2020

LightGWAS: A Novel Genome-Wide Association Study Procedure

Bruno Ambrozio
Technological University Dublin, d16128063@mytudublin.ie

LightGWAS: A Novel Genome-Wide Association Study Procedure

Bruno Ambrozio

A dissertation submitted in partial fulfilment of the requirements of
Technological University Dublin for the degree of
M.Sc. in Computing (Data Analytics)

28 September 2020.

Declaration

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Technological University Dublin and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: Bruno Ambrozio.

Date: 28 September 2020.

Abstract

This dissertation proposes LightGWAS, a novel machine learning procedure for genome-wide association study (GWAS) based on LightGBM and k-fold cross-validation. The conducted literature review identified that the currently available GWAS implementations rely on massive manual quality control steps to address statistical issues, such as controlling for false-positive inflation and power reduction. It also showed they demand a specific GWAS method for each type of genomic dataset morphology, which consequently increases the human dependency and open margins for misleadings. LightGWAS is a potential single, resilient, autonomous and scalable solution to address such concerns. Through this research, LightGWAS was contrasted against the current state-of-the-art for GWAS throughout secondary research method. It has been compared with a GWAS implementation based on general linear model (GLM) with support to Firth regularisation. Quantitative empirical tests and deductive reasoning have been employed to reach and evaluate the results. The models were submitted to balanced ($case : control = 1 : 1$), imbalanced ($case : control = 1 : 10$), and high-imbalanced ($case : control = 1 : 100$) genomic datasets of binary phenotypes. The results from statistical tests denoted that LightGWAS performs equivalently to the compared GLM method for balanced dataset scenarios, and outperforms for imbalanced and high-imbalanced datasets. The assessed metrics were *weighted average of the precision and recall (F1)*, *recall*, *average precision score (APS)*, *receiver operating characteristic (ROC)/area under the curve (AUC)*, *accuracy*, and *precision*.

Keywords: LightGWAS, LightGBM, Genome-wide association study (GWAS), Machine Learning (ML).

Acknowledgments

Firstly, I would like to thank IBM for sponsoring me financially.

Still, as a primary thank, my dissertation advisor, Dr Lucas Rizzo. Despite the worldwide declared pandemic, which restricted social contact, Rizzo did not measure efforts to help. He was always there, through the most diverse available online channels of communication. He allowed this paper to be the product of my own effort and steered me in the correct path whenever he thought I needed it.

I would also like to thank my first IBM manager, Mr Colm Farrell, and the team leader, Mr David McDonagh, for recognising the potential of my previous experiences and hiring me when I was still living in Brazil. Since I started working for IBM Ireland, nearly to five years ago, they always incentivised me to go further, including the support to start this masters course. I must also thank Mr Anthony Kelly, my current manager, for all the support during my last academic year. Teammates like them make me prouder to be an IBMer every day.

Finally, I must express my very profound gratitude to my parents, brother, and my partner, Magdalena Kajinić, for providing me with unfailing support and continuous encouragement throughout my years of study and within the process of researching and writing this dissertation. This accomplishment would not have been possible without them. Thank you all.

The author, Bruno Ambrozio.

Contents

Declaration **I**

Abstract **II**

Acknowledgments **III**

Contents **IV**

List of Figures **VII**

List of Tables **VIII**

List of Equations **IX**

List of Acronyms **X**

1 Introduction **2**

1.1 Background 5

1.1.1 What is Genome-Wide Association Study 5

1.1.2 Feature Selection Vs. GWAS 8

1.1.3 Feature Extraction for GWAS 8

1.1.4 What is LightGBM 9

1.2 Research Problem 10

1.3 Research Objectives 12

1.4 Research Methodologies 12

1.5 Scope and Limitations 13

1.6 Document Outline 14

2 Literature Review and Related Work **15**

2.1 The State-of-the-art for GWAS 15

2.1.1 Linear Mixed Model 16

2.1.2 Scalable and Accurate Implementation of Generalized Mixed Model 17

2.1.3 General Linear Model 17

2.2 LightGBM Inference 19

2.3 K-Fold Cross-Validation 22

2.3.1 Cross-Validation for Hyperparameters Tuning 24

2.3.2 Cross-Validation for Model Selection 25

2.4 Bootstrap to Find Confidence Intervals 26

2.5 Test of Normality 28

2.6 Power Transformation (the Box-Cox) 29

2.7 Test of Paired Mean Differences 30

2.7.1 Dependent Student's T-test 30

2.7.2 Wilcoxon Signed-Rank Test 31

3 Experiment Design and Methodology **32**

3.1 Design Context 34

3.2 Hypotheses 37

3.3 Participants 38

3.4 Datasets and Variables of Interest 39

3.5 Procedure 41

4 Results, Evaluation and Discussion **48**

4.1 Results and Evaluation 49

4.1.1 Dataset ds1_1: Normality Test 49

4.1.2 Dataset ds1_1: Mean/Median Test 51

4.1.3 Dataset ds1_1: Discovery of Causal-SNPs 52

4.1.4 Dataset ds1_10: Normality Test 52
4.1.5 Dataset ds1_10: Mean/Median Test 54
4.1.6 Dataset ds1_10: Discovery of Causal-SNPs 55
4.1.7 Dataset ds1_100: Normality Test 55
4.1.8 Dataset ds1_100: Mean/Median Test 57
4.1.9 Dataset ds1_100: Discovery of Causal-SNPs 58
4.2 Discussion 59

5 Conclusion **62**
5.1 Research Overview 63
5.2 Problem Definition 64
5.3 Design/Experiments, Evaluation & Results 64
5.4 Contributions and Impact 68
5.5 Future Work 69

References **70**

Appendices **79**

A Suplementary Material **80**

List of Figures

1.1 Manhattan plot. 3
1.2 Annotated Manhattan plot, depicting a GWAS analysis. 7
1.3 LightGBM Leaf-wise tree growth. 9
1.4 State-of-the-art for GWAS Vs. LightGWAS. 12
1.5 The research onion. 13

2.1 Exclusive feature bundling pseudocode. 21
2.2 K-fold Cross-Validation. 23

3.1 Design and implementation workflow diagram. 33

4.1 ds1_1 histograms. 49
4.2 ds1_10 histograms. 52
4.3 ds1_100 histograms. 56

A.1 ds1_1 5k bootstraps: pairwise KDE relationships. 81
A.2 ds1_10 5k bootstraps: pairwise KDE relationships. 82
A.3 ds1_100 5k bootstraps: pairwise KDE relationships. 83

List of Tables

3.1 Research objectives and experiments/tasks. 35
3.2 Cohorts' phenotype distribution. 38
3.3 Phenotype ratios for genetic dataset build-up. 39
3.4 Variables of interest in the *.fam files. 40
3.5 Variables of interest in the *.bim files. 40
3.6 Variables of interest in the *.raw files. 41
3.7 LightGBM's hyperparameters selected through 5-folds cross-validation. 43
3.8 Parameters for the logistic regression common classifier. 44

A.1 LightGBM hyperparameters' ranges for the 5-fold CV. 80
A.2 50-folds CV: Raw results from dataset ds1_1. 84
A.3 50-folds CV: Transformed (Box-Cox) raw results from dataset ds1_1. . . 85
A.4 50-folds CV: Raw results from dataset ds1_10. 86
A.5 50-folds CV: Transformed (Box-Cox) raw results from dataset ds1_10. . 87
A.6 50-folds CV: Raw results from dataset ds1_100. 88
A.7 50-folds CV: Transformed (Box-Cox) raw results from dataset ds1_100. 89

List of Equations

2.1 General linear model. 17
2.2 Sparse Learning Based Method. 18
2.3 Logistic regression. 18
2.4 Penalized Likelihood Regression. 18
2.5 Logistic regression with Firth penalization. 18
2.6 Gradient Boosted Decision Trees model from LightGBM. 19
2.7 Loss function minimization in LightGBM. 19
2.8 Variance gain over split subsets in LightGBM. 20
2.9 95% CI representation. 26
2.10 Number of bootstraps equation. 27
2.11 Box-Cox transformation. 30
2.12 Paired t-test. 31
2.13 Wilcoxon Matched Pairs Signed Rank Test. 31

3.1 Metrics to evaluate the GWAS models applied to binary phenotype. . . 36

List of Acronyms

AI artificial intelligence. 22

aka also known as. 5

APS average precision score. II, 37, 47, 50–52, 54–58, 60, 61, 67, 68

AUC area under the curve. II, 37, 47, 50–52, 54–57, 59–61, 67, 68

CI confidence interval. 26, 27, 46–49, 52, 53, 55, 56, 58, 59, 67, 69

CV cross-validation. II, 4, 11, 12, 16, 23–27, 35, 36, 43, 46–49, 60, 62–64, 66, 67, 69, 70

