value* of 2.43E-19. This observation aligns with its reported protective effect against MS in the literature[167] [168] . *HLA-A*02:01* was also recurrently selected across all folds and methods with the RFE and RFECV techniques in the previous section “3.4 Implementation of feature selection techniques”, emphasizing the relevance of this variant for predicting MS outcomes in the UKB cohort. The *HLA-A* gene belongs to the MHC class I, a group of cell surface proteins that play a crucial role in the immune system, recognizing intracellular pathogens and distinguishing between self and non-self cells. It is worth noting that the most significant genetic factor associated with MS, as reported in the literature, is *HLA-DRB1*15:01*[169], a predisposing HLA variant belonging to the MHC class II. In the UKB cohort, *HLA-DRB1*15:01* exhibited the most

significant differences in allele frequency between individuals with MS and controls, with a Fisher test *p-value* of 2.77E-101, being more prevalent in cases than controls, consistent with its predisposing role. However, when considering rankings across all chromosomes, this variant was ranked 23rd and, therefore, does not appear in the top ten in Table 13.

The SNVs rs7665090 and rs2248359 listed in Table 13 are located downstream and upstream of the genes *MANBA* and *CYP24A1*, respectively. Specifically, rs7665090 serves as both an sQTL and eQTL for *MANBA*, while rs2248359 functions as an eQTL for *CYP24A1*. *MANBA* is an exoglycosidase found in the lysosome and is present in immune system pathways. Notably, the variant rs7665090 has been linked to a reduction in *MANBA* transcript expression and enzymatic activity, along with the occurrence of neurological abnormalities and recurrent infections[170]. On the other hand, the *CYP24A1* gene encodes a protein involved in the catabolism of the active form of vitamin D. There is genetic and epidemiological evidence suggesting that vitamin D insufficiency contributes to MS. In this regard, the expression of *CYP24A1* and other genes associated with MS risk in peripheral blood indicates a response to vitamin D and showed different expression patterns in individuals with MS compared to controls in a published study[171].

The SNV rs180515 is situated in the 3' UTR of *RPS6KB1*. This SNV is also annotated as both an eQTL and sQTL for *RPS6KB1*. This gene is actively involved in immune response pathways, particularly in the IL-4 signalling pathway, which has been associated with the progression of MS in several studies[172] [173] [174] [175]. Moreover, another study revealed an upregulation of *RPS6KB1* transcript expression in whole blood samples from Iranian patients with MS when compared to healthy controls[176].

The variants rs1800693, rs2283792, and rs7200786, are located within the intronic regions of the genes *TNFRSF1A*, *MAPK1*, and *CLEC16A*, respectively. The SNV rs1800693 functions as both an eQTL and sQTL for *TNFRSF1A*. This gene encodes a member of the TNF receptor superfamily of proteins and is known to play a role in regulating the immune system and the initiation of inflammatory reactions[177]. Also, rs1800693 has been consistently linked to MS in various studies and is

hypothesized to influence the magnitude of monocyte responses to TNF-α stimulation[178] [179]. The SNV rs2283792 serves as an eQTL for *MAPK1*, which is linked to MS due to its involvement in the MAPK pathways. Notably, studies have shown that the overactivity of the MAPK ERK pathway in microglia can indirectly lead to demyelination, a defining characteristic of MS[180]. The SNV rs7200786 is both an eQTL and sQTL for *CLEC16A*. SNVs in the *CLEC16A* gene, specifically within its intronic regions, were some of the earliest non-HLA genetic variants to be established as having an association with MS[181] [182]. These SNVs have also been linked to other autoimmune diseases such as type 1 diabetes, rheumatoid arthritis, and primary biliary cirrhosis[182].

In the seventh position of Table 13 there is rs11586238, which is situated in an intergenic region on chromosome 1. This SNV serves as an eQTL for the *CD101* gene, which encodes a protein expressed on various immune cell populations. While the connection between *CD101* and MS remains unclear, it's worth noting that the transcripts of this gene have been observed to be upregulated in monozygotic twins with prodromal MS[183].

Regarding the remaining SNVs in Table 13, specifically rs2255214, an intronic variant in the *ILDR1* gene, and rs4285028, located in the 3' UTR of the *SLC15A2* gene, I was unable to find published works indicating their molecular association with MS in the literature.

dbSNP ID	Gene	Chromosome	LR Rank	GB Rank	ET Rank	RF Rank	FFN Rank	CNN Rank	Sum of Ranks
HLA-A*02:01	HLA-A	chr6	1	9	8	8	12	46	1
rs2255214	ILDR1	chr3	10	17	15	13	1	56	2
rs7665090	MANBA	chr4	16	5	112	18	32	54	3
rs180515	RPS6KB1	chr17	27	12	29	58	29	92	4
rs1800693	TNFRSF1A	chr12	35	93	36	29	3	68	5
rs2248359	CYP24A1	chr20	18	6	126	64	4	49	6
rs11586238		chr1	4	82	35	54	48	127	7
rs2283792	MAPK1	chr22	78	48	113	99	23	9	8
rs7200786	CLEC16A	chr16	17	27	17	30	42	239	9
rs4285028	SLC15A2	chr3	22	13	144	34	6	156	10

Table 13 lists the top ten best-ranked genomic features across all methods with the corresponding ranks assigned by each ML method. The values of the prioritized ranks are highlighted in red.

Overall, these results remark the polygenicity of MS and the variability in the assigned feature ranks across different methods. The majority of the highest-prioritized variants were identified as eQTL or sQTL located in

non-coding regions within or near genes associated with the immune response and MS. This observation supports the notion that the risk of MS is primarily influenced by many subtle alterations in gene regulation that gradually accumulate over time, eventually driving the system into a pathological state, instead of missense variants that would have a major predicted impact on protein structure.

Several SNVs were prioritized by the models but were not annotated as missense variants, eQTLs, or sQTLs affecting relevant genes. This could be partially attributed to the presence of LD, which results in highly correlated genotypes for variants located close together. While this correlation among features confers similar predictive power in the models, it does not necessarily imply that each individual variant is relevant to MS; it only indicates that at least one of the genomic variants in LD might be. It is important to note that prioritizing genomic variants on chromosome 6 associated with MS is challenging. This is because, in addition to the presence of LD, a large amount of these variants are located near the MHC class I and class II genes, considered the most polymorphic region in the human genome[28]. These difficulties will be further developed in the discussion.

3.7 Synergies among the prioritized genomic variants in multiple sclerosis

In this study, I tried to identify the synergistic effects that exist among the prioritized genomic features. This is because the ML methods I used, with the exception of LR, have the ability to capture complex patterns involving interactions.

To explore this, I tested all the possible pairwise interactions among the genomic features prioritized by each model independently, and used an harmonic mean p-value (HMP) cut-off of less than 0.01 to select statistically significant interactions. More details on the methods employed for the statistical test of interactions can be found in the section 5.5 of the Methods. I excluded interactions where the pair of genomic features exhibited correlation due to LD, with an r^2 greater than 0.2, as well as interactions involving different types of genomic features, such as the

combination of SNV and an HLA type. This decision is based on the fact that HLA types were imputed based on SNVs located on chromosome 6, and therefore, the two types of features are potentially confounded.

After identifying the most significant interactions, I mapped them into single proteins or protein complexes, thereby adding molecular annotations to these interactions. Figure 29 represents the types of interactions with molecular annotations in red, while those without annotations are shown in blue. It is important to note that the interactions without molecular annotations do not necessarily lack a molecular context. Conversely, these interactions may be part of indirect synergies where the proteins do not directly interact in complexes, and as a result, went unnoticed using the approach I employed.

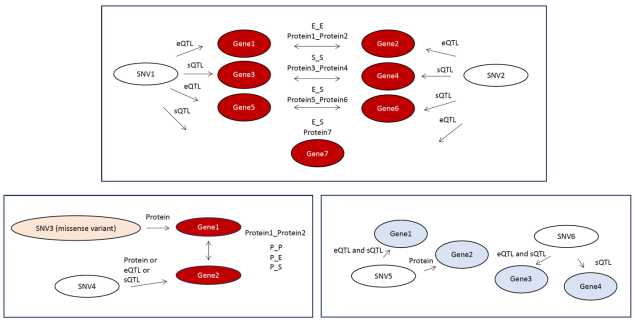

Figure 29 exemplifies the different scenarios in which the interactions between genomic variants are characterized with molecular annotations (in red), and those interactions lacking molecular annotations (in blue).

In Figure 30, I present the number of significant pairwise interactions between genomic variants for each method, highlighting those with molecular annotations in red. Among the prioritized genomic variants, there were twelve significant pairwise interactions for ET, making it the method with the highest number of interactions. LR followed with nine interactions, FFN with seven, GB with five, and CNN with two. There were no interactions found among the prioritized variants with RF.

Surprisingly, LR, which is not explicitly designed to capture interaction patterns, was the second method in terms of the number of interactions among the prioritized variants. This is likely because, in the case of LR, the individual predictive effects of the genomic features involved in the interactions were still big regardless of any synergistic effects. Consequently, these features were probably prioritized due to their independent predictive power.

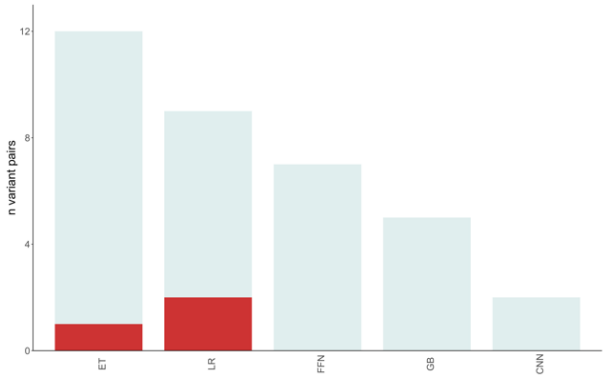

Figure 30 shows the number of significant pairs of interactions selected for each ML method using an HMP lower than 0.01. The interactions with molecular annotations are highlighted in red.

There were three interactions with molecular annotations highlighted in red in Figure 30. One of these interactions was identified by ET, involving the SNVs rs6903608 and rs3130299, both located in intergenic regions in the 6p21.32 cytoband. The other two interactions were detected by LR, and both involved the SNV rs760293, an intron variant in *BAG6* gene in cytoband 6p21.33, interacting with rs615672 and rs3135363, two intergenic SNVs in 6p21.32. Notably, rs615672 and rs3135363 exhibit moderate LD, with a r^2 of 0.36. Therefore, it is unclear whether the two interactions detected in LR are entirely independent.

Interestingly, all the SNVs involved in the interactions with molecular annotations were located at cytoband 6p21.3, which, as noted in the previous section, is the strongest MS susceptibility locus identified

genome-wide. Additionally, these SNVs were predicted to act as eQTL and sQTL for many other genes, especially HLA genes. However, with the exception of one SNV, rs760293 in the *BAG6* gene, the remaining SNVs involved in the interactions with molecular annotations were located in intergenic regions, and establishing the link of these intergenic genomic variants with the regulation of gene expression and splicing is particularly challenging. In this context, even if those variants were annotated with QTLs, it is important to clarify that these annotations may result from being in LD with other variants located in regulatory regions of disease-related genes.

In Figure 31 I represented all the genes potentially involved in interactions at the protein level based on the molecular annotations associated with the three pairs of SNVs described earlier. All these genes were situated on chromosome 6, and nearly half of the annotations were linked to genes from the HLA family. Furthermore, there were annotations of complement factor genes, including *C4A, C4B*, and *CFB*, which are also part of immune system pathways.

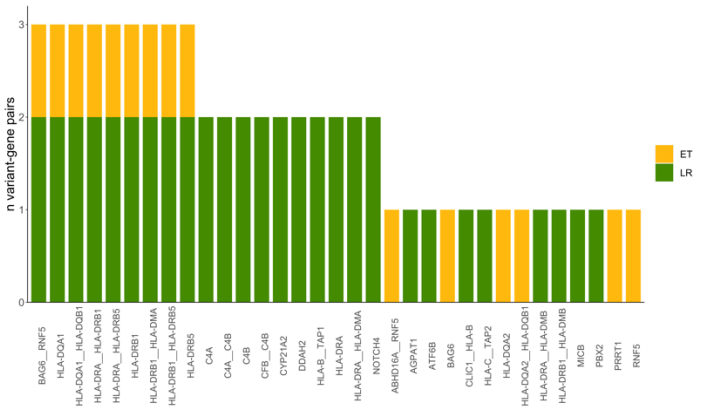

Figure 31 shows the genes involved in the interactions with molecular annotations. In yellow and in green, interactions found among the prioritized genomic variants by ET and LR are depicted, respectively. The y-axis indicates the number of different interacting pairs of genomic variants.

In summary, the strategy I employed to select interactions among the prioritized variants did not reveal a significant enrichment of statistically

significant interactions in GB, RF, FFN or CNN methods compared to LR, which is not designed to capture epistatic events among predictors. Among these methods, ET had the highest number of statistically significant interactions among the prioritized predictors. However, based on the results obtained, I lack sufficient evidence to demonstrate that tree-based methods or DL methods prioritized variants due to their involvement in epistasis, and that these interactions conferred an advantage over LR in the classification.

4. Discussion

In this work I investigated different aspects concerning the application of ML methods for predicting complex diseases based on genomic features. It is important to acknowledge that assessing genomic predisposition to complex diseases, which do not adhere to classic Mendelian inheritance patterns, present challenges. In this regard, ML methods have the ability to identify complex relationships in the data, and higher order interactions, that the traditional statistical methods may overlook. Consequently, I hypothesized that these tools could perform well at classifying individuals with complex diseases.

A general rule in ML methods is that increasing sample size enhances model reliability[184 185]. For that reason, I selected four complex diseases for the analysis —multiple sclerosis (MS), Alzheimer's disease (AD), schizophrenia (SC), and Parkinson's disease (PD)—from the UKB, each with over 900 cases.

Evaluation of the performance of ML models

The performance of models was evaluated and compared across folds, methods, and diseases. Lower variability across folds is often desirable in ML analyses, as it suggests more stable and reliable performance of the models. Notably, DL methods exhibited the highest variability in performance. This could be attributed to the relatively modest sample size employed in this study, posing challenges for generalization, especially when leveraging the deeper connections inherent in DL models. Related to this, several studies have suggested that traditional ML methods tend to outperform DL methods when dealing with small sample sizes[184 186].

Genomic variants located in proximity are usually correlated due to LD. Therefore, I hypothesized that the spatial representation of genomic variants and the use of convolutional layers in CNN may help to disentangle the information present in hot spots with highly correlated features. However, the inclusion of information on chromosome and position in CNN models did not seem to improve the results compared to other ML models.

Another important limitation of DL methods is that, using the same input dimensions, they take longer times to run compared to the other ML methods, even when employing graphics processing units (GPUs), which are frequently used to reduce computational times. Also, the flexibility of DL methods comes with an elevated number of tuneable parameters, making hyperparameter selection complex. In this work, I used the grid search approach for hyperparameter tuning, which involves defining a set of values for each hyperparameter and testing all possible combinations. While this method is exhaustive and often identifies optimal hyperparameter sets, it is computationally expensive, especially with a large hyperparameter space, and time-consuming when applied to DL methods. Therefore, for the purpose of efficiency, simple ML methods, excluding DL methods, are likely more suitable for disease classification with sample sizes similar to this study. Nevertheless, it is important to stress that these observations could be context-specific and may not apply universally in all cases.