DNA deoxyribonucleic acid. 2, 3, 5–7, 10, 11, 35, 41, 65

DT decision tree. 4, 9, 66, 70

EFB exclusive feature bundling. VII, 4, 9, 20, 21, 35, 64

EHR electronic health records. 35, 39

F1 weighted average of the precision and recall. II, 37, 47, 50, 52, 54–58, 61, 67, 68

FPR false-positive rate. 10, 37–39

GBDT gradient boosted decision trees. 4, 9–11, 14, 20, 35, 43, 46, 64, 66, 69

GBM gradient boosting machine. 4, 9, 64, 66, 69

GLM general linear model. I, II, V, 3, 16, 18, 19, 35, 36, 40, 43, 49, 64, 66, 70

GOSS gradient-based one-side sampling. 9, 20, 35, 64

GPU graphics processing unit. 69

GWAS genome-wide association study. I, II, 2–8, 10–19, 33, 35–41, 43, 45, 47, 60–70

IDS intrusion detection system. 22

KDE kernel density estimation. 53, 56, 59

L2 ridge regularisation. 18, 45

LASSO least absolute shrinkage and selection operator (aka l1). 18

LD linkage disequilibrium. 11

LGBM light gradient boosting machine (aka lightgbm). 43

LL lower limit. 26, 28, 47, 53, 56, 59

LMM linear mixed model. V, 3, 16, 17, 70

LR logistic regression. I, VIII, 16, 18, 19, 40, 43, 45, 64, 67

MAC minor allele count. 16, 43

MAF minor allele frequency. 6, 7, 11, 40, 65

MD mean absolute difference. 52

ML machine learning. II, 4–6, 12, 15, 24–27, 41, 46, 63, 64, 67, 69

PC principal component. 3, 4, 8, 17, 65

PCA principal component analysis. 3, 8, 17, 65

QC quality control. II, 4, 5, 9, 11, 35, 65, 69

ROC receiver operating characteristic. II, 37, 47, 50–52, 54–57, 59–61, 67, 68

SAIGE scalable and accurate implementation of generalized mixed model. V, 3, 16–18, 70

SNP single-nucleotide polymorphism. 2–8, 10, 11, 16, 17, 35, 36, 39–43, 45–47, 53, 56, 59–61, 64–68, 70

SPA saddlepoint approximation. 17, 18

TPR true-positive rate. 37

TSV tab-separated values. 41, 42

UL upper limit. 26, 28, 47, 53, 56, 59

VCF variant call format. 40, 42

Chapter 1

Introduction

Living organisms of the same species are distinguished from each other by their deoxyribonucleic acid (DNA). The most common type of genetic variant among humans is the single-nucleotide polymorphism (SNP) (Sebastiani et al., 2009). SNPs are responsible for physical trait differences. For example, the SNP $rs12913832$, causes a phenotype (trait) change from brown to blue eyes, respectively (White & Rabago-Smith, 2010). Besides mutations, SNPs also account for many of the known complex diseases, such as Type-2 diabetes or Coronary Heart Disease (Mills & Rahal, 2019). Whenever a SNP is responsible for a phenotype, it is denominated as a causal-SNP. Therefore, identifying causal-SNPs is an effective way to understand, prevent, or treat illnesses.

There are many methods to identify causal-SNPs, including genome-wide association study (GWAS). GWAS is an essential technique due to its achievements since the completion of the human genome sequence in 2003 (Mills & Rahal, 2019). For example, it contributed to the identification of the age-related macular degeneration (AMD) gene, called Factor H (Bush & Moore, 2012), an eye disease that may become worse over time. According to Pearson (2008), and Mills & Rahal (2019), about $3,700$ GWASs contributed to discovering thousands of genetic risk causal-SNPs and their biological function over the last decade.

GWAS is roughly analogue to, or a type of feature selection: each SNP is a feature (independent variable), and the phenotype is the class (target, or outcome variable).

It also extracts principal components (PCs) (typically from principal component analysis (PCA)) as part of its procedure, and use them as covariants for the underlying association model (Price et al., 2006). There are many GWAS implementations (Bush & Moore, 2012; Jiang et al., 2019). The state-of-the-art is based on the following statistical association models: general linear model (GLM), linear mixed model (LMM), and scalable and accurate implementation of generalized mixed model (SAIGE) (Loh et al., 2018). Their applicability depends on the phenotype type, sample size, and class distribution across the genomic dataset. However, all of them share the same primarily goal: to calculate the probability of association between each SNP and the underlying phenotype (evaluated through a p-value score interpretation). Consequently, they disclose which SNPs are likely to cause the investigated trait. A commonly used data visualisation in GWAS is the Manhattan Plot. Such a chart is a derivation of scatter plots, which depicts the p-values in $\log_{10}$ scale, highlighting the potential causal-SNP. Figure 1.1 exemplifies it.

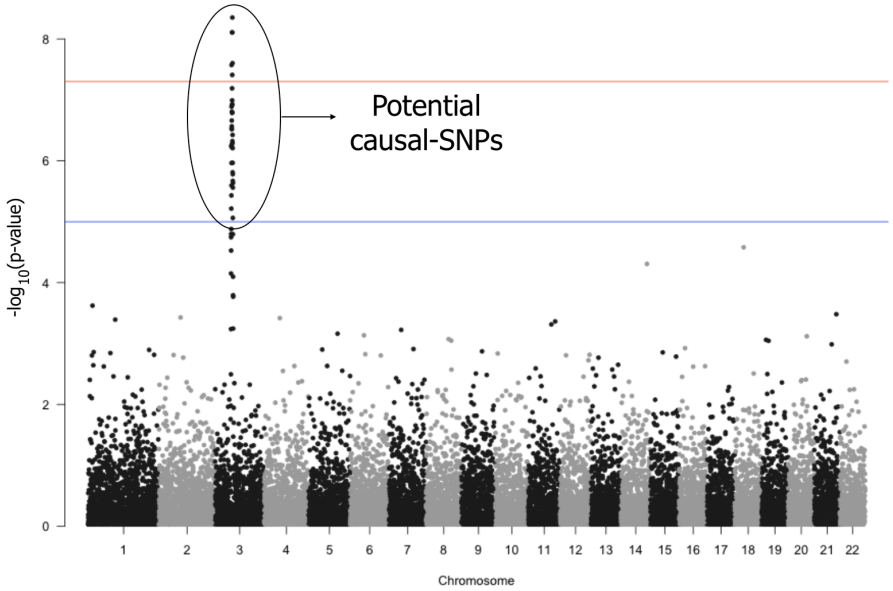

Figure 1.1: Manhattan Plot. Each dot is a SNP.

Although GWAS has proven to be an efficient method to discover causal-SNPs, the state-of-the-art begins to face inevitable bottlenecks that demand evolvements to stick with its relevance for such an end. For example, the costs with DNA sequencing

have reduced over the years, causing an exponential growth of the underlying genomic datasets (Pérez-Enciso & Zingaretti, 2019). Such a fact created a computational-cost increase. On top of that, genomic datasets have morphological aspects that make GWAS more challenging. The expansion of SNP's dataset contributes to overwhelming the data sparsity, with millions of SNPs (variables), and a few patients (samples) (Lubke et al., 2013). Another point of concern is the imbalanced distribution between rare cases and several controls (Zhou et al., 2018). Such challenges rely on cumbersome manual approaches, performed by analysts through several quality control (QC) steps, in an attempt to reduce the false-positives (Bush & Moore, 2012) caused by such scenarios.

Machine learning (ML) has non-parametric algorithms that, potentially, address GWAS issues better than the current state-of-the-art. They can self-adapt to different data structures by adding bias and controlling variance over the training process (Schratz et al., 2019). Consequently, it would improve precision, and increase statistical power, independently of human intervention. Therefore, this dissertation proposes a novel procedure for GWAS. It is assembled over decision tree (DT) bosted by gradient boosting machine (GBM), whose implementation comes from the LightGBM framework (Ke et al., 2017). It also employes k-fold cross-validation (CV) for hyperparameter selection, ensuring adaptability to the most diverse genomic data structures. Such a proposed method has been named LightGWAS.