Conversely, LR exhibited stable performance across folds and diseases, and was consistently positioned among the top-performing methods. LR is known for its relatively simple algorithm and ease of implementation. The popularity of LR stems, also, from its ability to perform well with only one or two tuned hyperparameters[63]. In addition, the marginal effects estimates, which are defined as the impact of small changes in specific predictors on the probability of the outcome variable, are less influenced by the limitations of a small sample size in LR[187]. Nevertheless, regression models such as LR, by default, are designed to detect linear additive associations, preventing them from capturing interactions between any two input variables. In this respect, this limitation did not seem to negatively impact the results of the current work.

Methodologically, tree-based ensemble methods (GB, ET, and RF) differ in how they introduce randomness during tree construction and in how they combine the predictions of individual trees. These variations may lead to differences in their performance. In this study, for example, tree-based ensemble methods exhibited variability in evaluation metrics within and across diseases. In AD, RF demonstrated the best performance in the comparison across methods, followed by ET, while GB ranked as the

second-worst method. Conversely in MS, GB was the second-best method after LR, and outperformed RF and ET. This variability is a factor that should be considered when choosing the ML methods to test in the studies.

In addition, tree-based ensemble methods have been shown to work better with tabular data and with small sample sizes in other studies, whereas DL methods performed better on structured data with larger sample sizes[188]. This is also observed in the current study using tabular data, where GB, ET and RF generally showed to be more stable and have better evaluation metrics compared with FFN.

When comparing across diseases, AD exhibited the best classification results, with balanced accuracy values ranging from 0.63 to 0.69, followed by MS with values from 0.61 to 0.64. Conversely, the performance of the other diseases, SC and PD, fell below balanced accuracy values of 0.6 across all methods. These differences in predictiveness may be attributed to various factors. Notably, SC had the smallest sample size of cases in UKB, with 988 cases compared to the 2020, 2490, and 3126 in MS, AD, and PD, respectively. Additionally, SC and PD are diseases difficult to diagnose due to several factors. The complexity and heterogeneity of symptoms, the overlapping nature of symptoms with other conditions, and the challenges associated with distinguishing SC and PD from coexisting conditions contribute to the difficulty of an accurate diagnosis[189 190]. This may result in subjects being misdiagnosed and incorrectly tagged as cases, negatively influencing the ability of ML models to discover generalizable genomic patterns associated with these diseases. Stratifying PD and SC subjects based on different symptoms and disease courses, with diagnoses from experts, instead of grouping all subjects into the same disease category obtained from clinical records, may help reduce heterogeneity and improve the results presented here. However, for stratifying individuals within the same disease, ideally, a larger sample size and an accurate diagnosis would be required.

As described in the introduction, overfitting is a common problem when using ML methods. In this work I used nested CV in order to prevent overfitting, which has been proven to give good results in other studies[55].

Importantly, the higher the ratio of features to sample size, the more likely it is that a model will fit the noise in the data instead of capturing the underlying patterns associated with the disease, and consequently, there is a higher risk of overfitting[156 8]. Also, higher dimensionality significantly increases computation times. These facts support the decision I made to include in the ML models only the genomic features reported to be associated with the disease in curated databases, rather than using all the genomic variants in the genotyping array. Alternatively, dimensionality reduction techniques, such as principal component analysis, can be used to decrease the number of features. However, when applying these techniques, the original features are transformed into a new set of variables, making it impossible to trace back the individual effect of each genomic variant. This limitation would have prevented the subsequent application of XAI tools to rank variants by importance, and the prioritization of the most informative genomic features in the classification.

To ensure that the models did not exhibit major overfitting, an external validation was performed on diseases with the best performance, MS and AD, using datasets obtained from independent studies. The models were tested on cohorts with individuals from the United States (US) and the United Kingdom (UK).

For MS, the performance in the external validation cohorts was either equivalent or superior to that in the UKB cohort. In the IMSGC MS cohort comprising individuals from UK, sensitivity was even better than in the UKB cohort with all methods, except for the CNN method, which showed high variability across folds. As discussed in the results, this may indicate that individuals with MS were more accurately diagnosed in the study conducted by the IMSGC, which was a MS dedicated study[154]. This advantage is likely enhanced with the IMSGC MS cohort, given that together with the UKB cohort used for training, both cohorts comprised individuals from UK.

Regarding AD, the performance of ML models did not worsen when tested on individuals from US in ADNI, demonstrating the ability of the AD models to generalize to cohorts with US population. Nevertheless, caution should

be taken when interpreting the AD results, as ML models heavily relied on a single SNV for making predictions.

Overall, the results on the external validation datasets were positive as they not only demonstrated that the performance did not worsen in other cohorts, but also, in some tests with the IMSGC cohorts the performance was better. However, as noted in the results, the evaluation in the IMSGC dataset was incomplete, as only the positive class could be tested due to the nature of the IMSGC study.

Influence of the variable sex in the model predictions

As the sex feature was recognized as a relevant factor influencing the outcome and progression of some of the diseases under study[141 155 191 192], I was interested in determining if models performed better in predicting one sex over the other, particularly for diseases showing the highest sex bias, MS and PD. In this context, when using the same sample size and independent models for females and males, males with MS exhibited greater specificity than females with the GB method. However, no significant differences were observed between both sexes in the remaining combinations of disease and ML methods, hindering the formulation of conclusive findings.

In addition, to assess the significance of the variable sex in the decisions made by the models, I constructed models only using the sex feature and compared the results with the original models. Upon doing so, I observed that for diseases with the lowest genomic predictiveness, such as PD and SC, the classification was predominantly influenced by the sex feature. In this study, where I aimed to evaluate the predictive power of genomic variants for complex diseases, this fact may lead to confusion, as in PD and SC, the genomic features appeared to have a low influence on the decisions made by the models.

Comparison of ML methods with PRS

Another relevant aspect of using ML methods is to investigate how they compare to PRS, which is the most widely used tool in population genomics to quantify an individual's genetic predisposition to a disease based on multiple genomic variants. Despite being applied to solve similar

problems, PRS differ from ML methods in several aspects, one of them being the approach for attributing importance to the genomic features. PRS are computed using statistical methods, specifically a linear additive model that involves summing the effect sizes of multiple genomic variants associated with the disease of interest. In the case of PRS, the weights assigned to genomic variants are derived from GWAS summary statistics. In contrast, in ML methods, the importance assigned to genomic variants is defined during training.

In this study, the *p-values* assigned to genomic variants for the PRS calculation, indicating the significance of their association with the disease, and sourced from GWAS summary statistics, were generally lower in AD and MS compared to SC and PD, suggesting a stronger genetic risk association in the former two diseases. Therefore, the statistics derived from independent GWAS studies aligned with the different performances observed across diseases with ML methods in the UKB cohort.

When comparing performances at the 99^{th} percentile with individuals ranked based on the continuous values quantifying disease risk, PRS consistently demonstrated average performance compared with the other methods, but never clearly reached the best performance. In the case of MS, for example, PRS ALL models including all the genomic variants in the array had a mean value of RR equal to 4.0, compared to the best performing method, FFN, which had an RR of 4.1, also achieving the lowest standard deviation across folds.

In fact, the case of FFN in MS is particularly interesting because, contrary to the results obtained at the 99^{th} percentile, it was the method exhibiting the highest variability across folds when considering the default cut-off of probability 0.5 to classify cases and controls. This fact indicates that FFN was more robust at classifying MS individuals with the highest probabilities to develop the disease. As suggested in the Results section 3.3, setting a flexible cut-off for classifying cases and controls, instead of using the default cut-off of 0.5, may, in some cases, reduce the variability found in FFN.

Once the performance across different methods was compared, and the conclusion was drawn that PRS was comparable to ML methods, the question arises of whether to select one method over the other. In this context, there are several advantages and disadvantages summarized in Table 14 and discussed in the following lines.

Firstly, PRS are limited to capturing only linear relations with the disease, as its core algorithm is a linear additive model. In contrast, ML methods, with the exception of LR, can capture complex interactions and nonlinearities. This ability could be especially valuable for detecting synergisms between genomic variants that may go unnoticed with PRS. Additionally, in the case of ML, the interpretability of the model is more flexible compared to PRS. This is because PRS provide a risk score based on a set of genomic variants with their associated weights obtained from GWAS summary statistics, and these weights remain unmodified. However, as mentioned earlier, ML have the capacity to learn from the data in the training set and refine the weights assigned to genomic variants. Finally, ML methods can produce a variety of outputs including the predicted class for each individual and its associated probability. As probabilities are easy to interpret and range from 0 to 1, the classification of the same individual can be directly compared across methods and experiments. In contrast, PRS themselves are not classifiers, and their output is a single numerical score for each individual, representing the cumulative genetic risk for a specific trait or disease. However, this score alone cannot be compared across different PRS studies, and the risk of individuals developing the disease is typically interpreted using quantiles, with the highest PRS values indicating higher genetic predisposition to the disease[44].

PRS	ML
It captures linear relations with the disease and does not search for genomic interactions	Have the ability to capture complex genomic patterns
The interpretability of the model depends on the GWAS summary statistics	The interpretability of the model can be more flexible
A risk score is a relative value and cannot be directly compared across different models	Models output probabilities that can be directly compared with the output of other models and methods
There is no need for a large dataset with individualized genomic data to build the models. Only GWAS summary statistics are required	Large datasets with individualized genomic data are needed to build robust models
You can use as many genomic variants as you want	Using too many features is problematic; a pre-filter is required to avoid dimensionality issues
It takes a short time to run	Some models take a long time to run, especially DL models
PRS models typically exhibit stable performance, with their performance falling within the average range when compared across other methods	The performance varies across ML methods depending on the context in which they are applied

Table 14 lists the advantages (in green) and disadvantages (in red) when using PRS or ML methods.

Conversely, several reasons can explain why PRS is the most widely used tool for disease risk stratification among the scientific community working on population genomics and has not been replaced by ML methods yet. One of the most important advantages of PRS is that there is no need to access large datasets with individualized genomic data to build the models. This is because PRS require GWAS summary statistics instead of individualized genomic data, and summary statistics are typically anonymized to protect the privacy and confidentiality of the study

participants. Therefore, there is no need to restrict access to such data with specific privacy and legal clauses, facilitating their public availability[6]. Additionally, in PRS there are no limits to the number of genetic variants to include in the models, as the scores of the variants are calculated as weighted counts of thousands of risk variants identified in GWAS. Therefore, some of the problems associated with the dimensionality of the data in ML are resolved in PRS. For example, the computation times of PRS will generally be reasonable regardless of the number of genomic variants used for the calculations due to its relatively simple core algorithm. This fact makes PRS accessible to scientists who do not have access to high-performance computing (HPC) platforms or computer servers with high capacities. Finally, PRS fell within the average performance when compared across methods and diseases. In contrast, the performance of ML methods was more variable, especially across tree-based ML and DL methods. It is worth noting that, apart from the methods used in this work, there is an extensive list of other ML methods that could be employed as classifiers, but it was not feasible to test all of them during my project. The uncertainty regarding the best ML method to use to solve a particular problem creates the necessity of testing several methods in the same study, adding complexity to the process of analysing the data. In this regard, PRS represents a safe choice in most of the cases, without the requirement of testing different methods, and with the support given by years of usage and published studies.

Implementation of feature selection techniques

In this study I employed different feature selection tools, specifically two recursive feature elimination methods, RFE and RFECV. The objective of testing these tools was to assess whether there exists a subset of predictors that could enhance the model's performance. Given that I chose genomic variants linked to the disease from curated databases to be predictors in the models, it is possible that some of these variants may not be informative for classifying individuals in the UKB cohort. As a result, the inclusion of non-informative features might introduce noise and impede the accurate classification of individuals.

Furthermore, some of the genomic predictors exhibited strong correlation due to LD, and the high correlation among multiple variants poses

challenges in identifying the specific variants associated with a disease or trait. In regression models, the incorporation of two or more features with a high degree of correlation is known as multicollinearity. For methods based on regression models, such as LR, multicollinearity can make it challenging to accurately estimate the true coefficients of the features[193], affecting the stability and robustness of the model. In this context, the use of RFE and RFECV tools is recommended. In the case of ensemble tree-based methods such as GB, ET, and RF models, even though these methods are more robust to feature redundancy, the presence of large groups of correlated features in the training data can generate misleading feature rankings[194]. Consequently, the use of RFE and RFECV could help optimize the attribution of importance to features. Finally, in DL, the direct application of recursive feature elimination techniques is less common. This is because DL models often have a large number of parameters and can learn intricate hierarchical representations, making feature selection less of a concern[195 196]. Given this circumstance and considering the high variability in the performance of DL methods across folds, I did not apply RFE and RFECV to these methods.

After applying RFE and RFECV, the number of features in the models decreased to varying extents, along with a reduction in the number of correlated features. Interestingly, there were no significant changes in the performance of the models after feature selection. The absence of performance improvement, despite the reduction in correlated features, suggests that the LD present among genomic variants did not have a major impact on the performance of the original models. This robustness towards LD was particularly unexpected for the LR method. As previously noted, methods based on regression models are typically unstable in the presence of multicollinearity. However, in this work, LR exhibited low variability across folds in the original models and showed no apparent improvement after the application of feature selection techniques.

In addition, the models developed for AD predominantly relied on a specific SNV, namely rs429358, for making predictions. After the application of RFECV techniques, some models exclusively depended on this genomic variant for classification, with no significant impact on performance. In fact, the strong consistency observed in classifying individuals with AD across

different methods was primarily due to the presence and influence of the rs429358 variant. Located on chromosome 19 in the *Apolipoprotein E* (*APOE*) gene, the allele (C) of rs429358 is one of the most extensively reported factors associated with AD risk and dementia, exhibiting an additive risk pattern[159]. In this regard, the recurrent selection of rs429358 across methods and folds, which also has substantial supporting evidence of an association with AD risk in the literature, over other variants in strong LD such as rs4420638 (r^2=0.708) and rs769449 (r^2=0.743) with less disease evidence, underscores the capability of RFE and RFECV to discern genomic variants with the most significant links to the disease, despite the presence of feature correlations.

The strong association between rs429358 (C) in *APOE* and the disease might overshadow the contributions of other weaker genetic risk factors in the AD models. Related to this, when trying to account for additional genomic variants conferring small risk effects to AD, it is a common practice to exclude the *APOE* region from GWAS and PRS calculations, treating the *APOE* locus as an independent factor or covariate[108 197 198].

In contrast, for MS, there was a recurrent selection of the HLA variant *HLA-A*02:01* across all folds and methods. However, none of the models entirely depended on this variant for predictions, as was the case with rs429358 in AD, supporting the presence of polygenicity in MS.

Interpretability of the models

As recurrently noted throughout the text, one of the advantages of using ML methods is that these methods are designed to learn patterns from the training data to make predictions. Unlike traditional statistical methods that often require assumptions about the underlying data distribution and rely on explicit statistical tests, ML models operate in a data-driven manner, and they can capture complex patterns and relationships.

In this study, I explored the variability in the ranking of genomic features assigned by different ML models for MS, a disease identified as having high polygenicity in previous analyses. The results clearly distinguished ensemble tree-based methods from the rest. The similarity in the rankings of features in tree-based methods could be explained by the use of the

same core algorithm, which are decision trees. Tree-based methods were also characterized by showing low variability in the ranking of variants across folds. On the other hand, there was a clear distinction between the two DL methods, FFN and CNN, with the former showing more stability and the latter showing high variability in feature ranks across folds. LR showed high variability in ranks across folds as well, with some folds clustering with CNN folds and showing negative correlation with the rest of the methods. For LR, the presence of multicollinearity might contribute to the high variability in the feature coefficients[193]. However, among the MS classifiers, LR generally demonstrated the lowest variability in performance across folds. Considering these findings, it appears that multicollinearity in LR may adversely affect the stability of feature coefficients without significantly impacting the performance of models.

The diversity in the ranking of features across methods and folds underscores the complexity of polygenic diseases, were genomic signals associated with the condition have multiple interpretations. Additionally, as previously noted, the presence of LD likely contributes to the instability of the importance assigned to the genomic predictors[194]. In this context, extracting general rules, such as a unique prioritization of features that applies to all methods, becomes challenging.

The next question after ranking genomic features was which individual variants were positioned in the top ranks, and therefore, were more informative to classify individuals with MS and healthy controls. Generally, the top genomic variants were located near genes involved in the immune response or associated with MS. In addition, most of these variants were annotated in GTEx as eQTLs or sQTLs to these genes in at least one tissue. However, QTL approaches can be influenced by LD as well, as any non-causal variant in high LD with a truly causal variant will likely show a statistical association with similar effects. Therefore, eQTL and sQTL annotation do not automatically imply that the variant is responsible for the differences in expression or splicing[23]. Also, QTLs are dependent on tissue, cell type, and cell state, and not all the tissues or cell types are equally represented in GTEx. The lack of equal representation hinders the discovery of QTLs across specific cellular contexts[23]. Consequently, QTL annotations should be considered as a guide rather than the ground truth.

Overall, in MS there were 16 different missense variants, also named as nonsynonymous variants, among the 362 genomic predictors, from which, only 3 were prioritized by at least one method. The remaining SNVs were synonymous variants, with no direct impact on the amino acid sequence of any protein. As explained in the Results section 3.6, this aligns with a polygenic and complex disease like MS, where the sum of many small effects across the genome is associated with the disease. In fact, there is currently no evidence for rare, high-impact disease variants in MS[161].

There was an enrichment of HLA gene annotations among the prioritized genomic variants in chromosome 6. Among the prioritized HLA genes, *HLA-A*, *HLA-B*, *HLA-C*, *HLA-E* and *HLA-F* belong to the MHC class I. MHC class I molecules are found on the cell surface of almost all nucleated cells in humans and are responsible of presenting peptide fragments of intracellular proteins to cytotoxic T cells (CD8+). This presentation triggers an immediate immune response when a specific non-self antigen is detected. MHC class I molecules play a central role in immune surveillance by presenting intracellular peptides to cytotoxic T cells, allowing the immune system to detect and eliminate infected or abnormal cells.

In contrast, the prioritized HLA genes *HLA-DRB1* and *HLA-DRB5* belong to the MHC class II, which is typically found only on professional antigen-presenting cells such as dendritic cells, mononuclear phagocytes, some endothelial cells, thymic epithelial cells, and B cells. The extracellular proteins are phagocytosed by these professional antigen-presenting cells, digested in lysosomes, and the resulting peptide fragments are loaded onto MHC class II molecules, which are then presented on the cell surface. MHC class II molecules primarily interact with immune cells, such as T helper cells (CD4+), and the presented peptide regulates how T cells respond to an infection. Therefore, MHC class II molecules are central to the immune system's ability to recognize and respond to antigens derived from extracellular pathogens, and they are critical for the activation of helper T cells, which coordinate and regulate various aspects of the adaptive immune response.

The region around *HLA-DRB1* consistently appears in the scientific literature as being the most strongly associated with MS, also involving other MHC class II genes. Specifically, the HLA variant *HLA-DRB1*15:01* is the strongest genetic determinant of MS, as defined in the literature[169] [166] [97] [199]. However, in the current study, the most consistently prioritized genomic variant in MS was *HLA-A*02:01*, which belongs to the MHC class I. Notably, there is evidence of the independent association of *HLA-A*02:01* and *HLA-DRB1*15:01* with MS[80] [167] [168], with the former considered protective and the latter predisposing to the disease. In this regard, the results in this study, showing prioritized variants affecting HLA genes from both MHC class I and class II, are consistent with findings from other studies that indicate alterations in both types of cell surface proteins in MS, involving different types of immune response and immune cells[200]. However, the strong LD present among HLA variants, such as the one observed between *HLA-DRB1*15:01* and *HLA-DRB5*Null*, complicates the use of fine mapping methods to identify the exact causal variants in MS. In conclusion, caution should be exercised when interpreting the list of prioritized genomic variants reported in this study, and results should be validated with other sources.

In the case of MS, the well-known region strongly associated with the disease in the vicinity of *HLA-DRB1*15:01* on chromosome 6 did not seem to overshadow the relevance of other variants located on different chromosomes. Consistent with this observation is the fact that, when exploring the genomic variants with the top 10 sum of ranks, only *HLA-A*02:01* belonged to chromosome 6, and it was not in strong LD with *HLA-DRB1*15:01*. Genomic variants near genes such as *MAPK1* in chromosome 22, *CYP24A1* in chromosome 20, *RPS6KB1* in chromosome 17, *CLEC16A* in chromosome 16, *TNFRSF1A* in chromosome 12 and *MANBA* in chromosome 4 were among the top 10 prioritized genomic variants, highlighting the distribution of prioritized genomic features across different chromosomes.

Given the polygenicity observed in MS, it is natural to hypothesize that certain genomic variants were prioritized due to the presence of epistasis among them. Despite the challenges associated with the complexity of chromosome 6, the most polymorphic region of the human genome, and

the presence of strong LD between HLA variants, other studies have reported some epistasis, typically involving the *HLA-DRB1*15:01* haplotype and other HLA variants[91 201].

In this work, I calculated the number of pairwise interactions among the prioritized features by each method in MS. Unfortunately, the results in this section were contradictory. ET was the method reporting the highest number of significant pairwise interactions across all ML methods with 12, followed by LR with 9. However, LR is not specifically designed to account for interactions between predictors. In this respect, the interactions obtained in LR were likely formed by genomic variants that had a strong individual association with the disease regardless of the presence of interactions[65]. In addition, in the case of RF, a method that has been used in other works to detect epistasis[70 202], interactions were missing among their prioritized variants.

Patterns captured by ML methods often lack associated biological mechanisms that could be related with the phenotype of interest[36]. For that reason, I added molecular annotations to the interactions to try to provide a possible molecular context in which these interactions may intervene. As a result, all the interactions with molecular annotations were located on chromosome 6, and half of the gene annotations specifically involved the HLA family and complement factor genes. This fact is consistent with MS being an autoimmune disease with an important role of HLA variants. Nevertheless, the reliability of the reported interactions with molecular annotations is difficult to assess due to the previously described reasons, and because the genomic variants involved in the interactions were eQTLs and sQTLs to many genes, complicating the interpretation of the molecular connections.

Limitations and ethical considerations of the study

It is important to clarify that modifying any of the parameters in this study, such as employing different ML methods, selecting different genomic features, or applying ML methods to other diseases, could potentially alter some of the conclusions drawn here. Indeed, one of the weaknesses of the self-learning methods studied in this work is their variability when certain conditions in the analysis are changed. Despite this limitation, ML

methods have proven to be powerful and efficient in various everyday applications, and properly applied with sufficient data, the advantages of these methods are undeniable. Therefore, the primary value of this work lies in highlighting general considerations for choosing and applying ML models in disease risk classification analyses using genomic data.

Several limitations have been encountered during the pursuit of the objectives of this work:

- The primary limitation is the restricted sample size. In the duration of this project, I encountered challenges in obtaining access to datasets with individualized genomic data. This is because the process of accessing such datasets is complex and involves several legal and privacy verification steps, requiring collaboration with the legal teams from both hosting and requesting institutions. Repositories like dbGAP[51] and EGA[203] host individualized human genomic and phenotypic data from various studies. In order to have access to the datasets, I had to comply with the terms outlined in the data transfer agreements. However, EGA provides limited guidance to scientists on the data request process; instead, they provide an email address with the contact information of the data owners for direct inquiries. Requests are initiated by sending an email expressing interest in accessing a specific dataset on EGA. Ideally, this email should initiate the communication between data owners and requesters, eventually leading to the exchange of the signed data access agreement. However, I found that direct communication with data owners was not always successful. Standardizing the legal clauses for data access and monitoring the responsiveness of researchers who own the data could potentially increase the sample sizes used in other studies. As a result, this could enhance the research in the application of ML models that require large amounts of genomic data to reach optimum performance.

- Gaining access to datasets does not guarantee the utility of the data, as datasets can be submitted to repositories without quality checks[204], highlighting an unresolved issue in the scientific

community. In this context, the low quality of data may result in poorer performance of ML models[205]. Another limitation is that the integration of diverse sequencing and genotyping technologies requires the inclusion of a step to homogenize and normalize the data, potentially creating batch effects that could introduce biases in the results. In this regard, the FAIR principles, defined as Findability, Accessibility, Interoperability, and Reusability, have emerged in recent years as a framework to ensure that scientific data is not only discoverable but also easily accessible, enabling interoperability across diverse platforms and systems[206].

- Rare and low-frequency variants (*MAF* < 0.05) are not included as features in the models due to several reasons. One reason for this exclusion is associated with the genotyping technology. Removing rare variants is a standard quality control step when working with genotyping arrays, as with this technology rare variants are susceptible to have an elevated false positive rate[207]. The other reason is that the insufficient genetic variation in rare variants can pose challenges in associating these specific locus with observed differences in traits or outcomes, because there is not enough diversity in the training set to extract any generalizable patterns[208].

Several ethical considerations are associated with studies using genomic data for similar applications as the one in this work:

- Privacy and informed consent are essential, as genomic data is highly sensitive and unique to individuals. Concerns exist about the potential use of this data for re-identification, leading to privacy breaches. Ensuring informed consent, explaining the risks, and providing clear information about how genomic data will be used are important ethical considerations. In this regard, compliance with relevant data protection and privacy regulations, such as the general data protection regulation (GDPR) for the European Union members, and the health insurance portability and accountability act (HIPAA) in the United States, is crucial for the ethical use of genomic data[209]. Following these directives, at the Istituto Italiano di Tecnologia (IIT), where the current study was conducted, it is

compulsory for all scientists working on sensitive data to take a course on GDPR.

- Safeguarding genomic data from unauthorized access, breaches, or misuse is essential. In relation to my work, the informatics team in the IIT implemented robust security measures to ensure the data protection of the genomic datasets used in this study. These measures are critical to maintaining trust in the use of ML methods[210].

- Determining who owns genomic data and who has control over its use is a complex ethical issue[211]. Clear policies should address issues of data ownership, access, and control, and individuals should be informed about how their data will be used. In addition, there should be a consent for the secondary use of the data, as individuals may have provided their genomic data for a specific purpose, but ML methods may involve secondary use of the data for different purposes. Regarding the datasets used in the current study, all the specifications concerning the use and ownership of the data were listed in the data transfer agreement required to be signed by both, the hosting and requesting institutions.

- Biases in genomic data can result in unfair and discriminatory outcomes of ML models[71]. Ensuring inclusivity in the training data and addressing biases in algorithms are essential to avoid perpetuating health disparities. In this respect, there is a general overrepresentation of studies with individuals of white European ancestry compared to other populations. If this is not considered when designing ML experiments, the predictive models may lead to inaccurate discoveries across underrepresented populations[212]. For example, the results in this study are likely to be more effectively extrapolated to white individuals from the UK, as in the UKB, other ethnicities were underrepresented.

- ML models can be complex and challenging to interpret. The lack of transparency may raise ethical concerns, especially when making decisions that impact individuals' health[213]. For that

reason, in this study I also focused on the interpretability of the models, choosing methods that, rather than being a black box, allow the application of XAI methods.

- The use of genomic data for disease classification may lead to stigmatization or discrimination based on genetic information. It is essential to consider the potential social and psychological impact of the predictive models on vulnerable individuals and communities[209]. In this work, I built models attempting to classify individuals with complex diseases from healthy controls. Given that individuals were anonymized, the results of this work have value in the context of research. However, the application of these models could also help in the diagnosis of complex diseases in the medical practice. In this context, the misuse of these tools could enhance the discrimination against individuals based on their genomic predisposition to develop a serious disease. Also, it may have a negative psychological impact on healthy individuals who receive information about their high probabilities of developing a disease with no cure.

- Some of the categories used to stratify the population in this study are discriminatory to minorities or for reasons of race. For example, the binary classification of sex is currently under discussion in the scientific community, as it can be discriminatory for transgender and non-binary individuals. Consequently, there is a need for a more inclusive classification of individuals who do not conform to traditional gender norms in scientific studies[214]. Additionally, the use of the term “Caucasian” to describe individuals with the white race is currently being questioned[215]. This term was originally invented by anthropologists who categorized humans into racial groups and created theories about white superiority in the 18th century. Despite of its origins, this term is still widely used to categorize white individuals, even in important databases such as UKB.

Future perspectives

Adherence to ethical guidelines and the maintenance of transparency are essential for the responsible use of ML methods with genomic data as disease classifiers. For future perspectives on the application of ML methods with genomic data, it is necessary to consider the previously described limitations and ethical concerns.

Some of the challenges associated with the use of self-learning methods on genomic data, such as sample size and privacy concerns, could potentially be addressed in the future by implementing federated learning (FL) techniques[72]. Even though FL is a relatively new tool, and research on the requirements to effectively run these processes is still in its early-stages, future advancements in this field will likely result in promising achievements.

In conclusion, with the reduction in genome sequencing costs and improvements in sequencing technologies, the volume of genomic data is expected to continue increasing in the coming years. Simultaneously, the rise in computational capacities and advancements in existing ML methods are likely to foster exciting discoveries in the field of population genomics. However, these promising developments should be accompanied by a careful consideration of the ethical concerns mentioned earlier to ensure the responsible and ethical use of these methods for the collective benefit.

5. Methods

5.1 Inclusion and exclusion criteria

For this work, I used information from individuals in the UK Biobank (UKB). The UK Biobank Axiom Array[11], a custom-designed array manufactured by Affymetrix, was the source of genomic data. This array contains nearly 820,000 genetic markers, including SNVs, and small insertion-deletion polymorphisms (indels).

The inclusion criteria used to select cases and controls in UKB were as follows:

- MS: Subjects with the ICD-10 code G35 in primary care data, hospital inpatient data, or mortality data.
- AD: Subjects with the ICD-10 code G30.9 in primary care data, hospital inpatient data, or mortality data. 78 subjects (3% of the total AD) had less than 65 years old and therefore, were probable EOAD. I decided to include the probable EOAD in the analysis under the premise that the models may be able to correctly classify them as AD, as both EOAD and LOAD share some genetic determinants. In any case, the age at first report was explored to evaluate potential biases in age among the true positives and false negatives as classified by the models.
- SC: Subjects with the ICD-10 code F20.9 in primary care data, hospital inpatient data, or mortality data.
- PD: Subjects with the ICD-10 code G20 in primary care data, hospital inpatient data, or mortality data.
- Controls: Subjects who are more than 75 years old and do not have any disease of the nervous system or mental, behavioral, and neurodevelopmental disorders (ICD-10 categories G00-G99 and F01-F99).

The exclusion criteria applied to subjects in UKB were as follows:

- Subjects without any clinical information.
- Subjects with more than one of the studied diseases.
- Genetic ethnic grouping not Caucasian (UKB Field ID 22006).

- Recommended genomic analysis exclusions due to poor heterozygosity/missingness (UKB Field ID 22010).
- Individuals with high heterozygosity rate (after correcting for ancestry) or high missing rates (UKB field ID 22018).
- Individuals with sex chromosome aneuploidy (UKB field ID 22019).
- Outliers for heterozygosity or missing rate (UKB field ID 22027).
- From the genetically related individuals, only one subject (preferentially with the disease) was included in the analysis (UKB field ID 22011).

	Cases		Controls	
	Female	**Male**	**Female**	**Male**
MS	1443	577	42154	37695
AD	1357	1133	42154	37691
PD	1161	1965	42146	37692
SC	411	577	42157	37698

Table 15 shows the distribution of individuals from UKB employed in this study, across diseases and sexes, after applying the selection criteria. Diseases MS and PD are highlighted in green, indicating the most pronounced sex imbalance.