LightGWAS conducts the SNP and phenotype association through the gradient boosted decision trees (GBDT) implementation available in the LightGBM framework. It replaces the need for covariants made of PCs when using one of the GWAS state-of-the-art models because it has an inbuilt feature-extraction algorithm, called exclusive feature bundling (EFB). EFB reduces the data dimension, diminishing the sparsity of the genomic dataset. Also, LightGWAS employes CV to obtain the optimal model's parameter (hyperparameters), which ensures it will self-adapt for different data structures. Once the hyperparameters are found, the GBDT is trained with a relevant portion of the genomic dataset. Subsequently, the model is fitted, and the importance of the features are scored. The sub-set of features with higher importance

scores are actually the set of potential causal-SNPs, just identified by the outlined procedure.

Besides an innovative alternative to GWAS implementations, LightGWAS also aims to be a scalable solution by reducing human intervention over the analysis, allowing the process to be fully automated through a computational pipeline if desired. According to Bush & Moore (2012), SNPs datasets tend to be replaced by whole-genome sequencing data (due to the earlier mentioned cost reduction with DNA sequencing). It means the datasets will be composed by billions of nucleotides (structural units of DNA) instead of only the SNPs (each SNP, or in this case, nucleotides, is a column in the genomic dataset), turning manual QC unfeasible. Therefore, a method capable of handling genomic's big data adaptively is vital.

1.1 Background

The following section covers in details what GWAS and LightGBM is. It makes a connection between the available GWAS methods with feature selection and feature extraction techniques. Moreover, it also explains LightGBM, the ML framework employed in the proposed novel procedure for GWAS, the LightGWAS. Such points are the baseline of this study. Accordingly, sufficient perception of them is essential to the progress and understanding of this dissertation.

1.1.1 What is Genome-Wide Association Study

GWAS is a hypothesis-free (or discovery-driven research) investigative technique to catalogue SNPs across populations and to identify genetic markers associated with a trait in a genetic region (*locus*) or on multiple regions (*loci*) (Bush & Moore, 2012; Farrell, 2017). When applied to human populations, GWAS aims to identify SNPs associated with one or more phenotypes (also known as (aka) causal-SNPs).

A phenotype is any observable characteristic or trait in a cohort (Hill et al., 2017). It can be a qualitative (binary) phenotype, such as eye colour, curly hair, or most commonly for GWAS analysis, a disease status (e.g., affected or not by Type-2 dia-

betes, COVID-19, coronary heart disease, and among others). Alternatively, it can be a quantitative (numeric) trait, for example, people's height, weight, body mass index, blood pressure, and so on.

SNP, on the other hand, is the name given to the genetic variation. They are single base-pair changes (*alleles*) in the DNA sequence that occur with high minor allele frequency (MAF) in the human genome (Bush & Moore, 2012). The convention threshold to qualify a DNA base-pair as a SNP variates from study by study. Usually, the adopted outset is $MAF \geq 0.01$ or $MAF \geq 0.05$ (Fadista et al., 2016). The "A haplotype map of the human genome" (2005) project, for instance, employed $MAF \geq 0.05$. The human DNA is composed of two strands held together by bonds between the bases adenine (A) to thymine (T) (or vice-versa), and cytosine (C) to guanine (G) (or vice-versa). Whenever those bonds vary in a specific *locus*, which means they are *alleles* in the DNA sequence (e.g., expected A as the dominant population hold, but found C in the examined sample), such *allele* (variant) is labelled as an SNP [1].

GWAS involves the comparison of two cohorts, one containing the phenotype object of the study, and another that do not have the trait. In a GWAS framework, such groups are called *cases*, and *controls*, respectively. They are analogue to the class label in a ML classification model (e.g., $Y = y; y \in \{0, 1\}$).

GWAS investigates how significantly an SNP is associated with the trait in the *cases* cohort, against how insignificant, or perhaps even null, is the same SNP in the *controls* cohort. With that, it is possible to presume that the found correlated SNP are the underlying phenotype's causal-SNP. As introduced in the begging of this chapter, the probability of association is calculated through a given statistical model. The returned scores (p-values) contribute to infer the existence of an association between the variance and trait, whenever they are below a specific threshold of significance ($p < \alpha$). The convention for GWAS is a threshold of $\alpha \approx 5 \times 10^{-8}$ (Fadista et al., 2016; Mills & Rahal, 2019). Therefore, an SNP is told to be associated with a phenotype whenever its p-value is lower than the predefined α. In other words, an SNP

[1] The definitions presented above are the minimum and high−level information needed to interpret this computing science material. Further biological details are beyond its scope.

is statistically significantly correlated with a phenotype whether $p \leq \alpha | \alpha \approx 5 \times 10^{-8}$, thus such an SNP is a causal-SNP.

GWAS is a crucial tool in the combat of diseases. It favours the development of drugs focused on a particular genetic aspect. Sometimes it even discovers that an existent drug assists in treating a new disease, just because the investigated disease has its causal-SNP associated with the same phenotype (illness, in this case) which already have a treatment developed. It implies that GWAS offers excellent potential to both help identify new therapeutic targets, and support the stratification of patients who would gain the most significant benefit from specific, and already existent, drug classes (Hill et al., 2017).

Figure 1.2 depicts the content approached in this section. It shows how a base-pair of DNA is labelled as SNPs by giving a MAF threshold, the cohorts division, the association analysis of SNPs with higher frequency in cases cohort than in controls, and the identification of potential causal-SNPs given a α cut-off.

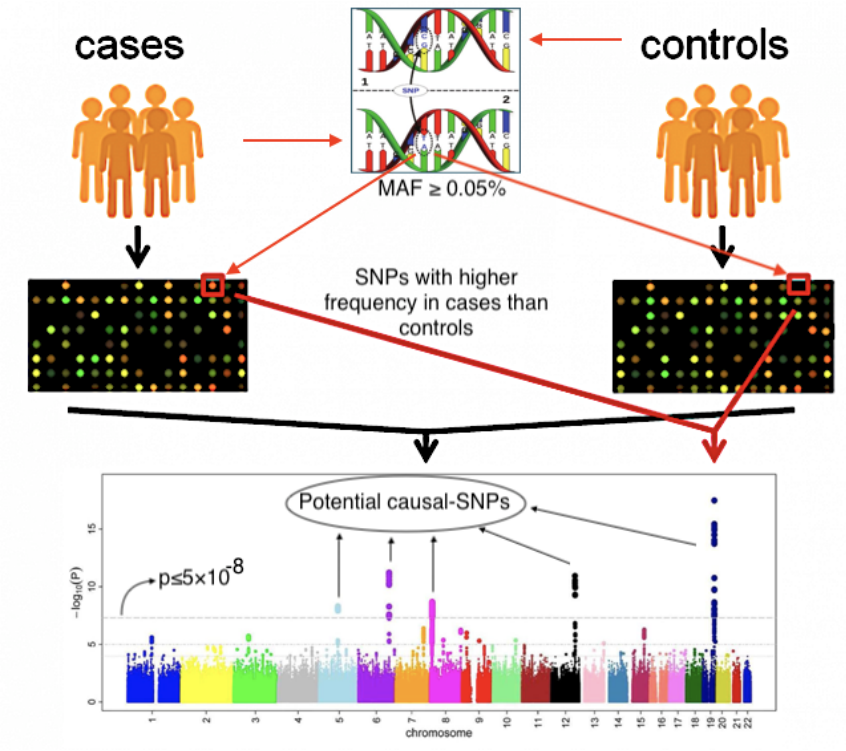

Figure 1.2: Annotated Manhattan plot, depicting a GWAS analysis. Adapted from EMBL-EBI, European Bioinformatics Institute (2020), and Sukhumsirichart (2018).

1.1.2 Feature Selection Vs. GWAS

GWAS process has similarities with feature selection techniques: Let each SNP be an independent variable, and the phenotype the dependent variable (target, or class) in a dataset. GWAS will select the features that could be seen as causal-SNPs since they can better predict the class variable (phenotypes).

Feature selection is a well-known pre-processing step of most of the ML or statistical regression models. It encompasses the process of deciding what are the relevant variables from a dataset, in terms of outcome class explanation (e.g., features abler to predict a target variable). In a classification model, for example, it aims to find the minimum set of non-redundant features that better predicts one or more outcome variables (Shah & Kusiak, 2004).

There are many different approaches and algorithms grounding feature selection (Guo et al., 2002). They can be univariate based methods, like chi-squared (χ^2) score (Liu & Setiono, 1995), and Fisher Score (Duda et al., 2012), or regression model-based, such as linear model penalized, and tree-based feature selection (Shah & Kusiak, 2004).

The first implementations of GWAS have been designed over univariate based feature selection methods. Examples include the popular PLINK1.7 tool (Purcell et al., 2007), MACH2qtl/dat, SNPTEST, ProbABEL, Beagle, BIMBAM, SNPMStat, and others (Pei et al., 2010). Still, the latest versions such as PLINK2 (Hill et al., 2017), Fast-GWAS (Yang et al., 2011), and others are grounded on regression model-based.