The ADNI dataset (adni.loni.usc.edu)[216] was used as an external validation dataset for AD. In this study I used the genomic data coming from whole-genome sequence (WGS) at high coverage. The inclusion criteria used to select cases and controls in ADNI were as follows:

- AD: Individuals with probable or possible diagnosis of AD (field name: DXAPP) and dementia due to AD (field name: DXDDUE) or mild cognitive impairement (MCI) due to AD (field name: DXMDUE).
- Controls: Individuals without MCI or dementia (field name: DIAGNOSIS), without probable or possible diagnosis of AD (field

name: DXAPP), without dementia due to AD (field name: DXDDUE) and without MCI due to AD (field name: DXMDUE).

The exclusion criteria applied to subjects in ADNI were as follows:

- Individuals without WGS data available.
- Individuals with missing data in the demographic variables sex and year of birth (field names: PTGENDER and PTDOBYY).
- Individuals with missing values in the examination date (field name: EXAMDATE).
- Individuals with less than 75 years old.

	Cases		Controls	
	Female	**Male**	**Female**	**Male**
ADNI	17	39	17	20

Table 16 shows the distribution of individuals in the ADNI dataset after applying the selection criteria.

I used the genotyping data from Affymetrix GeneChip® Human Mapping 500K arrays generated by the International Multiple Sclerosis Genetics Consortium (IMSGC) and available in dbGAP under the accession ID "phs000139.v1.p1" as external validation dataset for MS. The dataset consisted in two cohorts: named as "multiple sclerosis" (IMSGC MS), which included trio families recruited from across the UK, and "multiple sclerosis and related disorders" (IMSGC MSRD), comprising trio families recruited from across the US. Approximately 4% of the subjects in the latter cohort were diagnosed with clinically isolated syndrome (CIS) at the time of enrolment into the study. Additional information regarding the selection of participants can be found in the supplementary appendix of the original paper[154]. The inclusion criteria I used to select cases in the IMSGC dataset were as follows:

- Subjects with MS (variable name: AFFECTION_STATUS).
- Only one individual per family was included.

The exclusion criteria applied to subjects in the IMSGC dataset were as follows:

- The data consisted in family trios. Consequently, controls were excluded from the analysis as they had at least one relative with MS.
- Subjects with more than 20% of missing genotypes were also excluded.

	Cases	
	Female	**Male**
IMSGC MS	363	122
IMSGC MSRD	357	108

Table 17 shows the distribution of individuals in the IMSGC dataset after applying the selection criteria. In the case of IMSGC, two cohorts were available, IMSGC MS corresponding to individuals from the UK, and IMSGC MSRD corresponding to individuals from the US.

5.2 Pre-processing of genomic data

ML methods were employed to classify cases and controls using a set of genomic variants as features in the models. These genomic variants, were reported in ClinVar[52] with at least one level of review status or reported in DisGeNet[53] within the curated dataset. A binary feature indicating sex was also included in the models. When an HLA gene was associated with the disease, the imputed HLA variants for this gene obtained from UKB (UKB Field ID 22182) were included as predictors. The numbers of predictors used in each disease are shown in Table 18.

Genetic variants were encoded as 0, 1, 2 and 3 corresponding to missing value, the absence of the variant, the presence of the variant in one copy, and the presence of the variant in two copies, respectively, assuming an additive model. Genomic variants with the same values in all samples (monomorphic predictors) were excluded from the analysis.

PLINK[217] was used to apply an initial quality control and to pre-process the genomic raw data. SNVs with a Hardy-Weinberg equilibrium *p-value*

("HWE" in PLINK) lower than 1e-8, minor allele frequency ("MAF" in PLINK) lower than 0.05, missingness per marker ("geno" in PLINK) higher than 0.2, and samples with missingness per individual ("mind" in PLINK) higher than 0.2, were excluded. Additionally, PLINK was employed to compute the linkage disequilibrium (LD) statistics between genomic variants.

	MS	AD	SC	PD
Features by type	309 SNVs 53 HLA 1 sex	167 SNVs 2 HLA 1 sex	136 SNVs 32 HLA 1 sex	64 SNVs 2 HLA 1 sex
Total features	363	170	169	67

Table 18 indicates the number of features of each type used in the models for each disease.

In genotyping arrays and WGS, missing values are not randomly distributed, and specific tools can be used for the imputation[218] [219]. The pipeline for imputing missing values involved several steps, and only the genomic variants that were already present in the array but had missing genotypes in some samples (less than 20% of samples after QC filters) were imputed. The pre-processing of genomic files was performed using PLINK and bcftools[220], which included tasks such as strand flipping, genome build, and ID conversion. Haplotype phasing was performed using SHAPEIT4[221], while IMPUTE5[26] was employed for genomic imputation. The reference files for genomic imputation were obtained from the 1000 genomes phase3[222]. Imputed genotypes with less than 80% probability were considered as missing, and imputed genomic variants with a quality score lower than 0.90 were excluded from further analysis. In an attempt to impute the HLA genes, HIBAG[30] was applied to the dataset sourced from dbGAP phs000139.v1.p1 (GeneChip® Human Mapping 500K arrays). However, the quality of HLA imputation in this dataset did not meet the desired standards, and consequently, the imputed HLA types were not included in the analysis for the dbGAP dataset.

5.3 Machine learning models

Nested cross-validation (nested CV) was applied with 10 folds in the inner loop and 5 folds in the outer loop to select the optimum hyperparameter configuration and obtain an estimate of the model's generalization performance. For the hyperparameter selection, the grid search approach was employed, and the 10 evaluation scores obtained for each hyperparameter configuration in the inner loop were used to select the optimum hyperparameter configuration. The hyperparameter configurations were ranked in decreasing order using the mean of balanced accuracy across the 10 inner folds. From the top 10 hyperparameter configurations with higher values of balanced accuracy mean, the hyperparameter configuration with the highest value of sensitivity minus the standard deviation of sensitivity across the 10 folds was selected. For each fold in the outer loop, the selected hyperparameter configuration in the inner loop was applied in the outer loop using 80% of balanced samples for training and 20% of samples for testing. The strategy of nested CV used in this study is represented in Figure 3 in the Introduction section 1.3. The ML methods used, along with the corresponding hyperparameters considered in the grid search, are listed in Table 19.

The architecture of FFN with the list of fixed and tuned parameters used in this study is represented in Figure 7 in the Introduction section 1.3. The architecture of the CNN used in this study is represented in Figure 8 in the Introduction section 1.3. The convolutional block in the CNN employed three matrices as input channels. Matrix A represented the presence of genomic variants, like the matrices used in the other ML methods. Matrix B and matrix C represented the chromosome and position of the genomic variants, respectively. The sex variable was encoded in matrix A, with a value of 1 for females and 2 for males, while it was 0 in matrix B, and the lowest genomic position minus one in matrix C. The values in the three matrices were converted to the range of -1 to 1. Genomic variants were ordered by chromosome and position to represent their location over the entire genome.

	Python library	Hyperparameters
Gradient-Boosted Decision Trees (GB)	scikit-learn, GradientBoostingClassifier	• n_estimators (70, 80, 90, 100) • learning_rate (0.0001, 0.001, 0.01, 0.1, 1.0) • subsample (0.5, 0.7, 1.0) • max_depth (7, 9, 10, 12, 14) • loss ('log_loss', 'exponential') • balance (50, 60, 70) • sampling ('ENN', 'random', 'SMOTE_ENN', 'SMOTE_random')
Extremely Randomized Trees (ET)	scikit-learn, ExtraTreesClassifier	• n_estimators (50, 60, 70, 80, 100) • min_samples_split (2, 5, 8) • min_samples_leaf (1, 2, 5) • max_depth (None) • balance (50, 60, 70) • sampling ('ENN', 'random', 'SMOTE_ENN', 'SMOTE_random')
Random Forest (RF)	scikit-learn, RandomForestClassifier	
Logistic Regression (LR)	scikit-learn, LogisticRegression	• solver ('newton-cg', 'liblinear', 'sag', 'saga') • creg (0.00001, 0.0001, 0.001, 0.01, 1, 10, 100) • balance (50, 60, 70) • sampling ('ENN', 'random', 'SMOTE_ENN', 'SMOTE_random')
Feedforward networks (FFN)	PyTorch	• number of epochs (300, 400, 500) • learning rate (0.0001, 0.001, 0.01) • drop-out (0.1, 0.2, 0.4) • number of units nUnits (100, 200) • number of layers nLayers (1, 2, 3) • balance (50, 60, 70) • sampling ('ENN', 'SMOTE_ENN')
Convolutional Neural Networks (CNN)	PyTorch	

Table 19 showing the ML methods used in this study, along with the corresponding python libraries and functions used to build the models, as well as the tested hyperparamaters.

In addition to the hyperparameters related to the configuration of the ML methods above listed, other parameters associated with balancing and sampling strategies were tested. Varying degrees of class imbalance were used during training, where the number of cases remained constant, while the number of controls varied according to the following proportions:

- 50% cases and 50% of controls
- 40% cases and 60% controls
- 30% cases and 70% controls

As for the sampling strategies, four different approaches were tested:

- Random undersampling (random)
- Edited nearest neighbour undersampling (ENN)
- SMOTE oversampling 20% of cases + random undersampling (SMOTE_random)
- SMOTE oversampling 20% of cases + edited nearest neighbour undersampling (SMOTE_ENN)

The final hyperparameter configurations selected for each fold, method and disease are listed in Table 20 for MS and AD, and in Table 21 for SC and PD. The class imbalances 40%/60% and 30%/70% did not appear to confer any advantage to the model performance, as all the final hyperparameter configurations exhibited a class imbalance of 50%/50%. As for SMOTE oversampling, it was only selected in some of the hyperparameter configurations of DL methods.

	MS	AD
GB	**Fold1**: 100|0.01|0.5|12|exponential|50|random **Fold2**: 70|0.001|0.5|12|deviance|50|ENN **Fold3**: 70|0.0001|0.5|7|deviance|50|random **Fold4**: 100|0.1|1|7|exponential|50|random **Fold5**: 80|0.1|0.7|10|exponential|50|random	**Fold1**: 90|0.01|1|12|exponential|50|ENN **Fold2**: 100|1|0.7|14|exponential|50|random **Fold3**: 90|1|1|7|exponential|50|random **Fold4**: 80|1|1|12|exponential|50|ENN **Fold5**: 100|1|1|7|exponential|50|ENN
ET	**Fold1**: 50|5|2|None|50|ENN **Fold2**: 50|5|1|None|50|ENN **Fold3**: 60|5|2|None|50|ENN **Fold4**: 50|8|1|None|50|ENN **Fold5**: 50|2|2|None|50|random	**Fold1**: 60|5|1|None|50|ENN **Fold2**: 50|8|1|None|50|random **Fold3**: 50|5|1|None|50|ENN **Fold4**: 60|5|1|None|50|ENN **Fold5**: 70|5|1|None|50|random
RF	**Fold1**: 50|2|2|None|50|random **Fold2**: 50|8|2|None|50|ENN **Fold3**: 50|5|1|None|50|random **Fold4**: 50|2|2|None|50|ENN **Fold5**: 50|5|2|None|50|random	**Fold1**: 80|5|1|None|50|ENN **Fold2**: 50|2|2|None|50|ENN **Fold3**: 80|5|1|None|50|ENN **Fold4**: 60|5|2|None|50|ENN **Fold5**: 50|8|1|None|50|ENN
LR	**Fold1**: liblinear|1|50|random **Fold2**: newton-cg|0.01|50|ENN **Fold3**: saga|0.01|50|random **Fold4**: saga|1|50|random **Fold5**: newton-cg|0.01|50|ENN	**Fold1**: newton-cg|100|50|ENN **Fold2**: newton-cg|10|50|ENN **Fold3**: saga|100|50|ENN **Fold4**: sag|100|50|ENN **Fold5**: newton-cg|1|50|ENN
FFN	**Fold1**: 300|0.0001|0.1|100|1|ENN|50 **Fold2**: 400|0.0001|0.4|200|1|ENN|50 **Fold3**: 300|0.0001|0.2|100|1|ENN|50 **Fold4**: 500|0.01|0.2|100|1|ENN|50 **Fold5**: 300|0.0001|0.1|100|2|ENN|50	**Fold1**: 500|0.001|0.4|100|3|ENN|50 **Fold2**: 500|0.01|0.1|100|3|ENN|50 **Fold3**: 500|0.01|0.1|100|2|ENN|50 **Fold4**: 500|0.01|0.1|200|1|ENN|50 **Fold5**: 400|0.001|0.4|100|1|ENN|50
CNN	**Fold1**: 500|0.0001|0.1|100|2|SMOTE_ENN|50 **Fold2**: 500|0.0001|0.1|200|3|SMOTE_ENN|50 **Fold3**: 400|0.0001|0.1|200|1|ENN|50 **Fold4**: 500|0.0001|0.1|100|1|ENN|50 **Fold5**: 500|0.001|0.4|200|1|SMOTE_ENN|50	**Fold1**: 500|0.001|0.4|100|3|ENN|50 **Fold2**: 500|0.01|0.2|200|1|SMOTE_ENN|50 **Fold3**: 500|0.001|0.4|200|1|ENN|50 **Fold4**: 300|0.01|0.2|100|3|SMOTE_ENN|50 **Fold5**: 400|0.0001|0.1|100|3|SMOTE_ENN|50

Table 20 lists the hyperparameters selected for the final models in MS and AD. The five folds correspond to the outer loop of the nested CV. The parameters listed for GB, in order, are: n_estimators, learning_rate, subsample, max_depth, loss, balancing, and sampling. The parameters listed for ET and RF, in order, are: n_estimators, min_samples_split, min_samples_leaf, max_depth, balancing, and sampling. The parameters listed for LR, in order, are: solver, C, balancing, and sampling. The parameters listed for FFN and CNN, in order, are: number of epochs, learning rate, dropout probability, number of units, number of layers, sampling, and balancing.

	SC	PD
GB	**Fold1**: 70|0.01|0.5|14|exponential|50|ENN **Fold2**: 100|0.0001|0.5|12|deviance|50|ENN **Fold3**: 70|0.01|0.7|10|exponential|50|ENN **Fold4**: 100|0.01|0.7|7|exponential|50|ENN **Fold5**: 80|0.0001|0.7|7|exponential|50|ENN	**Fold1**: 100|0.01|0.5|12|exponential|50|ENN **Fold2**: 100|0.0001|0.7|14|deviance|50|ENN **Fold3**: 100|0.001|1.0|7|deviance|50|random **Fold4**: 100|0.0001|0.7|12|exponential|50|random **Fold5**: 70|0.01|1.0|7|deviance|50|random
ET	**Fold1**: 100|8|5|None|50|ENN **Fold2**: 60|8|2|None|50|ENN **Fold3**: 100|8|2|None|50|ENN **Fold4**: 70|5|1|None|50|ENN **Fold5**: 100|8|2|None|50|ENN	**Fold1**: 60|2|2|None|50|ENN **Fold2**: 80|5|5|None|50|ENN **Fold3**: 100|2|1|None|50|random **Fold4**: 100|5|5|None|50|random **Fold5**: 60|8|5|None|50|random
RF	**Fold1**: 100|8|2|None|50|ENN **Fold2**: 100|5|1|None|50|ENN **Fold3**: 70|5|2|None|50|ENN **Fold4**: 100|2|5|None|50|ENN **Fold5**: 100|5|1|None|50|random	**Fold1**: 100|5|2|None|50|ENN **Fold2**: 100|8|5|None|50|random **Fold3**: 100|2|5|None|50|random **Fold4**: 100|5|5|None|50|random **Fold5**: 80|2|5|None|50|random
LR	**Fold1**: newton-cg|0.01|50|ENN **Fold2**: newton-cg|1|50|ENN **Fold3**: sag|1|50|random **Fold4**: liblinear|1|50|ENN **Fold5**: sag|1|50|ENN	**Fold1**: newton-cg|1|50|random **Fold2**: liblinear|0.01|50|random **Fold3**: liblinear|1|50|random **Fold4**: newton-cg|10|50|random **Fold5**: newton-cg|1|50|random
FFN	**Fold1**: 300|0.0001|0.2|200|1|ENN|50 **Fold2**: 500|0.0001|0.2|100|2|ENN|50 **Fold3**: 300|0.0001|0.1|100|3|SMOTE_ENN|50 **Fold4**: 300|0.0001|0.1|100|1|ENN|50 **Fold5**: 300|0.0001|0.2|100|1|SMOTE_ENN|50	**Fold1**: 300|0.0001|0.2|200|3|ENN|50 **Fold2**: 400|0.0001|0.4|100|1|ENN|50 **Fold3**: 300|0.0001|0.2|100|1|SMOTE_ENN|50 **Fold4**: 400|0.0001|0.4|100|1|ENN|50 **Fold5**: 400|0.0001|0.4|200|1|ENN|50
CNN	**Fold1**: 300|0.001|0.2|100|1|ENN|50 **Fold2**: 400|0.001|0.2|200|3|ENN|50 **Fold3**: 300|0.001|0.4|200|2|ENN|50 **Fold4**: 500|0.001|0.2|100|2|ENN|50 **Fold5**: 300|0.0001|0.2|200|1|SMOTE_ENN|50	**Fold1**: 400|0.0001|0.2|100|1|ENN|50 **Fold2**: 400|0.0001|0.2|100|2|ENN|50 **Fold3**: 300|0.0001|0.1|200|1|ENN|50 **Fold4**: 400|0.0001|0.1|200|1|SMOTE_ENN|50 **Fold5**: 400|0.0001|0.1|200|1|SMOTE_ENN|50

Table 21 lists the hyperparameters selected for the final models in SC and PD, following the same structure as in Table 20.

The number of samples for each disease across the testing, validation, and training sets are as follows:

- MS: Testing (404 cases, 15970 controls); Validation (162 cases, 6388 controls); Training (1616 cases).
- AD: Testing (498 cases, 15969 controls); Validation (200 cases, 6387 controls); Training (1992 cases).
- SC: Testing (198 cases, 15971 controls); Validation (79 cases, 6389 controls); Training (790 cases).
- PD: Testing (625 cases, 15968 controls); Validation (251 cases, 6387 controls); Training (2250 cases).

The number of controls in the training sets varied and depended on the imbalance rate. The final evaluation performance for each method was obtained from the outer loop in the nested CV. This was done by calculating the mean and standard deviation across the five different folds.

Feature selection methods were used to identify a subset of predictors that could potentially enhance the performance of the models. Recursive feature elimination (RFE) was implemented using the sklearn.feature_selection.RFE function in Python. Different number of features were tested using the sklearn.model_selection.GridSearchCV function, with 20, 50, 100, 150, 200, and 250 for MS, and 5, 20, 50, 100 and 150 for AD. Additionally, recursive feature elimination with cross-validation (RFECV) was implemented using the sklearn.feature_selection.RFECV function. In RFE and RFECV, balanced accuracy was used as the scoring function.

5.4 Polygenic risk score

PRSice-2 was used to calculate the polygenic risk score (PRS)[45] for the disease of interest. PLINK files from UKB were used as target data. The summary statistics used as the base data were downloaded from the NHGRI-EBI GWAS Catalog[50] on 25/05/2023 for the studies GCST005531[89] related to MS and GCST007511[111] to AD. Summary statistics for SC were obtained from dbGaP at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000021.v3.p2[223]. Summary statistics for PD were obtained from the International Parkinson Disease Genomics Consortium (IPDGC) resources page (https://pdgenetics.org/resources)[146].

PRS was calculated using the average effect size function and considering an additive model for regression. PRS calculation was combined with *p-value* thresholding using the C+T (*clumping* + *thresholding*) method[43]. Following this approach, PRS was calculated several times comprising SNVs with increasing GWAS *p-value* thresholds, and the most predictive PRS was used for the final PRS calculation.

Genomic variants in the base data were filtered to exclude multi-allelic SNVs. Discrepancies caused by inverted effect alleles were resolved, and each rsID was linked to a single nucleotide change. The sex variable and the first 10 principal components (PC) available for researchers to download from UKB (UKB field ID 22009) were added as covariates in the PRS models.

PRS were calculated five times for each disease, including in the regression model the same samples used in the outer loop of the nested CV used for training the final ML models. The aim was to compare the performance of PRS with ML using the same samples for fitting and evaluation in each fold. I ran the experiment twice: first, using all the genomic variants present in the target and base data (PRS_ALL; 6007, 234084, 48197, 265035 variants after clumping for MS, AD, SC, and PD, respectively), and second, the disease related variants used in the ML models (PRS_RED; 78, 72, 42, 48 variants after clumping for MS, AD, SC, and PD, respectively). To convert PRS into binary categories, a threshold was established to distinguish the individuals with high risk to the disease. Individuals with a PRS above the 99th percentile were classified as high risk (positives)[44]. Similarly, the 99th percentile was applied to the probabilities obtained from ML models to classify high-risk individuals and to compare the results with the PRS models. The relative risk (RR) and odds ratio (OR) were used to evaluate the models, with the formulas provided below:

$$RR = \frac{P^{99th}/(P^{99th} + N^{99th})}{P'/(P' + N')}$$

$$OR = \frac{P^{99th}/N^{99th}}{P'/N'}$$

Where P^{99th} and N^{99th} represent the number of positives (individuals with the disease) and negatives (controls) present in the top 99th percentile with the highest PRS, or probabilities in the case of ML methods. P' and N' represent the number of positives and negatives present in the samples that were not in in the top 99th percentile.

5.5 Explainability methods applied to machine learning models

The importance measures assigned to the features in the classification were obtained through various approaches depending on the ML method. For the tree-based ensemble ML methods such as GB, ET and RF, feature

importance metrics were obtained from model statistics. In the case of LR, the coefficients of the features in the decision function were used. In DL methods, specifically FFN and CNN, importance metrics were derived using layer integrated gradients (LIG)[224], layer deeplift (DE)[225], saliency maps (SM)[226] and guided backpropagation (GBP)[227]. These methods provide a score for each sample and feature, and the resulting matrices share the same dimensions as the input matrices. To obtain a single importance value for each feature, the median of the absolute values of attributes was calculated for cases and controls, and both values were summed for each feature. In the case of CNN, this process was repeated for each matrix, and the resulting values from the three matrices were summed. To address the fact that the importance measures were obtained using different approaches, and consequently, had a different range of values, the predictors were ranked from the highest importance to the lowest importance using consecutive ordinal numbers for each method and fold.

The prioritization of genomic features was performed in the fold with the highest balanced accuracy for each ML method. The top 10% of the best-ranked features in each method were selected as the prioritized genomic variants indicating a stronger association with the disease. To add information on the predicted pathogenic effect of missense mutations to the protein, AlphaMissense was employed[160]. To incorporate information regarding the impact of SNVs on RNA expression and splicing, data on expression quantitative trait loci (eQTL) and splicing quantitative trait loci (sQTL) from the GTEx database were used.

With the aim of detecting pairwise interactions among the prioritized genomic variants, I used a full generalized lineal model (GLM) considering two genomic variants as independent variables with their individual effect and interaction to classify cases and controls, as well as a reduced GLM with the same variables but considering only the individual effect of each genomic variant without the interaction term to classify both classes. An analysis of deviance between the full GLM and reduced GLM, comparing the reduction in deviance with a chi-squared test was performed. The same test was applied 100 times, randomly selecting 1800 MS and controls in each iteration. The asymptotically exact harmonic mean p-

value (HMP) was used to summarize the *p-values* obtained across the iterations and correct for multiple comparisons. A HMP < 0.01 was employed to select significant pairwise interactions.

Abbreviations	Full description
P-P	The genomic variants with an annotated protein-protein interaction are both missense variants located in the transcripts coding for the proteins involved in the interaction. These are designated as direct interactions.
P-E	One genomic variant involved in the interaction is a missense variant located in the transcript of the protein, and the other affects the gene expression of the other protein involved in the interaction.
P-S	One genomic variant involved in the interaction is a missense variant located in the transcript of the protein, and the other affects the splicing of the transcript coding for the other protein involved in the interaction.
E-E	Both genomic variants with the annotated interaction affect the gene expression of the proteins involved in the interaction.
S-S	Both genomic variants with the annotated interaction affect the splicing of transcripts coding for the proteins involved in the interaction.
E-S	The genomic variants with the annotated interaction affect the spicing in one case and the gene expression in the other case, of the proteins involved in the interaction.

Table 22 lists the different types of interactions with molecular annotations.

The significant pairwise interactions involving the prioritized genomic variants were characterized with molecular annotations. The experimental-validated human protein-protein interactions were sourced from the Integrated Interactions Database[228]. Interactions of variants affecting genes coding for proteins involved in protein-protein interactions were characterized with the corresponding molecular annotations. Interactions of variants involving the same gene coding for the same protein were also characterized with molecular annotations. Table 22 describes the different types of interactions with molecular annotations.

6. References

1. Human Genome Project Fact Sheet. Accessed December 1, 2023. https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project
2. Akintunde O, Tucker T, Carabetta VJ. The evolution of next-generation sequencing technologies. *ArXiv*. Published online May 15, 2023. Accessed December 1, 2023. /pmc/articles/PMC10246072/
3. Pervez MT, Hasnain MJU, Abbas SH, Moustafa MF, Aslam N, Shah SSM. A Comprehensive Review of Performance of Next-Generation Sequencing Platforms. *Biomed Res Int*. 2022;2022. doi:10.1155/2022/3457806
4. Lamy P, Grove J, Wiuf C. A review of software for microarray genotyping. *Hum Genomics*. 2011;5(4):304. doi:10.1186/1479-7364-5-4-304
5. Verlouw JAM, Clemens E, de Vries JH, et al. A comparison of genotyping arrays. *Eur J Hum Genet 2021 2911*. 2021;29(11):1611-1624. doi:10.1038/s41431-021-00917-7
6. Tanjo T, Kawai Y, Tokunaga K, Ogasawara O, Nagasaki M. Practical guide for managing large-scale human genome data in research. *J Hum Genet 2020 661*. 2020;66(1):39-52. doi:10.1038/s10038-020-00862-1
7. He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. *Int J Mol Sci*. 2017;18(2). doi:10.3390/IJMS18020412
8. Hassan M, Awan FM, Naz A, et al. Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. *Int J Mol Sci*. 2022;23(9). doi:10.3390/IJMS23094645
9. Sherry ST, Ward MH, Kholodov M, et al. DbSNP: The NCBI database of genetic variation. *Nucleic Acids Res*. 2001;29(1):308-311. doi:10.1093/nar/29.1.308
10. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*. 2016;37(6):564-569. doi:10.1002/HUMU.22981
11. UK Biobank Axiom™ Array. Accessed December 7, 2023. https://www.thermofisher.com/order/catalog/product/es/en/902502
12. Naito T, Okada Y. HLA imputation and its application to genetic and molecular fine-mapping of the MHC region in autoimmune diseases. *Semin Immunopathol*. 2022;44(1):15. doi:10.1007/S00281-021-00901-9

13. Miko I. Gregor Mendel and the Principles of Inheritance. Nature Education. Published 2008. Accessed November 29, 2023. https://www.nature.com/scitable/topicpage/gregor-mendel-and-the-principles-of-inheritance-593/
14. User:Sciencia58 - Wikipedia. Accessed December 2, 2023. https://de.wikipedia.org/wiki/Benutzerin:Sciencia58
15. Perlman RL, Govindaraju DR. Archibald E. Garrod: the father of precision medicine. *Genet Med 2016 1811*. 2016;18(11):1088-1089. doi:10.1038/gim.2016.5
16. Borecki IB, Province MA. Genetic and genomic discovery using family studies. *Circulation*. 2008;118(10):1057-1063. doi:10.1161/CIRCULATIONAHA.107.714592
17. Wolf JB, Ferguson-Smith AC, Lorenz A. Mendel's laws of heredity on his 200th birthday: What have we learned by considering exceptions? *Hered 2022 1291*. 2022;129(1):1-3. doi:10.1038/s41437-022-00552-y
18. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15(8):1034-1050. doi:10.1101/GR.3715005
19. Acuna-Hidalgo R, Veltman JA, Hoischen A. New insights into the generation and role of de novo mutations in health and disease. *Genome Biol*. 2016;17(1). doi:10.1186/S13059-016-1110-1
20. Chen S, Francioli LC, Goodrich JK, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nat 2023*. Published online December 6, 2023:1-11. doi:10.1038/s41586-023-06045-0
21. Benton ML, Abraham A, LaBella AL, Abbot P, Rokas A, Capra JA. The influence of evolutionary history on human health and disease. *Nat Rev Genet 2021 225*. 2021;22(5):269-283. doi:10.1038/s41576-020-00305-9
22. Ma M, Ru Y, Chuang LS, et al. Disease-associated variants in different categories of disease located in distinct regulatory elements. *BMC Genomics*. 2015;16(8):1-13. doi:10.1186/1471-2164-16-S8-S3/TABLES/5
23. Uffelmann E, Huang QQ, Munung NS, et al. Genome-wide association studies. *Nat Rev Methods Prim 2021 11*. 2021;1(1):1-21. doi:10.1038/s43586-021-00056-9
24. Albert FW, Kruglyak L. The role of regulatory variation in complex traits and disease. *Nat Rev Genet 2015 164*. 2015;16(4):197-212. doi:10.1038/nrg3891
25. File:Chromosome icon.svg - Wikimedia Commons. Accessed December 5, 2023.

https://commons.wikimedia.org/wiki/File:Chromosome_icon.svg
26. Rubinacci S, Delaneau O, Marchini J. Genotype imputation using the Positional Burrows Wheeler Transform. *PLOS Genet*. 2020;16(11):e1009049. doi:10.1371/JOURNAL.PGEN.1009049
27. Browning BL, Zhou Y, Browning SR. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet*. 2018;103(3):338-348. doi:10.1016/j.ajhg.2018.07.015
28. Jin Y, Wang J, Bachtiar M, Chong SS, Lee CGL. Architecture of polymorphisms in the human genome reveals functionally important and positively selected variants in immune response and drug transporter genes. *Hum Genomics*. 2018;12(1). doi:10.1186/S40246-018-0175-1
29. Jia X, Han B, Onengut-Gumuscu S, et al. Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One*. 2013;8(6):e64683. doi:10.1371/JOURNAL.PONE.0064683
30. Zheng X, Shen J, Cox C, et al. HIBAG—HLA genotype imputation with attribute bagging. *Pharmacogenomics J 2014 142*. 2013;14(2):192-200. doi:10.1038/tpj.2013.18
31. Dilthey AT, Moutsianas L, Leslie S, McVean G. HLA*IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes. *Bioinformatics*. 2011;27(7):968. doi:10.1093/BIOINFORMATICS/BTR061
32. Hormozdiari F, Kostem E, Kang EY, Pasaniuc B, Eskin E. Identifying causal variants at loci with multiple signals of association. *Genetics*. 2014;198(2):497-508. doi:10.1534/GENETICS.114.167908/-/DC1
33. Benner C, Spencer CCA, Havulinna AS, Salomaa V, Ripatti S, Pirinen M. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics*. 2016;32(10):1493. doi:10.1093/BIOINFORMATICS/BTW018
34. Kichaev G, Yang WY, Lindstrom S, et al. Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLOS Genet*. 2014;10(10):e1004722. doi:10.1371/JOURNAL.PGEN.1004722
35. Wang G, Sarkar A, Carbonetto P, Stephens M. A Simple New Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping. *J R Stat Soc Ser B Stat Methodol*. 2020;82(5):1273-1300. doi:10.1111/RSSB.12388
36. Okazaki A, Ott J. Machine learning approaches to explore digenic inheritance. *Trends Genet*. 2022;38(10):1013-1018. doi:10.1016/j.tig.2022.04.009
37. Winham SJ, Motsinger-Reif AA. An R package implementation of multifactor dimensionality reduction. *BioData Min*. 2011;4(1):24.