1.1.3 Feature Extraction for GWAS

The term "feature extraction" sometimes is confused with "feature selection". Although the terms are sometimes used interchangeably, they do not represent the same thing. As outlined by Li et al. (2018), feature extraction accounts for dimension-reduction, which means transforming a set of high-dimensional features into a small set of new low-dimensional variables. An example of feature extraction is PCA. In GWAS procedures, PC's extraction is one of the manual steps encompassed by the

underlying QC process. They are used as covariates for the association model. Another example is EFB (Ke et al., 2017), the embedded feature extraction algorithm of the LightGBM framework. It bundles mutually exclusive features to reduce dimensionality, throughout an algorithm with a continuous approximation ratio.

1.1.4 What is LightGBM

LightGBM is a GBM framework, that implements a GBDT algorithm. It applies histogram-based algorithms to find the best split point of a tree. The information gain is estimated through what Ke et al. (2017) have called gradient-based one-side sampling (GOSS). GOSS is an algorithm that downsamples data instances, focusing on the accuracy of information gain estimation. It randomly drops instances with small gradients, yet, retaining the native data distribution.

LightGBM additionally encompasses a feature extraction technique called EFB, also designed by Ke et al. (2017). According to the authors, on large and sparse data structures, many features are regularly exclusive. Hence, EFB identify such variables and safely bundle them together. It fits a Graph Coloring algorithm-based and regular approximation ratio for the dimension reduction: Features become vertices, and for every two of them, non-mutually exclusive, edges are attached.

Three main characteristics distinguishing LightGBM's implementation from the other GBDT algorithms: (1) It grows trees leaf-wise instead of depth-wise as most of the DT algorithms. Figure 1.3 depicts it. (2) It uses GOSS to reduce the histogram building coast by sub-sampling the data without much intervention in the actual distribution, and (3) It employes EFB as feature extraction for dimension reduction, which in turn reduces the computational cost of finding the best split-points of the trees.

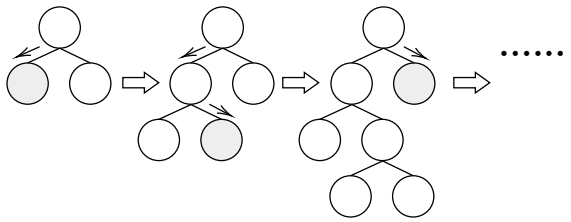

Figure 1.3: LightGBM Leaf-wise tree growth.

LightGBM has been created to address the main bottleneck of conventional GBDT algorithms: The accuracy, efficiency, and scalability when the feature dimension is high, and data size is large (Ke et al., 2017).

1.2 Research Problem

With the continuous costs reduction for DNA sequencing over the years, the genomic datasets have grown exponentially (Pérez-Enciso & Zingaretti, 2019), which in turn created a computational-cost increase for GWAS. Also, genomic datasets often have aspects that turn GWAS harder to correlate phenotypes with causal-SNPs, such as high-sparse data due to millions of SNPs, with a few samples (Lubke et al., 2013).

Another point of concern with GWAS implementations is dealing with imbalanced data. It is a quite common scenario for GWAS to be applied on datasets composed of many cases, and a few controls. Whenever the case-control ratio are imbalanced ($case : control = 1 : 10$) or high-imbalanced ($case : control = 1 : 100$) (Zhou et al., 2018), current state-state-of-the-art GWAS algorithms often introduces bias that inflates the false-positive rate (FPR), implying on statistical power reduction (Sebastiani et al., 2009; Lee et al., 2010; Price et al., 2006; Reed et al., 2015; Spencer et al., 2009). Such a scenario causes Type 1 error (it rejects the hypothesis that states no association when de facto there is none). To address such an issue, the analysts usually employ some imputation method, to either rebalance the samples and increase statistical power. However, this is another manual step, that is part of the QC process. It relies on professional skills, which in turn, opens margins for human mistakes. Imputation methods also increase computational coast and often inflates false-positives (Spencer et al., 2009). The data comes from external sources, such as the 1,000 Genomes project (Auton et al., 2015), which in turn, is beyond the cases and controls pre-filtered cohorts.

Statistical power is also affected by a high degree of homogeneity (e.g., when MAF is too low (e.g., $MAF < 1\%$) (Reed et al., 2015) at SNPs across records. Alternatively, when within a given population, the alleles are more correlated than would be expected

whether by chance, meaning the data is in linkage disequilibrium (LD) (Grinberg et al., 2019), which also implies on type 1 error.

Population stratification is another point of concern in GWAS. For example, allele frequency differences between cases and controls due to systematic ancestry differences can cause spurious associations with traits (Price et al., 2006).

And last but not least, cloud computing has enabled affordable hardware access so that working with a whole-genome through the entire DNA sequence will soon become a reality. It means that manual QC steps will no longer be possible (Bush & Moore, 2012), forcing GWAS to evolve towards autonomous systems.

For all of those reasons, current state-of-the-art for GWAS has struggled to ensure acceptable precision on causal-SNPs selection. They depend on meticulous manual QC steps, which are compromised by data scale, human mistakes, and lack of automation.

As can be perceived, GWAS shares from most of the same challenges as the GBDT implementations, now addressed by LightGBM framework. Such a perception has motivated to employ LightGBM as a potential solution for the aforementioned problems. Therefore, this dissertation proposes the LightGWAS, a new procedure for GWAS, based on LightGBM and CV. By this mean, it aims to answer the following research question:

- *Can LightGWAS be an alternative method to the state-of-the-art for genome-wide association studies, by increasing statistical power on causal-SNP detection, and reduction of manual quality control steps?*

The figure 1.4 illustrates the components' differences between the current state-of-the-art for GWAS and LightGWAS. Each independent box represents a human intervention. The boxes "GLM", "LMM", and "SAIGE" are mutually exclusive. The analyst must decide which one to apply, depending on the underlying data structure. LightGWAS aims to be a self-contained and autonomous ML framework for GWAS.

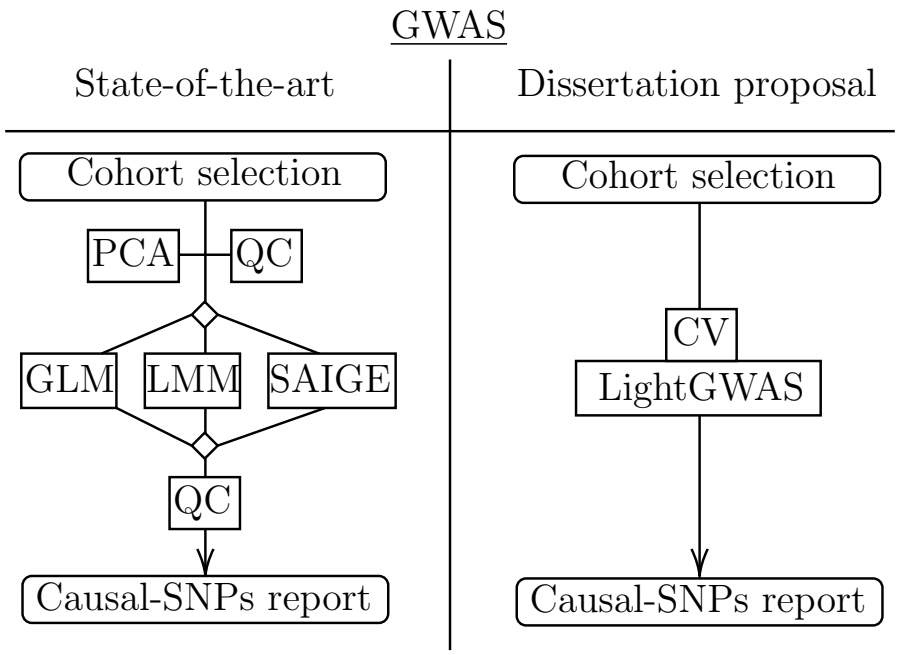

Figure 1.4: State-of-the-art for GWAS Vs. LightGWAS.