doi:10.1186/1756-0381-4-24
38. Moore JH, Andrews PC. Epistasis analysis using multifactor dimensionality reduction. *Methods Mol Biol*. 2015;1253. doi:10.1007/978-1-4939-2155-3_16
39. Zhang Q, Long Q, Ott J. AprioriGWAS, a new pattern mining strategy for detecting genetic variants associated with disease through interaction effects. *PLoS Comput Biol*. 2014;10(6). doi:10.1371/JOURNAL.PCBI.1003627
40. Borgelt C. An implementation of the FP-growth algorithm. *Proc ACM SIGKDD Int Conf Knowl Discov Data Min*. Published online 2005:1-5. doi:10.1145/1133905.1133907
41. Nasreen S, Azam MA, Shehzad K, Naeem U, Ghazanfar MA. Frequent Pattern Mining Algorithms for Finding Associated Frequent Patterns for Data Streams: A Survey. *Procedia Comput Sci*. 2014;37:109-116. doi:10.1016/J.PROCS.2014.08.019
42. Chen Y, Xu F, Pian C, et al. EpiMOGA: An Epistasis Detection Method Based on a Multi-Objective Genetic Algorithm. *Genes 2021, Vol 12, Page 191*. 2021;12(2):191. doi:10.3390/GENES12020191
43. Choi SW, Mak TSH, O'Reilly PF. A guide to performing Polygenic Risk Score analyses. *Nat Protoc*. 2020;15(9):2759. doi:10.1038/S41596-020-0353-1
44. Collister JA, Liu X, Clifton L. Calculating Polygenic Risk Scores (PRS) in UK Biobank: A Practical Guide for Epidemiologists. *Front Genet*. 2022;13:105. doi:10.3389/FGENE.2022.818574/BIBTEX
45. Choi SW, O'Reilly PF. PRSice-2: Polygenic Risk Score software for biobank-scale data. *Gigascience*. 2019;8(7). doi:10.1093/GIGASCIENCE/GIZ082
46. Craig JE, Han X, Qassim A, et al. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nat Genet*. 2020;52(2):160-166. doi:10.1038/S41588-019-0556-Y
47. Bigdeli TB, Voloudakis G, Barr PB, et al. Penetrance and Pleiotropy of Polygenic Risk Scores for Schizophrenia, Bipolar Disorder, and Depression Among Adults in the US Veterans Affairs Health Care System. *JAMA Psychiatry*. 2022;79(11):1092-1101. doi:10.1001/JAMAPSYCHIATRY.2022.2742
48. Zhang JP, Robinson D, Yu J, et al. Schizophrenia Polygenic Risk Score as a Predictor of Antipsychotic Efficacy in First Episode Psychosis. *Am J Psychiatry*. 2019;176(1):21. doi:10.1176/APPI.AJP.2018.17121363
49. Clark K, Leung YY, Lee WP, Voight B, Wang LS. Polygenic Risk

Scores in Alzheimer's Disease Genetics: Methodology, Applications, Inclusion, and Diversity. *J Alzheimers Dis*. 2022;89(1):1-12. doi:10.3233/JAD-220025
50. Sollis E, Mosaku A, Abid A, et al. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res*. 2023;51(D1):D977. doi:10.1093/NAR/GKAC1010
51. Mailman MD, Feolo M, Jin Y, et al. The NCBI dbGaP database of genotypes and phenotypes. *Nat Genet*. 2007;39(10):1181-1186. doi:10.1038/NG1007-1181
52. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. doi:10.1093/NAR/GKX1153
53. Piñero J, Bravo Á, Queralt-Rosinach N, et al. DisGeNET: A comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res*. 2017;45(D1):D833-D839. doi:10.1093/nar/gkw943
54. Lin J, Ngiam KY. How data science and AI-based technologies impact genomics. *Singapore Med J*. 2023;64(1):59-66. doi:10.4103/SINGAPOREMEDJ.SMJ-2021-438
55. Vabalas A, Gowen E, Poliakoff E, Casson AJ. Machine learning algorithm validation with a limited sample size. *PLoS One*. 2019;14(11):e0224365. doi:10.1371/JOURNAL.PONE.0224365
56. Zhu C, Byrd RH, Lu P, Nocedal J. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans Math Softw*. 1997;23(4):550-560. doi:10.1145/279232.279236
57. Morales JL, Nocedal J. Remark on algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound constrained optimization. *ACM Trans Math Softw*. 2011;38(1). doi:10.1145/2049662.2049669
58. Schmidt M, Le Roux N, Bach F. Minimizing finite sums with the stochastic average gradient. *Math Program*. 2017;162(1-2):83-112. doi:10.1007/S10107-016-1030-6/FIGURES/3
59. Defazio A, Bach F, Lacoste-Julien S. SAGA: A Fast Incremental Gradient Method With Support for Non-Strongly Convex Composite Objectives. *Adv Neural Inf Process Syst*. 2014;2(January):1646-1654. Accessed December 9, 2023. https://arxiv.org/abs/1407.0202v3
60. Friedman JH. Greedy function approximation: A gradient boosting machine. *https://doi.org/101214/aos/1013203451*. 2001;29(5):1189-1232. doi:10.1214/AOS/1013203451
61. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5-32. doi:10.1023/A:1010933404324/METRICS

62. Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. *Mach Learn*. 2006;63(1):3-42. doi:10.1007/S10994-006-6226-1/METRICS
63. Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press; 2016. http://www.deeplearningbook.org
64. Papadimitriou S, Gazzo A, Versbraegen N, et al. Predicting disease-causing variant combinations. *Proc Natl Acad Sci U S A*. 2019;116(24):11878-11887. doi:10.1073/PNAS.1815601116/-/DCSUPPLEMENTAL
65. Ott J, Park T. Overview of frequent pattern mining. *Genomics Inform*. 2022;20(4). doi:10.5808/GI.22074
66. Lipton ZC. The Mythos of Model Interpretability. *Commun ACM*. 2016;61(10):35-43. doi:10.1145/3233231
67. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey Of Methods For Explaining Black Box Models. *ACM Comput Surv*. 2018;51(5). doi:10.1145/3236009
68. Molnar C, Casalicchio G, Bischl B. Interpretable Machine Learning -- A Brief History, State-of-the-Art and Challenges. *Commun Comput Inf Sci*. 2020;1323:417-431. doi:10.1007/978-3-030-65965-3_28
69. Bhat A, Lucek PR, Ott J. Analysis of complex traits using neural networks. *Genet Epidemiol*. 1999;17 Suppl 1(SUPPL. 1). doi:10.1002/GEPI.1370170781
70. Mukherjee S, Cogan JD, Newman JH, et al. Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network. *Am J Hum Genet*. 2021;108(10):1946-1963. doi:10.1016/J.AJHG.2021.08.010
71. Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. *ACM Comput Surv*. 2021;54(6). doi:10.1145/3457607
72. Alvarellos M, Sheppard HE, Knarston I, et al. Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics. *Front Genet*. 2022;13. doi:10.3389/FGENE.2022.1045450
73. Boscarino N, Cartwright RA, Fox K, Tsosie KS. Federated learning and Indigenous genomic data sovereignty. *Nat Mach Intell 2022 411*. 2022;4(11):909-911. doi:10.1038/s42256-022-00551-y
74. Kolobkov D, Sharma SM, Medvedev A, Lebedev M, Kosaretskiy E, Vakhitov R. Efficacy of federated learning on genomic data: a study on the UK Biobank and the 1000 Genomes Project. *medRxiv*. Published online February 9, 2023:2023.01.24.23284898. doi:10.1101/2023.01.24.23284898

75. Chen M, Shlezinger N, Vincent Poor H, Eldar YC, Cui S. Communication-efficient federated learning. *Proc Natl Acad Sci U S A*. 2021;118(17):e2024789118. doi:10.1073/PNAS.2024789118/SUPPL_FILE/PNAS.2024789118.SAPP.PDF
76. Liu P, Xu X, Wang W. Threats, attacks and defenses to federated learning: issues, taxonomy and perspectives. *Cybersecurity*. 2022;5(1):1-19. doi:10.1186/S42400-021-00105-6/TABLES/4
77. Ye M, Fang X, Du B, Yuen PC, Tao D. Heterogeneous Federated Learning: State-of-the-art and Research Challenges. *ACM Comput Surv*. 2023;56(3):1-44. doi:10.1145/3625558
78. McGinley MP, Goldschmidt CH, Rae-Grant AD. Diagnosis and Treatment of Multiple Sclerosis: A Review. *JAMA*. 2021;325(8):765-779. doi:10.1001/JAMA.2020.26858
79. Hollenbach JA, Oksenberg JR. The Immunogenetics of Multiple Sclerosis: A Comprehensive Review. *J Autoimmun*. 2015;64:13. doi:10.1016/J.JAUT.2015.06.010
80. Sawcer S, Hellenthal G, Pirinen M, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. *Nature*. 2011;476(7359):214-219. doi:10.1038/NATURE10251
81. Wang R. Mendelian randomization study updates the effect of 25-hydroxyvitamin D levels on the risk of multiple sclerosis. *J Transl Med*. 2022;20(1). doi:10.1186/S12967-021-03205-6
82. Almohmeed YH, Avenell A, Aucott L, Vickers MA. Systematic review and meta-analysis of the sero-epidemiological association between Epstein Barr virus and multiple sclerosis. *PLoS One*. 2013;8(4). doi:10.1371/JOURNAL.PONE.0061110
83. Biström M, Jons D, Engdahl E, et al. Epstein–Barr virus infection after adolescence and human herpesvirus 6A as risk factors for multiple sclerosis. *Eur J Neurol*. 2021;28(2):579-586. doi:10.1111/ENE.14597
84. Houen G, Trier NH, Frederiksen JL. Epstein-Barr Virus and Multiple Sclerosis. *Front Immunol*. 2020;11. doi:10.3389/FIMMU.2020.587078
85. Afrasiabi A, Parnell GP, Swaminathan S, Stewart GJ, Booth DR. The interaction of Multiple Sclerosis risk loci with Epstein-Barr virus phenotypes implicates the virus in pathogenesis. *Sci Reports 2020 101*. 2020;10(1):1-11. doi:10.1038/s41598-019-55850-z
86. Soldan SS, Lieberman PM. Epstein–Barr virus and multiple sclerosis. *Nat Rev Microbiol*. 2023;21(1):51. doi:10.1038/S41579-022-00770-5

87. Lanz T V., Brewer RC, Ho PP, et al. Clonally expanded B cells in multiple sclerosis bind EBV EBNA1 and GlialCAM. *Nat 2022 6037900*. 2022;603(7900):321-327. doi:10.1038/s41586-022-04432-7
88. Hauser SL, Cree BAC. Treatment of Multiple Sclerosis: A Review. *Am J Med*. 2020;133(12):1380-1390.e2. doi:10.1016/J.AMJMED.2020.05.049
89. Beecham AH, Patsopoulos NA, Xifara DK, et al. Analysis of immune-related loci identifies 48 new susceptibility variants for multiple sclerosis. *Nat Genet 2013 4511*. 2013;45(11):1353-1360. doi:10.1038/ng.2770
90. De Jager PL, Jia X, Wang J, et al. Meta-analysis of genome scans and replication identify CD6, IRF8 and TNFRSF1A as new multiple sclerosis susceptibility loci. *Nat Genet*. 2009;41(7):776. doi:10.1038/NG.401
91. Patsopoulos NA, Baranzini SE, Santaniello A, et al. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science*. 2019;365(6460). doi:10.1126/SCIENCE.AAV7188
92. Shams H, Shao X, Santaniello A, et al. Polygenic risk score association with multiple sclerosis susceptibility and phenotype in Europeans. *Brain*. 2023;146(2):645-656. doi:10.1093/BRAIN/AWAC092
93. Breedon JR, Marshall CR, Giovannoni G, Van Heel DA, Dobson R, Jacobs BM. Polygenic risk score prediction of multiple sclerosis in individuals of South Asian ancestry. *Brain Commun*. 2023;5(2). doi:10.1093/BRAINCOMMS/FCAD041
94. Fuh-Ngwa V, Zhou Y, Melton PE, et al. Ensemble machine learning identifies genetic loci associated with future worsening of disability in people with multiple sclerosis. *Sci Reports 2022 121*. 2022;12(1):1-13. doi:10.1038/s41598-022-23685-w
95. Ghafouri-Fard S, Taheri M, Omrani MD, Daaee A, Mohammad-Rahimi H. Application of Artificial Neural Network for Prediction of Risk of Multiple Sclerosis Based on Single Nucleotide Polymorphism Genotypes. *J Mol Neurosci*. 2020;70(7):1081-1087. doi:10.1007/S12031-020-01514-X/FIGURES/5
96. Burnard SM, Lea RA, Benton M, et al. Capturing SNP Association across the NK Receptor and HLA Gene Regions in Multiple Sclerosis by Targeted Penalised Regression Models. *Genes (Basel)*. 2022;13(1):87. doi:10.3390/GENES13010087/S1
97. Briggs FBS, Sept C. Mining Complex Genetic Patterns Conferring Multiple Sclerosis Risk. *Int J Environ Res Public Health*. 2021;18(5):1-12. doi:10.3390/IJERPH18052518

98. Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E. Alzheimer's disease. *Lancet*. 2011;377(9770):1019-1031. doi:10.1016/S0140-6736(10)61349-9
99. Boller F, Forbes MM. History of dementia and dementia in history: An overview. *J Neurol Sci*. 1998;158(2):125-133. doi:10.1016/S0022-510X(98)00128-2
100. Cacace R, Sleegers K, Van Broeckhoven C. Molecular genetics of early-onset Alzheimer's disease revisited. *Alzheimer's Dement*. 2016;12(6):733-748. doi:10.1016/j.jalz.2016.01.012
101. Rabinovici GD. Late-onset Alzheimer disease. *Contin Lifelong Learn Neurol*. 2019;25(1):14-33. doi:10.1212/CON.0000000000000700
102. Toyota Y, Ikeda M, Shinagawa S, et al. Comparison of behavioral and psychological symptoms in early-onset and late-onset Alzheimer's disease. *Int J Geriatr Psychiatry*. 2007;22(9):896-901. doi:10.1002/gps.1760
103. Deture MA, Dickson DW. The neuropathological diagnosis of Alzheimer's disease. *Mol Neurodegener*. 2019;14(1). doi:10.1186/s13024-019-0333-5
104. Gatz M, Reynolds CA, Fratiglioni L, et al. Role of genes and environments for explaining Alzheimer disease. *Arch Gen Psychiatry*. 2006;63(2):168-174. doi:10.1001/archpsyc.63.2.168
105. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: The AlzGene database. *Nat Genet*. 2007;39(1):17-23. doi:10.1038/ng1934
106. de la Fuente J, Grotzinger AD, Marioni RE, Nivard MG, Tucker-Drob EM. Integrated analysis of direct and proxy genome wide association studies highlights polygenicity of Alzheimer's disease outside of the APOE region. *PLOS Genet*. 2022;18(6):e1010208. doi:10.1371/JOURNAL.PGEN.1010208
107. Andrews SJ, Fulton-Howard B, Goate A. Interpretation of risk loci from genome-wide association studies of Alzheimer's disease. *Lancet Neurol*. 2020;19(4):326-335. doi:10.1016/S1474-4422(19)30435-1
108. Bellenguez C, Küçükali F, Jansen IE, et al. New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet 2022 544*. 2022;54(4):412-436. doi:10.1038/s41588-022-01024-z
109. Jansen IE, Savage JE, Watanabe K, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet 2019 513*. 2019;51(3):404-413. doi:10.1038/s41588-018-0311-9