1.3 Research Objectives

This work is aimed at initiating what may become a new GWAS framework. It endeavours to assemble a novel GWAS procedure based on LightGBM and CV. As previously stated, such a proof-of-concept has been named LightGWAS. Subsequently, to validate its relevance, LightGWAS has been compared with one of the current state-of-the-art implementations. Therefore, two objectives are defined:

1. To evaluate whether LightGWAS is a suitable GWAS method.

2. To assess if LightGWAS outperforms the compared state-of-the-art method.

1.4 Research Methodologies

This dissertation has been built through a secondary research method. Quantitative empirical experiments, examined through deductive reasoning, have been proposed in order to achieve the previously defined objectives. Hence, according to the "research

onion methodology" (Saunders et al., 2009) (see figure 1.5 below), this study could be classified as follows:

- Philosophical stance: Positivism. The research encompasses a research question with testable hypotheses.
- Approach: Deductive. The study has been driven by the search for the answer to the research question.
- Strategy: Experiment. A set of reproducible technical tests have been employed to reach the established objectives of this research.
- Choice: Mono method. The research works with quantitative data only.
- Time horizon: Cross-sectional. The utilized data originated from a specific group of individuals, at a single point in time.
- Technique and procedure: Data collection and data analysis - The research involves genomic datasets and statistical tests.

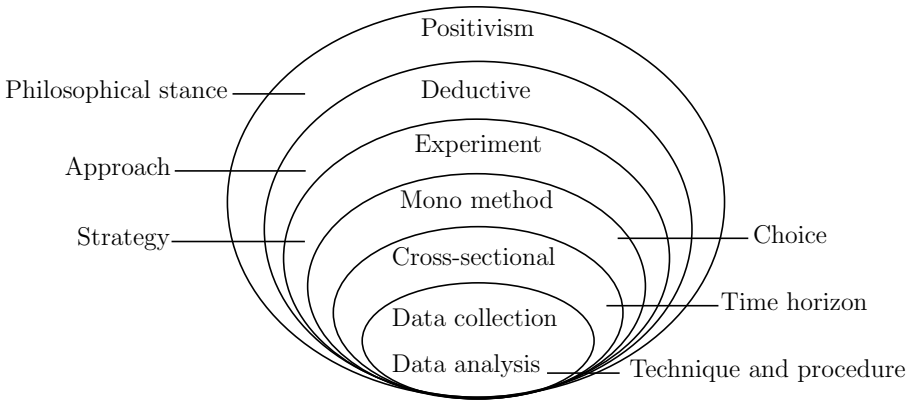

Figure 1.5: The research onion. Adapted from Saunders et al. (2009) diagram. The image displays only the employed items from each layer of the research onion diagram.

1.5 Scope and Limitations

The present research conducts a literature review about the state-of-the-art for GWAS. It provides an understanding of the current and likely future limitations surrounding

the available methods. From that point, LightGWAS has been designed, implemented, and assessed whether a potential solution to address the identified gaps. Given the time limitation to conclude this research, the experiments have been restricted to synthetic datasets, over three different genomic data structures (balanced, imbalanced, and high-imbalanced class) for qualitative phenotypes only. The study also embraces a comparison of the proposed LightGWAS with one of the GWAS's implementations available in the market. Such an implementation is an instance of the considered state-of-the-art for the data structures and phenotype category above-mentioned.

1.6 Document Outline

Chapter 2 brings in details the state-of-the-art for GWAS. It explores each of the leading methods available to address each of the known GWAS's scenarios. It also explains how LightGBM works, the GBDT implementation employed in the LightGWAS procedure, and how similar problems across other research areas have also been addressed through such a ML framework. Definitions surrounding the statistical techniques to either assemble LightGWAS and test it against the state-of-the-art is also given over this chapter.

In chapter 3, the design and methodology applied is detailed as much as the hypotheses aimed to be tested. Details about how LightWAS has been developed and applied are disclosed. It also explains how the involved datasets and models were prepared, fitted, tested, compared, and evaluated.

Chapter 4, in turn, carries over the outcomes from the experiments. Three main scenarios have been explored; therefore, such a section is composed of three result sets, followed by their underlying evaluations and discussions.

Chapter 5 concludes the dissertation. Firstly, an overview concerning the entire research is given. Secondly, a summary of the identified problems with GWAS's state-of-the-art is discussed, followed by the evaluation of the proposed design to address them. Thirdly the contributions and impact of this work are reviewed, and lastly, a list of suggested future studies is recommended.

Chapter 2

Literature Review and Related Work

In this chapter, the literature review grounding this dissertation is presented. Each of the methods that compose the current state-of-the-art for GWAS is examined. LightGBM framework and k-fold CV are also explored. They are the core of the proposed method for GWAS, the LightGWAS. Moreover, every statistical approach to execute the experiments have been outlined. Their definitions are elementary to base the decisions taken upon the dissertation.

2.1 The State-of-the-art for GWAS

Three main methods compose the GWAS's state-of-the-art: GLM with Firth support (Ma et al., 2013), LMM (Loh et al., 2018), and SAIGE (Zhou et al., 2018). According to Loh et al. (2018), the applicability of these methods should consider the following criteria: (a) LMM (e.g., BOLT-LMM implementation) for datasets bigger than five thousand samples and quantitative traits type. Whether qualitative phenotype, the dataset should be in a normal distribution; otherwise, the SNP's probability scores can become miscalibrated. (b) GLM (e.g., GCTA-fastGWAS (Jiang et al., 2019) or Plink (version $>$ 1.9) implementations (Hill et al., 2017)) for quantitative traits, up to five thousand samples. When qualitative phenotypes, the logistic regression (LR)

implementation should include Firth regularisation support for cases where minor allele count (MAC)< 400; otherwise the results may suffer type 1 error (caused by false-positive inflation), and consequently, statistical power reduction (Ma et al., 2013). (c) SAIGE for high-imbalanced *case* : *control* ratio of qualitative traits.

For all the cases, PCs (from PCA, for instance) should be utilised as covariants in the association model (Price et al., 2006). The convention is to use the first ten eigenvalues when using PCA (Price et al., 2006; Chen & Ishwaran, 2012). The following subsections will detail what is and how each of the listed GWAS methods work.

2.1.1 Linear Mixed Model

LMM is an algorithm derivated from the traditional linear models. In GWAS context, it allows the fusion of either fixed effect and random effect to estimate the correlation between classes (phenotypes) and features (SNPs). According to Fitzmaurice & Laird (2015), mixed model-based algorithms offers flexibility over the correlation analysis, outstanding its parent linear model, mainly for unbalanced regression studies. The BOLT-LMM, for example, is a GWAS method based on mixed models. According to its creators Loh et al. (2018), BOLT-LMM differentiates from its predecessors by assuming a Bayesian mixture-of-normals prior to the random effect associated with the SNPs. It infers the standard "infinitesimal" mixed model employed by previous mixed-model association methods. So that it increases power while controlling for false-positives. The authors also demonstrated that BOLT-LMM is faster than eigendecomposition-based methods (eigenvalues), either when using the Bayesian mixture model or specialised to LMM association. BOLT-LMM is state-of-the-art for GWAS across datasets bigger than five thousand samples of quantitative trait type. It can also be used against qualitative traits, as long as the distribution is Gaussian (Loh et al., 2018).

2.1.2 Scalable and Accurate Implementation of Generalized Mixed Model

SAIGE is the state-of-the-art for GWAS for imbalanced genomic datasets of qualitative phenotypes. It has been developed over LMM along with saddlepoint approximation (SPA) to score the association probabilities. According to Zhou et al. (2018), SAIGE's authors, SPA can calibrate imbalanced case-control proportions across association tests, which accounts for statistical power increment. It addresses false-positives and reduces type-1 errors. The authors have provided statistical evidence that SAIGE results in accurate probabilities even when case-control ratios are extremely imbalanced ($case : control \leq 1 : 100$).

2.1.3 General Linear Model

GLM grounds many of the known statistical tests (Urso et al., 2019), such as ANOVA, logistic regression, and linear regression. They all derivate from the same structure as $Data = Model + Error$. GLM can be represented as follow:

General linear model.

$$\hat{Y} = \beta_0 + \beta_1 X \tag{2.1}$$

where $\hat{Y}$ is the class (dependent variable), β_0 is the constant intercept, β_1 is the slope or weight that get stimulated to fit the model, and X is an independent (feature) variable.

The LR algorithm with Firth, for instance, when applied for GWAS, is a category of Sparse-Learning-Based feature selection methods. According to Guo et al. (2002), such a method category aims to *"minimize the fitting errors along with some sparse regularisation terms"*, such as least absolute shrinkage and selection operator (aka l1) (LASSO), ridge regularisation (L2), or Firth (Heinze, 2006) for rare feature variance. Equation 2.2 is the mathematical representation of sparse learning-based methods.

Sparse Learning Based Method.

$$\begin{aligned} \|W\|_p &= \left(\sum_{i=1}^{d} \|W\|^P\right)^{\frac{1}{P}} \\ min_w &= loss(w; X, y) + \alpha\|W\|_p \end{aligned} \tag{2.2}$$

Where $\|W\|_p$ is a sparse regularisation term, $loss(\bullet)$ is a loss function, such as logistic loss, and α is a regularisation parameter to balance the contribution of the loss function, and also the sparse regularisation term for feature selection.