110. Marioni RE, Harris SE, Zhang Q, et al. GWAS on family history of Alzheimer's disease. *Transl Psychiatry 2018 81*. 2018;8(1):1-7. doi:10.1038/s41398-018-0150-6
111. Kunkle BW, Grenier-Boley B, Sims R, et al. Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat Genet 2019 513*. 2019;51(3):414-430. doi:10.1038/s41588-019-0358-2
112. Leonenko G, Sims R, Shoai M, et al. Polygenic risk and hazard scores for Alzheimer's disease prediction. *Ann Clin Transl Neurol*. 2019;6(3):456-465. doi:10.1002/ACN3.716
113. de Rojas I, Moreno-Grau S, Tesi N, et al. Common variants in Alzheimer's disease and risk stratification by polygenic risk scores. *Nat Commun 2021 121*. 2021;12(1):1-16. doi:10.1038/s41467-021-22491-8
114. Tan CH, Hyman BT, Tan JJX, et al. Polygenic hazard scores in preclinical Alzheimer'sdisease. *Ann Neurol*. 2017;82(3):484. doi:10.1002/ANA.25029
115. De Velasco Oriol J, Vallejo EE, Estrada K, Taméz Peña JG, Disease Neuroimaging Initiative TA s. Benchmarking machine learning models for late-onset Alzheimer's disease prediction from genomic data. *BMC Bioinformatics*. 2019;20(1):709. doi:10.1186/s12859-019-3158-x
116. Romero-Rosales BL, Tamez-Pena JG, Nicolini H, Moreno-Treviño MG, Trevino V. Improving predictive models for Alzheimer's disease using GWAS data by incorporating misclassified samples modeling. *PLoS One*. 2020;15(4). doi:10.1371/journal.pone.0232103
117. Gyawali PK, Le Guen Y, Liu X, et al. Improving genetic risk prediction across diverse population by disentangling ancestry representations. *Commun Biol*. 2023;6(1):964. doi:10.1038/S42003-023-05352-6
118. Jemimah S, AlShehhi A. c-Diadem: a constrained dual-input deep learning model to identify novel biomarkers in Alzheimer's disease. *BMC Med Genomics 2023 162*. 2023;16(2):1-13. doi:10.1186/S12920-023-01675-9
119. Chandrashekar PB, Alatkar S, Wang J, et al. DeepGAMI: deep biologically guided auxiliary learning for multimodal integration and imputation to improve genotype–phenotype prediction. *Genome Med*. 2023;15(1):88. doi:10.1186/S13073-023-01248-6
120. Vivek S, Faul J, Thyagarajan B, Guan W. Explainable variational autoencoder (E-VAE) model using genome-wide SNPs to predict dementia. *J Biomed Inform*. 2023;148. doi:10.1016/J.JBI.2023.104536

121. Chang YC, Wu JT, Hong MY, et al. GenEpi: Gene-based epistasis discovery using machine learning. *BMC Bioinformatics*. 2020;21(1):1-13. doi:10.1186/S12859-020-3368-2/FIGURES/5
122. Bayat A, Hosking B, Jain Y, et al. Fast and accurate exhaustive higher-order epistasis search with BitEpi. *Sci Reports 2021 111*. 2021;11(1):1-12. doi:10.1038/s41598-021-94959-y
123. Lundberg M, Sng LMF, Szul P, et al. Novel Alzheimer's disease genes and epistasis identified using machine learning GWAS platform. *Sci Rep*. 2023;13(1):17662. doi:10.1038/S41598-023-44378-Y
124. Johannessen JO. Review: lifetime prevalence of schizophrenia and related disorders is about 5.5 per 1000, but there is significant variation between regions. *Evid Based Ment Health*. 2003;6(3):74. doi:10.1136/EBMH.6.3.74
125. Kraguljac NV, Lahti AC. Neuroimaging as a Window Into the Pathophysiological Mechanisms of Schizophrenia. *Front Psychiatry*. 2021;12:613764. doi:10.3389/FPSYT.2021.613764
126. Shnayder NA, Novitsky MA, Neznanov NG, et al. Genetic Predisposition to Schizophrenia and Depressive Disorder Comorbidity. *Genes (Basel)*. 2022;13(3). doi:10.3390/GENES13030457
127. Power RA, Verweij KJH, Zuhair M, et al. Genetic predisposition to schizophrenia associated with increased use of cannabis. *Mol Psychiatry*. 2014;19(11):1201-1204. doi:10.1038/MP.2014.51
128. Alnæs D, Kaufmann T, Van Der Meer D, et al. Brain Heterogeneity in Schizophrenia and Its Association With Polygenic Risk. *JAMA psychiatry*. 2019;76(7):739-748. doi:10.1001/JAMAPSYCHIATRY.2019.0257
129. Pardiñas AF, Smart SE, Willcocks IR, et al. Interaction Testing and Polygenic Risk Scoring to Estimate the Association of Common Genetic Variants With Treatment Resistance in Schizophrenia. *JAMA Psychiatry*. 2022;79(3):260-269. doi:10.1001/JAMAPSYCHIATRY.2021.3799
130. Richards AL, Pardiñas AF, Frizzati A, et al. The Relationship Between Polygenic Risk Scores and Cognition in Schizophrenia. *Schizophr Bull*. 2020;46(2):336-344. doi:10.1093/SCHBUL/SBZ061
131. Habtewold TD, Liemburg EJ, Islam MA, et al. Association of schizophrenia polygenic risk score with data-driven cognitive subtypes: A six-year longitudinal study in patients, siblings and controls. *Schizophr Res*. 2020;223:135-147. doi:10.1016/J.SCHRES.2020.05.020
132. Pillinger T, Osimo EF, de Marvao A, et al. Effect of polygenic risk

for schizophrenia on cardiac structure and function: a UK Biobank observational study. *The Lancet Psychiatry*. 2023;10(2):98-107. doi:10.1016/S2215-0366(22)00403-5

133. Bracher-Smith M, Crawford K, Escott-Price V. Machine learning for genetic prediction of psychiatric disorders: a systematic review. *Mol Psychiatry*. 2021;26(1):70-79. doi:10.1038/s41380-020-0825-2
134. Vivian-Griffiths T, Baker E, Schmidt KM, et al. Predictive modeling of schizophrenia from genomic data: Comparison of polygenic risk score with kernel support vector machines approach. *Am J Med Genet B Neuropsychiatr Genet*. 2019;180(1):80-85. doi:10.1002/AJMG.B.32705
135. Yang H, Liu J, Sui J, Pearlson G, Calhoun VD. A Hybrid Machine Learning Method for Fusing fMRI and Genetic Data: Combining both Improves Classification of Schizophrenia. *Front Hum Neurosci*. 2010;4. doi:10.3389/FNHUM.2010.00192
136. Li G, Han D, Wang C, Hu W, Calhoun VD, Wang YP. Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. *Comput Methods Programs Biomed*. 2020;183. doi:10.1016/J.CMPB.2019.105073
137. Rahaman MA, Chen J, Fu Z, Lewis N, Iraji A, Calhoun VD. Multi-modal deep learning of functional and structural neuroimaging and genomic data to predict mental illness. *Annu Int Conf IEEE Eng Med Biol Soc IEEE Eng Med Biol Soc Annu Int Conf*. 2021;2021:3267-3272. doi:10.1109/EMBC46164.2021.9630693
138. van Hilten A, Kushner SA, Kayser M, et al. GenNet framework: interpretable deep learning for predicting phenotypes from genetic data. *Commun Biol 2021 41*. 2021;4(1):1-9. doi:10.1038/s42003-021-02622-z
139. Jo Y, Webster MJ, Kim S, Lee D. Interpretation of SNP combination effects on schizophrenia etiology based on stepwise deep learning with multi-precision data. *Brief Funct Genomics*. Published online September 21, 2023. doi:10.1093/BFGP/ELAD041
140. Jankovic J. Parkinson's disease: clinical features and diagnosis. *J Neurol Neurosurg Psychiatry*. 2008;79(4):368-376. doi:10.1136/JNNP.2007.131045
141. Cerri S, Mus L, Blandini F. Parkinson's Disease in Women and Men: What's the Difference? *J Parkinsons Dis*. 2019;9(3):501. doi:10.3233/JPD-191683
142. Meder D, Herz DM, Rowe JB, Lehéricy S, Siebner HR. The role of dopamine in the brain - lessons learned from Parkinson's disease. *Neuroimage*. 2019;190:79-93.

doi:10.1016/J.NEUROIMAGE.2018.11.021
143. Do CB, Tung JY, Dorfman E, et al. Web-Based Genome-Wide Association Study Identifies Two Novel Loci and a Substantial Genetic Component for Parkinson's Disease. *PLOS Genet*. 2011;7(6):e1002141. doi:10.1371/JOURNAL.PGEN.1002141
144. Pan H, Liu Z, Ma J, et al. Genome-wide association study using whole-genome sequencing identifies risk loci for Parkinson's disease in Chinese population. *npj Park Dis 2023 91*. 2023;9(1):1-11. doi:10.1038/s41531-023-00456-6
145. Blauwendraat C, Heilbron K, Vallerga CL, et al. Parkinson disease age at onset GWAS: defining heritability, genetic loci and α-synuclein mechanisms. *Mov Disord*. 2019;34(6):866. doi:10.1002/MDS.27659
146. Nalls MA, Blauwendraat C, Vallerga CL, et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol*. 2019;18(12):1091-1102. doi:10.1016/S1474-4422(19)30320-5
147. Nussbaum RL. The Identification of Alpha-Synuclein as the First Parkinson Disease Gene. *J Parkinsons Dis*. 2017;7(Suppl 1):S43. doi:10.3233/JPD-179003
148. Jacobs BM, Belete D, Bestwick J, et al. Parkinson's disease determinants, prediction and gene-environment interactions in the UK Biobank. *J Neurol Neurosurg Psychiatry*. 2020;91(10):1046-1054. doi:10.1136/JNNP-2020-323646
149. Koch S, Laabs BH, Kasten M, et al. Validity and Prognostic Value of a Polygenic Risk Score for Parkinson's Disease. *Genes (Basel)*. 2021;12(12). doi:10.3390/GENES12121859
150. Koch S, Schmidtke J, Krawczak M, Caliebe A. Clinical utility of polygenic risk scores: a critical 2023 appraisal. *J Community Genet*. Published online 2023. doi:10.1007/S12687-023-00645-Z
151. Nguyen TT, Huang JZ, Wu Q, Nguyen TT, Li MJ. Genome-wide association data classification and SNPs selection using two-stage quality-based random forests. *BMC Genomics*. 2015;16(Suppl 2). doi:10.1186/1471-2164-16-S2-S5
152. Rahman MA, Liu J. A genome-wide association study coupled with machine learning approaches to identify influential demographic and genomic factors underlying Parkinson's disease. *Front Genet*. 2023;14:1230579. doi:10.3389/FGENE.2023.1230579/BIBTEX
153. Rodrigo LM, Nyholt DR. Imputation and Reanalysis of ExomeChip Data Identifies Novel, Conditional and Joint Genetic Effects on Parkinson's Disease Risk. *Genes (Basel)*. 2021;12(5).

doi:10.3390/GENES12050689
154. Hafler DA, Compston A, Sawcer S, et al. Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med*. 2007;357(9):851-862. doi:10.1056/NEJMOA073493
155. Alvarez-Sanchez N, Dunn SE. Immune Cell Contributors to the Female Sex Bias in Multiple Sclerosis and Experimental Autoimmune Encephalomyelitis. *Curr Top Behav Neurosci*. 2023;62:333-373. doi:10.1007/7854_2022_324
156. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *npj Digit Med 2021 41*. 2021;4(1):1-8. doi:10.1038/s41746-021-00521-5
157. Sawcer S, Jones HB, Feakes R, et al. A genome screen in multiple sclerosis reveals susceptibility loci on chromosome 6p21 and 17q22. *Nat Genet*. 1996;13(4):464-468. doi:10.1038/NG0896-464
158. Chartier-hariln MC, Parfitt M, Legrain S, et al. Apolipoprotein E, ε4 allele as a major risk factor for sporadic early and late-onset forms of Alzheimer's disease: analysis of the 19q13.2 chromosomal region. *Hum Mol Genet*. 1994;3(4):569-574. doi:10.1093/HMG/3.4.569
159. Huang YWA, Zhou B, Wernig M, Südhof TC. ApoE2, ApoE3, and ApoE4 Differentially Stimulate APP Transcription and Aβ Secretion. *Cell*. 2017;168(3):427-441.e21. doi:10.1016/j.cell.2016.12.044
160. Cheng J, Novati G, Pan J, et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science*. 2023;381(6664):eadg7492. doi:10.1126/SCIENCE.ADG7492/SUPPL_FILE/SCIENCE.ADG7492_DATA_S1_TO_S9.ZIP
161. Goris A, Vandebergh M, McCauley JL, Saarela J, Cotsapas C. Genetics of multiple sclerosis: lessons from polygenicity. *Lancet Neurol*. 2022;21(9):830-842. doi:10.1016/S1474-4422(22)00255-1
162. Lee LF, Logronio K, Tu GH, et al. Anti-IL-7 receptor-α reverses established type 1 diabetes in nonobese diabetic mice by modulating effector T-cell function. *Proc Natl Acad Sci U S A*. 2012;109(31):12674-12679. doi:10.1073/PNAS.1203795109
163. Douroudis K, Nemvalts V, Rajasalu T, Kisand K, Uibo R. The CD226 gene in susceptibility of type 1 diabetes. *Tissue Antigens*. 2009;74(5):417-419. doi:10.1111/J.1399-0039.2009.01320.X
164. Meyer A, Parmar PJ, Shahrara S. Significance of IL-7 and IL-7R in RA and autoimmunity. *Autoimmun Rev*. 2022;21(7). doi:10.1016/J.AUTREV.2022.103120

165. Tan RJL, Gibbons LJ, Potter C, et al. Investigation of rheumatoid arthritis susceptibility genes identifies association of AFF3 and CD226 variants with response to anti-tumour necrosis factor treatment. *Ann Rheum Dis*. 2010;69(6):1029-1035. doi:10.1136/ARD.2009.118406
166. Caillier SJ, Briggs F, Cree BAC, et al. Uncoupling the roles of HLA-DRB1 and HLA-DRB5 genes in multiple sclerosis. *J Immunol*. 2008;181(8):5473-5480. doi:10.4049/JIMMUNOL.181.8.5473
167. Brynedal B, Duvefelt K, Jonasdottir G, et al. HLA-A Confers an HLA-DRB1 Independent Influence on the Risk of Multiple Sclerosis. *PLoS One*. 2007;2(7):e664. doi:10.1371/JOURNAL.PONE.0000664
168. Bergamaschi L, Leone MA, Fasano ME, et al. HLA-class I markers and multiple sclerosis susceptibility in the Italian population. *Genes Immun*. 2010;11(2):173-180. doi:10.1038/GENE.2009.101
169. Menegatti J, Schub D, Schäfer M, Grässer FA, Ruprecht K. HLA-DRB1*15:01 is a co-receptor for Epstein–Barr virus, linking genetic and environmental risk factors for multiple sclerosis. *Eur J Immunol*. 2021;51(9):2348-2350. doi:10.1002/EJI.202149179
170. González-Jiménez A, López-Cotarelo P, Agudo-Jiménez T, et al. Impact of Multiple Sclerosis Risk Polymorphism rs7665090 on MANBA Activity, Lysosomal Endocytosis, and Lymphocyte Activation. *Int J Mol Sci*. 2022;23(15). doi:10.3390/IJMS23158116
171. Law SPL, Gatt PN, Schibeci SD, et al. Expression of CYP24A1 and other multiple sclerosis risk genes in peripheral blood indicates response to vitamin D in homeostatic and inflammatory conditions. *Genes Immun*. 2021;22(4):227-233. doi:10.1038/S41435-021-00144-6
172. Gadani SP, Cronk JC, Norris GT, Kipnis J. Interleukin-4: A Cytokine to Remember. *J Immunol*. 2012;189(9):4213. doi:10.4049/JIMMUNOL.1202246
173. Wang K, Song F, Fernandez-Escobar A, Luo G, Wang JH, Sun Y. The Properties of Cytokines in Multiple Sclerosis: Pros and Cons. *Am J Med Sci*. 2018;356(6):552-560. doi:10.1016/J.AMJMS.2018.08.018
174. Kallaur AP, Oliveira SR, Simao ANC, et al. Cytokine profile in relapsing-remitting multiple sclerosis patients and the association between progression and activity of the disease. *Mol Med Rep*. 2013;7(3):1010-1020. doi:10.3892/MMR.2013.1256
175. Vogelaar CF, Mandal S, Lerch S, et al. Fast direct neuronal signaling via the IL-4 receptor as therapeutic target in

neuroinflammation. *Sci Transl Med*. 2018;10(430). doi:10.1126/SCITRANSLMED.AAO2304
176. Akbarian F, Tabatabaiefar MA, Shaygannejad V, et al. Upregulation of MTOR, RPS6KB1, and EIF4EBP1 in the whole blood samples of Iranian patients with multiple sclerosis compared to healthy controls. *Metab Brain Dis*. 2020;35(8):1309-1316. doi:10.1007/S11011-020-00590-7/FIGURES/2
177. Ward-Kavanagh LK, Lin WW, Šedý JR, Ware CF. The TNF Receptor Superfamily in Co-stimulating and Co-inhibitory Responses. *Immunity*. 2016;44(5):1005-1019. doi:10.1016/J.IMMUNI.2016.04.019
178. Javor J, Shawkatová I, Ďurmanová V, et al. TNFRSF1A polymorphisms and their role in multiple sclerosis susceptibility and severity in the Slovak population. *Int J Immunogenet*. 2018;45(5):257-265. doi:10.1111/IJI.12388
179. Ottoboni L, Frohlich IY, Lee M, et al. Clinical relevance and functional consequences of the TNFRSF1A multiple sclerosis locus. *Neurology*. 2013;81(22):1891. doi:10.1212/01.WNL.0000436612.66328.8A
180. ten Bosch GJA, Bolk J, 't Hart BA, Laman JD. Multiple sclerosis is linked to MAPKERK overactivity in microglia. *J Mol Med*. 2021;99(8):1033-1042. doi:10.1007/S00109-021-02080-4
181. Leikfoss IS, Keshari PK, Gustavsen MW, et al. Multiple Sclerosis Risk Allele in CLEC16A Acts as an Expression Quantitative Trait Locus for CLEC16A and SOCS1 in CD4+ T Cells. *PLoS One*. 2015;10(7). doi:10.1371/JOURNAL.PONE.0132957
182. Berge T, Leikfoss IS, Harbo HF. From Identification to Characterization of the Multiple Sclerosis Susceptibility Gene CLEC16A. *Int J Mol Sci*. 2013;14(3):4476. doi:10.3390/IJMS14034476
183. Beltrán E, Gerdes LA, Hansen J, et al. Early adaptive immune activation detected in monozygotic twins with prodromal multiple sclerosis. *J Clin Invest*. 2019;129(11):4758-4768. doi:10.1172/JCI128475
184. Ng W, Minasny B, de Sousa Mendes W, Melo Demattê JA. The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data. *SOIL*. 2020;6(2):565-578. doi:10.5194/SOIL-6-565-2020
185. Raudys SJ, Jain AK. Small Sample Size Effects in Statistical Pattern Recognition: Recommendations for Practitioners. *IEEE Trans Pattern Anal Mach Intell*. 1991;13(3):252-264. doi:10.1109/34.75512

186. Dong Y, Zhou S, Xing L, et al. Deep learning methods may not outperform other machine learning methods on analyzing genomic studies. *Front Genet*. 2022;13. doi:10.3389/FGENE.2022.992070/FULL
187. Bergtold JS, Yeager EA, Featherstone AM. Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. *J Appl Stat*. 2018;45(3):528-546. doi:10.1080/02664763.2017.1282441
188. Xu H, Kinfu KA, LeVine W, et al. When are Deep Networks really better than Decision Forests at small sample sizes, and how? Published online 2021.
189. Tolosa E, Garrido A, Scholz SW, Poewe W. Challenges in the diagnosis of Parkinson's disease. *Lancet Neurol*. 2021;20(5):385-397. doi:10.1016/S1474-4422(21)00030-2
190. Sharaev MG, Malashenkova IK, Maslennikova V, et al. Diagnosis of Schizophrenia Based on the Data of Various Modalities: Biomarkers and Machine Learning Techniques (Review). *Mod Technol Med*. 2022;14(5):53. doi:10.17691/STM2022.14.5.06
191. Li R, Ma X, Wang G, Yang J, Wang C. Why sex differences in schizophrenia? *J Transl Neurosci*. 2016;1(1):37. Accessed November 14, 2023. /pmc/articles/PMC5688947/
192. Lindamer LA, Lohr JB, Harris MJ, McAdams LA, Jeste D V. Gender-related clinical differences in older patients with schizophrenia. *J Clin Psychiatry*. 1999;60(1):61-67. doi:10.4088/JCP.V60N0114
193. Vatcheva KP, Lee M, McCormick JB, Mohammad RH. Multicollinearity in Regression Analyses Conducted in Epidemiologic Studies. *Epidemiol (Sunnyvale, Calif)*. 2016;6(2). doi:10.4172/2161-1165.1000227
194. Toloşi L, Lengauer T. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*. 2011;27(14):1986-1994. doi:10.1093/BIOINFORMATICS/BTR300
195. Alzubaidi L, Zhang J, Humaidi AJ, et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data 2021 81*. 2021;8(1):1-74. doi:10.1186/S40537-021-00444-8
196. Shrestha A, Mahmood A. Review of deep learning algorithms and architectures. *IEEE Access*. 2019;7:53040-53065. doi:10.1109/ACCESS.2019.2912200
197. Schwartzentruber J, Cooper S, Liu JZ, et al. Genome-wide meta-analysis, fine-mapping and integrative prioritization implicate new Alzheimer's disease risk genes. *Nat Genet*. Published online

February 15, 2021:1-11. doi:10.1038/s41588-020-00776-w
198. Ware EB, Faul JD, Mitchell CM, Bakulski KM. Considering the APOE locus in Alzheimer's disease polygenic scores in the Health and Retirement Study: a longitudinal panel study. *BMC Med Genomics*. 2020;13(1):1-13. doi:10.1186/S12920-020-00815-9/TABLES/2
199. Misicka E, Davis MF, Kim W, et al. A higher burden of multiple sclerosis genetic risk confers an earlier onset. *Mult Scler*. 2022;28(8):1189-1197. doi:10.1177/13524585211053155
200. Traugott U. Multiple sclerosis: relevance of Class I and Class II MHC-expressing cells to lesion development. *J Neuroimmunol*. 1987;16(2):283-302. doi:10.1016/0165-5728(87)90082-8
201. Moutsianas L, Jostins L, Beecham AH, et al. Class II HLA interactions modulate genetic risk for multiple sclerosis. *Nat Genet*. 2015;47(10):1107. doi:10.1038/NG.3395
202. Versbraegen N, Gravel B, Nachtegael C, et al. Faster and more accurate pathogenic combination predictions with VarCoPP2.0. *BMC Bioinformatics*. 2023;24(1):1-19. doi:10.1186/S12859-023-05291-3/TABLES/4
203. Freeberg MA, Fromont LA, D'Altri T, et al. The European Genome-phenome Archive in 2021. *Nucleic Acids Res*. 2022;50(D1):D980-D987. doi:10.1093/NAR/GKAB1059
204. Gronemeyer H, Souren NY. Big Data: The good, the bad and the ugly. *Int J cancer*. 2021;148(12):2870-2871. doi:10.1002/IJC.33466
205. Budach L, Feuerpfeil M, Ihde N, et al. The Effects of Data Quality on Machine Learning Performance. Published online July 29, 2022. Accessed November 10, 2023. https://arxiv.org/abs/2207.14529v4
206. Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data 2016 31*. 2016;3(1):1-9. doi:10.1038/sdata.2016.18
207. Wright CF, West B, Tuke M, et al. Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am J Hum Genet*. 2019;104(2):275-286. doi:10.1016/J.AJHG.2018.12.015
208. Virgolin M, Alderliesten T, Bosman PAN. On explaining machine learning models by evolving crucial and compact features. *Swarm Evol Comput*. 2020;53:100640. doi:10.1016/J.SWEVO.2019.100640
209. Bonomi L, Huang Y, Ohno-Machado L. Privacy challenges and research opportunities for genomic data sharing. *Nat Genet 2020*

527. 2020;52(7):646-654. doi:10.1038/s41588-020-0651-0
210. Mohammed Yakubu A, Chen YPP. Ensuring privacy and security of genomic data and functionalities. *Brief Bioinform*. 2020;21(2):511-526. doi:10.1093/BIB/BBZ013
211. Dankar FK, Gergely M, Dankar SK. Informed Consent in Biomedical Research. *Comput Struct Biotechnol J*. 2019;17:463. doi:10.1016/J.CSBJ.2019.03.010
212. Bentley AR, Callier SL, Rotimi CN. Evaluating the promise of inclusion of African ancestry populations in genomics. *NPJ genomic Med*. 2020;5(1). doi:10.1038/S41525-019-0111-X
213. Stiglic G, Kocbek P, Fijacko N, Zitnik M, Verbert K, Cilar L. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2020;10(5):e1379. doi:10.1002/WIDM.1379
214. Moseson H, Zazanis N, Goldberg E, et al. The Imperative for Transgender and Gender Nonbinary Inclusion: Beyond Women's Health. *Obstet Gynecol*. 2020;135(5):1059. doi:10.1097/AOG.0000000000003816
215. Shamambo LJ, Henry TL. Rethinking the Use of "Caucasian" in Clinical Language and Curricula: a Trainee's Call to Action. *J Gen Intern Med*. 2022;37(7):1780-1782. doi:10.1007/S11606-022-07431-6
216. Saykin AJ, Shen L, Yao X, et al. Genetic studies of quantitative MCI and AD phenotypes in ADNI: Progress, opportunities, and plans. doi:10.1016/j.jalz.2015.05.009
217. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet*. 2007;81(3):559. doi:10.1086/519795
218. Lin WY, Liu N. Reducing bias of allele frequency estimates by modeling snp genotype data with informative missingness. *Front Genet*. 2012;3(JUN):107. doi:10.3389/FGENE.2012.00107/BIBTEX
219. Wang C, Schroeder KB, Rosenberg NA. A maximum-likelihood method to correct for allelic dropout in microsatellite data with no replicate genotypes. *Genetics*. 2012;192(2):651-669. doi:10.1534/GENETICS.112.139519
220. Danecek P, Bonfield JK, Liddle J, et al. Twelve years of SAMtools and BCFtools. *Gigascience*. 2021;10(2):1-4. doi:10.1093/GIGASCIENCE/GIAB008
221. Delaneau O, Zagury JF, Robinson MR, Marchini JL, Dermitzakis ET. Accurate, scalable and integrative haplotype estimation. *Nat Commun 2019 101*. 2019;10(1):1-10. doi:10.1038/s41467-019-13225-y

222. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
223. Shi J, Levinson DF, Duan J, et al. Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature*. 2009;460(7256):753-757. doi:10.1038/NATURE08192
224. Sundararajan M, Taly A, Yan Q. Axiomatic Attribution for Deep Networks. *34th Int Conf Mach Learn ICML 2017*. 2017;7:5109-5118. Accessed April 4, 2023. https://arxiv.org/abs/1703.01365v2
225. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. *34th Int Conf Mach Learn ICML 2017*. 2017;7:4844-4866. Accessed April 4, 2023. https://arxiv.org/abs/1704.02685v2
226. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *2nd Int Conf Learn Represent ICLR 2014 - Work Track Proc*. Published online December 20, 2013. Accessed April 4, 2023. https://arxiv.org/abs/1312.6034v2
227. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller M. Striving for Simplicity: The All Convolutional Net. *3rd Int Conf Learn Represent ICLR 2015 - Work Track Proc*. Published online December 21, 2014. Accessed April 4, 2023. https://arxiv.org/abs/1412.6806v3
228. Kotlyar M, Pastrello C, Sheahan N, Jurisica I. Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic Acids Res*. 2016;44(D1):D536-D541. doi:10.1093/NAR/GKV1115

7. List of publications

Vandelli A, **Arnal Segura M**, Monti M, et al. *The PRALINE database: protein and Rna humAn singLe nucleotIde variaNts in condEnsates.* Bioinformatics. 2023;39(1). doi:10.1093/BIOINFORMATICS/BTAC847

Arnal Segura M, Bini G, Krithara A, Paliouras G, Gaetano Tartaglia G. *Evaluation of Genomic Data-Based Machine Learning Methods for Classifying Complex Diseases*. bioRxiv 2024.03.18.585541; doi: https://doi.org/10.1101/2024.03.18.585541

Abstracts:

RNA transcripts of amyloid, stress granule proteins and neurodegenerative related genes present an enrichment of interactions with RNA-binding proteins with distinct regulatory patterns.
Arnal Segura M, Gaetano Tartaglia G.
Poster presentation at SIBBM 2022 “Frontiers in Molecular Biology, The RNA World 3.0” (June 20th - 22nd 2022), Sapienza University of Rome, (Rome, Italy)

8. Acknowledgements

I would like to express my gratitude to Gian Gaetano Tartaglia for providing me with the opportunity to pursue the Ph.D in his group, and to Claudia Giambartolomei, who, together with him, supervised my work.

I am also thankful to all the lab members who supported me during these years. Special thanks to Giorgio Bini, my computational colleague in Genova, from whom I learned a lot and who provided strong support throughout this journey. I am grateful to Jackob Ruppert, who made me laugh when I was down and showed me the best walking paths in Liguria. A special thanks also to Nuria Crua and Marta Bernad, my Catalan friends in Genova. Their generosity and friendliness made me feel like I had a family abroad. I appreciate Elsa Zacco, a great researcher and a strong woman who also took the time to organize activities to bring the group together.

PhD can be a difficult path sometimes, but it has positive things as well. Some of the people you meet during this journey are specially linked to you. I hope that these friendships will last forever.

Many thanks to Eduardo and Yeraldin, who shared the workspace with us at IIT and were always kind and supportive. I extend my appreciation to George Paliouras, Anastasia Krithara, and the entire team at NCSR Demokritos, who gave me the opportunity to visit their group in Athens and learn techniques that were crucial for the completion of this work.

Finally, I would like to express my heartfelt gratitude to my pillars, my family. It is thanks to them that I have achieved some of the most significant goals in my career. They have always believed in my potential, even during the periods when I felt broken. To them, my PhD represents something greater than I may never fully understand, and for this, I feel honored. I would specially like to express my gratitude to Antonio, who met me when I was starting this adventure in Genova, and he tried hard with me since the beginning. I have often heard people say that one finds love when least expected. In my case, this proved true. Without him, everything would have been much more challenging.