LR applicability, as explained in Heinze (2006), can be observed in equation 2.3:

Logistic regression.

$$Pr(Y = 1) = \pi = \left[1 + e^{(-X\beta)}\right]^{-1} \tag{2.3}$$

Where $e^{(\beta)} = \frac{Pr(Y=1|X=x_0+1)/Pr(Y=0|X=x_0+1)}{Pr(Y=1|X=x_0)/Pr(Y=0|X=x_0)}$
Likelihood: $L(\beta|X) = \prod_{i=1}^{n} \hat{\pi}_i^{y_i}(1 - \hat{\pi}_i)^{1-y_i}$. (See equation 2.4 below)

Penalized Likelihood Regression.

$$\log L^*(\beta) = \log L(\beta) + A(\beta) \tag{2.4}$$

where $A(\beta)$ imposes prior on model coefficients, such as Firth-type, Jeffereys prior (FIRTH, 1993).
Firth type penalization: $A(\beta) = \frac{1}{2}\log\det(I(\beta))$.

With that, the equation 2.4 represents Firth-Jeffreys plugged for invariant before the likelihood regression equation:

Logistic regression with Firth penalization.

$$L^*(\theta) = L(\theta)\det(I(\theta))^{\frac{1}{2}} \tag{2.5}$$

Where $I(\theta)$ is the Fisher information matrix.

PLINK2 GLM[1] is one of the state-of-the-art for GWAS implementations that apply GLM with Firth regularization as above explained. Hence, it has been selected to compare with the LightGWAS over this study.

2.2 LightGBM Inference

LightGBM is a GBDT framework based on histogram algorithms that grows decision trees leaf-wise, and uses GOSS, and EFB as outlined below. It was designed three years ago by Ke et al. (2017), aiming to offer a highly efficient GBDT, in terms of accuracy, computational resources (such as reduction of memory consumption), and faster training for data in large-scale.

GBDT Histogram based algorithms costs $\mathcal{O}(\#data \cdot \#feature)$ for histogram building and $\mathcal{O}(\#bin \cdot \#feature)$ for split point finding (see equation 2.6). Computational complexity became higher as $bin \ll data$. LightGBM addresses it by downsampling data and reducing feature dimension with GOSS and EFB, respectively.

Gradient Boosted Decision Trees model from LightGBM.

$$F(x; w) = \sum_{t=0}^{T} \alpha_1 h_t(x; w) \tag{2.6}$$

Where function $h_t(\bullet)$ represents the tth decision tree model, function $F(\bullet)$ denotes the predictive values of the GBDT model, x is the input samples, w is the parameter of the decision tree, and α is the weight of each tree.

By minimizing the loss function $L(\bullet)$ for mapping space x to space y, the optimal model is solved through the equation 2.7:

Loss function minimization in LightGBM.

$$F^* = \arg\min_{F} \sum_{i=0}^{N} L(y, F(x; w)) \tag{2.7}$$

[1] `https://www.cog-genomics.org/plink/2.0/assoc#glm`

GOSS retains large gradients samples, while samples with small gradient are randomly selected, given constant weights. With that, GOSS concentrates on undertrained samples without altering the distribution of raw data. The equation 2.8 defines the variance gain of splitting the instances over subsets A's and B's features.

Variance gain over split subsets in LightGBM.

$$\tilde{V}_j(d) = \frac{1}{2}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_r^j(d)}\right) \tag{2.8}$$

Where $A_l = \{x_i \in A : x_{ij} \leqslant d\}$, $A_r = \{x_i \in A : x_{ij} \ > d\}$, $B_l = \{x_i \in B : x_{ij} \leqslant d\}$, $B_r = \{x_i \in B : x_{ij} > d\}$, n is the dimension of the characteristic x, a, b, and d are constants. In each iteration of gradient boosting, the negative gradients of the loss function concerning the output of the GBDT model denote as g_i. Subset A consists of the $top_a \times 100\%$ samples with the large gradients. A_c represents the $(1-a) \times 100\%$ samples, and subset B is discretionarily selected with size $b \times |A_c|$ (Wang et al., 2019).

EFB, in turn, is a feature extraction technique, based on graph coloring problem, which also contributes to reducing the histogram building complexity. It deals with the sparsity of the data, where $\#bundle \ll \#feature$, by grouping many independent features to the dense features, avoiding unnecessary computation with features that do not account for the outcome variable (the variables with zero gain score). Therefore, the complexity $\mathcal{O}(\#data)$ becomes $\mathcal{O}(\#non\_zero\_data)$. The pseudocode in figure 2.1 represents an abstract implementation of EFB for LightGWAS.

Greedy Bundling

Input: F: features, K: max conflict count
Construct graph G
searchOrder $\leftarrow$ G.sortByDegree()
bundles $\leftarrow$ {}, bundlesConflict $\leftarrow$ {}
for i ***in*** *searchOrder* **do**
  needNew $\leftarrow$ True
  for $j = 1$ ***to*** *len(bundles)* **do**
    cnt $\leftarrow$ ConflictCnt(bundles[j],F[i])
    if *cnt + bundlesConflict[i]* $\leq K$ **then**
      bundles[j].add(F[i]), needNew $\leftarrow$ False
      break
  if *needNew* **then**
    Add $F[i]$ as a new bundle to *bundles*
Output: *bundles*

Merge Exclusive Features

Input: $numData$: number of data
Input: F: One bundle of exclusive features
binRanges $\leftarrow$ {0}, totalBin $\leftarrow$ 0
for f ***in*** F **do**
  totalBin += f.numBin
  binRanges.append(totalBin)
newBin $\leftarrow$ new Bin(numData)
for $i = 1$ ***to*** $numData$ **do**
  newBin[i] $\leftarrow$ 0
  for $j = 1$ ***to*** *len(F)* **do**
    if *F[j].bin[i]* $\neq$ *0* **then**
      newBin[i] $\leftarrow$ F[j].bin[i] + binRanges[j]
Output: $newBin$, $binRanges$

Figure 2.1: Exclusive feature bundling pseudocode. Created by Ke et al. (2017).

LightGBM is a young framework, thereupon, there are no many publications yet exploring its features. However, from the few published studies available, it's possible to see how efficient has been LightGBM inference over big datasets with sparse features in the most diverse fields of science. For example, Mo & Li (2019) have proposed an efficient model based on Auto-Encoder and LightGBM to classify network traffics, aiming to address some of the cybersecurity-related issues, such as the challenges with intrusion detection systems (IDSs) that continuously suffer from the network criminals renovating their attack means. The involved datasets are composed of many data sources, which causes sparse data with many features, and a vast amount of irrelevant dimensions that harm the available models' accuracy. The authors identified that LightGBM would be the right choice, given the morphology of their datasets. With LightGBM, they managed to address the feature selection by consuming low memory, and satisfactory accuracy ratio, through a fast training process.

In another paper, Wang et al. (2019) designed a transient stability assessment method based on LightGBM. The research addresses the challenges involving large-scale and high dimension of data in the involved datasets for artificial intelligence (AI) grounded on transient stability assessment systems. They adopted LightGBM mainly as a feature selector, that allowed the proposed model to work with only the relevant variables and dimensions. Such a model has applicability to reduce physical risks with

renewable power systems like wind and photovoltaic that face challenges regarding volatility, randomness, and low-inertia introduced by renewable generation resources.

Song et al. (2019), in turn, presents a regression model based on LightGBM to predict the probability of diseases such as cardiovascular and cerebrovascular, through double-high biochemical indicators. The dataset involves the user's physical examination information and five biochemical signs. The research demonstrated that the proposed model has higher stability and better generalization performance compared to other available approaches for the same end.

To conclude this section, Singh et al. (2020) have proposed a prediction model based on LightGBM to identify potential applicants who are likely to take admission in a university. The research reported 95% accuracy with the LightGBM model, while the other compared models reached 82.4% using logistics regression, and 86.5% with neural networks model.

2.3 K-Fold Cross-Validation

Cross-validation (CV) is a technique of model evaluation based on subsampling with no replacements (Shao, 1993). Folds in this context represents the equally (or approximately equal) size division of the dataset into k subsets that do not overlap. The disjoint into k folds is the product of a random selection, which means that the dataset is firstly shuffled to then be split. The model is trained with all the folds, but one $(k-1)$, which in turn is employed as a validation set. The left-out set is utilised to test the trained model and generate the metric to measure its effectiveness. The whole process is repeated until every fold had a chance to be out as a validation set. Therefore, when the CV concludes, a metric result set is generated composed of k score records. The mean of such a score set is the performance of the operated CV. Figure 2.2 illustrates a 10-folds CV process.

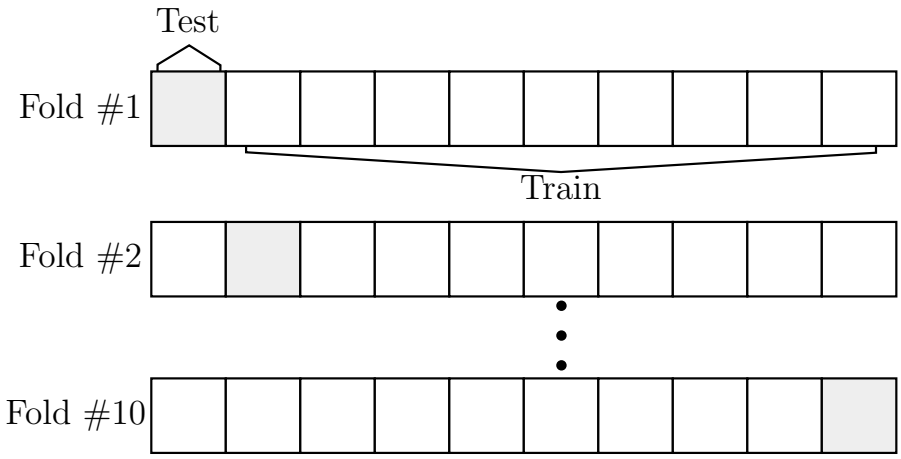

Figure 2.2: K-fold Cross-Validation. Adapted from Berrar (2019).

Given that $k = 10$, each iteration will have a train composed of 90% of the data and test with the 10% left behind. Over each fold, a different subset is separated to be the validation sample. At the end of the k iterations (10 folds), every subset has been once a test group and $k - 1$ times part of the training group.

An important point about fold sampling (sometimes is overlooked) is ensuring stratification of the outcome variable. Stratified random sampling is the process to assuring the test fold will have the same proportion of cases-controls of the model's class. Consider a dataset with 100 samples and a binary class variable. Let this class variable has 20 1s and 80 0s. If no stratification is employed during the sampling, there is a risk that the test set will contain samples of only one case (or control), which will result on inaccurate model results. Stratification sampling, in turn, will ensure the test sets will be composed of 20% of 1s for this example. In other words, the test sets will have 2 1s and 8 0s over each fold. Therefore, each subsample will comply with the whole class' dataset distribution.

Last but not least is the number of folds to be used. According to Kohavi (1995), stratified 10-fold CV is a reasonable fit for most of the real-world datasets. They demonstrated that as higher is the k, lower is the bias, and higher is the variance. $K = 10$ is the best fit for general purposes, as it saves 10% of the dataset for validation, given margin enough to calculate the average of the folds result sets. Kuhn & Johnson (2013) aggress with that. They also add that $k = 5$ may also be a good fit, leaving

20% of the data for validation over each fold. However, some other studies have contested it (Bengio & Grandvalet, 2003; Zhang & Yang, 2015), and claimed through empirical tests that the variation on the choice of k may depend on the underlying model and dataset structure. Bengio & Grandvalet (2003) explains that whether CV was employed to calculate the mean of independent estimates, $k = N$, where N is the total number of samples in the whole dataset (Aso known as Leave one out CV (LOOCV)) will reduce the variance significantly, and increase the bias. However, they have also proved through empirical tests that this is not true when the training set is highly correlated. Commonly 5-folds is applied for hyperparameter tunning, 10-folds for model's evaluation score (eg., Accuracy), and $k > 10$, such as LOOCV ($k = n$) for model's selection (eg., comparing multiple models to figure out which one has the best performance against a common dataset).

2.3.1 Cross-Validation for Hyperparameters Tuning

ML models depend on parameters that control their learning process. Such parameters are identified as hyperparameters. Hyperparameters are not learnt within the estimators; they are the arguments to trigger the learning process. Some models rely on dozens of parameters, whose which can receive infinity variation of information (e.g. (fraction of) numbers, ranges, categories, etc.), and be combined among each other for different ends. Hence, finding out the best combination with the best values of each parameter is a vital step to build a model that optimally generalises a specific problem. CV for hyperparameters tunning involves nested loops over the k-folds. The process to find the optimal parameters are costly from the computational point of view. Each parameter is cross-validated against each provided value (depending on the implementation, it may be a random sample instead all of them), usually iterated over nested loops across each fold, and validated in the hold-off validation subset (Cawley & Talbot, 2010). For example, in a k-folds scenario, each of the folds will be composed of another l-folds, where $l = k - 1$. The inner folds are cross-validated, and the average result is employed to determine the best combination of hyperparameters trained over that outer fold. The process repeats over each k-fold (outer loop),

which in turn, also averages the final result to find the set of parameters that better generalised the model.

2.3.2 Cross-Validation for Model Selection

Model selection in ML is the process of identifying which model provides higher statistical power in the generalisation of a problem. For example, in a classification scenario, many models can be used to predict a class, such as logistic regression, decision trees, support vector machine, among others. Depending on the dataset morphology, size, and context, a specific model may perform significantly better than another, and identifying such a model is a crucial step. According to Cox (2006), neglecting the process of finding the correct model to a specific problem is a majority mistake in statistical inference. K-fold CV is a reasonable method for that.

K-fold will be created for each given model. After the cross-validation, the averages are compared, and the one with a higher score is usually the best model for the underlying problem. For this context, sometimes it is also a good practice to test how significant is the performance of each model. It may help to decide whether worth the effort of a specific model over another for a specific problem. Depending on the purpose of a model, insignificant statistical differences between a soft-learning and a deep-learning option might not compensate for the tread-off of timing and computational consumption. Therefore, statistical tests to measure the means of each result set group can be applied (Berrar, 2019), ensuring the differences are statistically significant. For instance, in a scenario where three different models have been cross-validated over 30 folds, each 30-folds CV will result in a result set with the chosen metric to measure their performance (e.g., accuracy). Next, those three distributions of accuracy scores can be submitted to a paired t-test (whether in a normal distribution) to find out how significantly is the measured differences.

2.4 Bootstrap to Find Confidence Intervals

The confidence interval (CI) is the range between a lower limit (LL) and upper limit (UL) of a data distribuition, given an specific α (usually 5%). The estimation of value to be within a CI depends on the probability of LL be lower than such a value, which in turn, is lower than the UL. For example, let α be the likelihood desired for CI and θ be the measured value. Therefore, the probability that the interval contains the true value is at least $1 - \alpha$ for $LL \leq \theta \leq UL$ (Snijders, 2001).

Bootstrap is a statistical method derivated from the "jackknife resampling" technique. Similar to CV, it is useful for variance and bias estimation. It performs multiple estimations upon a predefined number of bootstraps (analogue to the k from the CV, however with replacement sampling) and averages the result set. Bootstrap applied to CI evaluation plays the rule of computing the before mentioned result variation, and check whether they fall within the CI (Cameron et al., 2005). In a ML classification problem, for example, it may determine that a model has 95% likelihood of classification accuracy between 82% and 93% (assuming that the calculated $LL = 0.82$ and $LL = 0.93$ for $\alpha = 0.05$). As exemplified in Dekking et al. (2006), consider the data: x_1, x_2, ..., x_n. Assuming this data has been extracted from $N(\mu, \sigma^2)$, where μ is the unknown mean of the data, and σ^2 is a known variance. Given that, the 95% CI of the mean can be calculated through the equation below (2.9).

95% CI representation.

$$\left[\overline{x} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{x} + 1.96\frac{\sigma}{\sqrt{n}}\right] \tag{2.9}$$

Where $\overline{x}$ is the sample mean, σ^2 is the variance, and n is the sample size.

However, if the data has been sampled from an unknown distribution, it is still possible to use the sample mean x to estimate μ but to find the CI surrounding μ will demand multiple iterations through subsamples. For such an achievement, bootstrap is employed.

Calculating the CI of one or more metrics from of a ML model is done as follows: First, the number of bootstraps, or resamplings (BS) is defined. It is analogue to the k from the k-fold CV. It is also a stratified subsample, but it allows replacements. Therefore, the number of bootstraps can be as large as computational capacity supports. Wilcox (2010) has proven through empirical experiments that $BS = 599$ is a reasonable number. Davidson & MacKinnon (2000) has clarified that such a specific number originated from the Monte Carlo tests, over the following scenario: in an exact test, let α be the significance level, and β the number of samples, then $\alpha \cdot (1 + \beta) = integer$. Considering the commonly significance levels $\alpha_1 = 0.01$, or $\alpha_2 = 0.05$, the number of boostraps will be as per equation 2.10:

Number of bootstraps equation.

$$\beta_1 = \frac{integer}{0.01} - 1, \ \beta_2 = \frac{interger}{0.05} - 1 \tag{2.10}$$

The -1 in the equation causes the number of bootstraps to be such as the "599", instead of 600, for example. The book Hair et al. (2017) advises to use about 500 samplings for initial analysis. When $\alpha = 0.05$, at least 1500, and for final analysis with $\alpha = 0.01$, at least 5000 subsamples.

Once the number of bootstraps is defined, the next step is to define the size of each bootstrap. Usually, the sampling is composed of 50% (sometimes 80%) of the data. As mentioned earlier, the subsampling is with replacements. Therefore it does not matter how many times it happens. Naturally, the rule of stratification also applies to bootstrap. It is essential to ensure that the class distribution reflects the distribution existent in the parent dataset. The third step is the iteration over each sample. The subsampling is once again split into train and test; the model is refit, trained, and validated. The outcome result is persisted in a separated array, then the loop repeats. Fourth and last step, already out of the loop, the LL and UL are calculated from the result set, against the pre-defined α. Such a range is, in fact, the confidence interval for $1 - \alpha$.

2.5 Test of Normality

Whenever it is desired to measure how significant is the difference between the means of two or more datasets, a statistical test is applied. A crucial step is to identify what statistical test is the most appropriated to a specific problem. One of the first decisions to be made is whether the test will be parametric or nonparametric. Such a decision is based on the distribution of the data. Whenever the data structure holds a normal distribution (e.g., Gaussian form), parametric tests should be used, and nonparametric otherwise. There are many techniques to identify whether a data distribution attends the requirements of normality. A common approach is a visual evaluation of histograms, Q-Q plots, or box-plots. As explained in Yap & Sim (2011), such a strategy offers reasonable information for a deductive conclusion. However, it is not possible to measure the accuracy of such a decision. Sometimes it may lead to misinterpretation, depending on how the images have been displayed. To address this problem, the employment of statistical tests to measure the level of normality is also recommended. It helps to interpret the graphical representation throughout measurable evidence to support the decision. The tests of normality work over the null hypothesis that states the data was drawn from a Gaussian distribution; therefore, it is normal. The statistical test must define a cut-off α, and whenever the test results in a p-value whose $p \leq \alpha$, it should reject such a hypothesis as it was found evidence that the data distribution does not comply with normality. Some of the statistical tests available to measure whether the data deviates from a Gaussian distribution are the Anderson-Darling test, Shapiro-Wilk through Kolmogorov-Smirnov (aka Goodness-Of-Fit) algorithm and D'Agostino-Pearson normality test.

Anderson-Darling test uses the cumulative distribution of the dataset upon the ideal cumulative distribution of a normal distribution to calculate the probability of the data not be within Gaussian form.

Shapiro-Wilk normality test (through Kolmogorov-Smirnov implementation), in turn, has some limitations. It does work well when values are unique across the data because the p-value is computed from a single value: the most significant discrepancy

between the cumulative distribution of the data and the cumulative normal distribution. Hence, it is not a practical way to assess normality. According to D'Agostino (1986), "*The Kolmogorov-Smirnov test is only a historical curiosity. It should never be used.*"

Last but not least, D'Agostino-Pearson normality test. This technique is based on skewness and kurtosis analysis. Skew quantifies how much of the data is in one of the sides of the data (left or right). The kurtosis, in turn, is the quantification of the distribution in the tail. Such a test calculates how far a normal distribution is, in terms of symmetry and shape. The most used implementation is the omnibus K^2 test. Such a test is well recommended over the literatures, as it takes into consideration the graphical format of the data. Some "rules of thumb" have also been derivated from skewness and kurtosis analysis to determine whether the data holds a normal distribution or not. For example, Darren George (2011) suggest the distribution can be assumed as normal whether the relevant standardised scores for skewness and kurtosis fall within the range ± 2. Also, Andy Field (2012) advises a distribution is normal whether 95% of the scores fall within the bounds of ± 3.29, for datasets larger than 80 cases.

2.6 Power Transformation (the Box-Cox)

Power transformation is applied whenever a more Gaussian-like distribution is desired. Usually, power transformation is employed before choosing to use a nonparametric statistical tests. It gives a chance to the data to fit into a normal distribution, and whenever it is the case, a parametric statistical test can safely be applied. There are many data transformation algorithm groups available, and Box and Cox (Box & Cox, 1964) (also known as Box-Cox transformation) is one of the most relevant groups of algorithms (Ruppert, 2001). The famous *log transformation* is part of the Box-Cox family of algorithms for power transformation. The Box-Cox is applied over positive outcome variables. Its mathematical definition can be observed below (equation 2.11):

Box-Cox transformation.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(y) & \text{if } \lambda = 0 \end{cases} \tag{2.11}$$

Where λ varies from -5 to 5.

The λ range is usually determined through maximum likelihood estimation, which depends on the underlying implementation. For example, some implementations may consider a fraction range such as $\lambda = [... - 1.5, -1, 0.5, 0, 0.5, 1, 1.5...]$ All of the inner values are iterated over the formula until an optimal value is discovered. The optimal value is the one that put the data distribution closer to a normal distribution.

2.7 Test of Paired Mean Differences

Among the statistical tests to compare populations are the dependent tests. They are applied whenever the samples are originated from the same population. There are many statistical tests for such an end, like the Dependent Student's T-test and Wilcoxon signed-rank test. Those mentioned tests should be employed whenever two dependent populations need to be compared. The T-test is applied for parametric data and Wilcoxon for nonparametric data. Next two subsections approach how each of them works, and how should they be interpreted. The theory presented in this section has been extracted from Gauthier & Hawley (2015).

2.7.1 Dependent Student's T-test

The Dependent Student's T-test, also known as paired t-test is a parametric statistical test to compare the mean of the differences between samples of dataset pairs. It works upon the null hypothesis (H_0) that states that the population mean of the differences between each data pair is equal to zero. So that, $H_0 : \mu_d = 0$. The statistic (t) is calculated as per equation 2.12.

Paired t-test.

$$t = \frac{d}{\frac{S_d}{\sqrt{n}}} \tag{2.12}$$

$$\text{Where } S_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n-1}}$$

Where d_i is the difference between the ith data pair and d is the sample mean of the differences between each data pair. The paired t-test assumes that the values of d_i are normally distributed. Given that, the t statistic can be compared with a tabular t value (α) with degrees of freedom of $n - 1$. The table will give the range of the probability falls, given the t and the degrees of freedom. With that, the null hypothesis can be rejected if the t exceeds the tabular value (or $p \leq \alpha$).

2.7.2 Wilcoxon Signed-Rank Test

The Wilcoxon signed-rank test (Wilcoxon, 1945) (also known as Wilcoxon Matched Pairs Signed Rank Test) is the nonparametric equivalent to the parametric paired student's t-test. It means that such a statistical test is appropriated to test two groups of data from the same population, whether they do not comply with the thresholds that qualify a normal distribution. Wilcoxon signed-rank test tests the null hypothesis that states that the difference between the groups follows a symmetric distribution around zero. Its statistic is calculated as per equation 2.13 below.

Wilcoxon Matched Pairs Signed Rank Test.

$$W = \sum_{i=1}^{N_r} [sng(x_{2,i} - x_{1,i}) \cdot R_i] \tag{2.13}$$

Similar to the t-test, a reference table gives critical values to interpret W statistic. The null hypothesis is rejected whether $|W|$ is higher than the underlying critical value. Another way to interpret it is through the p-value score. The null hypothesis is rejected whether $p \leq \alpha$, where α is a pre-determined threshold, usually 1% or 5%.

Chapter 3

Experiment Design and Methodology

This chapter covers how the research was conducted. It goes over every technical step taken to execute the experiments and achieve the objectives previously listed in section 1.3 (page 12). The chapter has been divided into five main sections: The first one (3.1) provides a context regarding the proposed design, which is a preparation for the hypothesis statement. The second section (3.2) contains the hypotheses aimed to be tested in order to answer the research question stated in section 1.2 (page 10). The third section (3.3) carries the steps to acquire and prepare the involved datasets. The fourth section (3.4) presents the technical details about the data modelling for each GWAS model, as much as the dataset's attributes. The fifth section (3.5) discloses the technical design and the steps taken to execute the models, collect the results, and evaluate the outcomes. The diagram below (figure 3.1) has a graphical representation of its content.

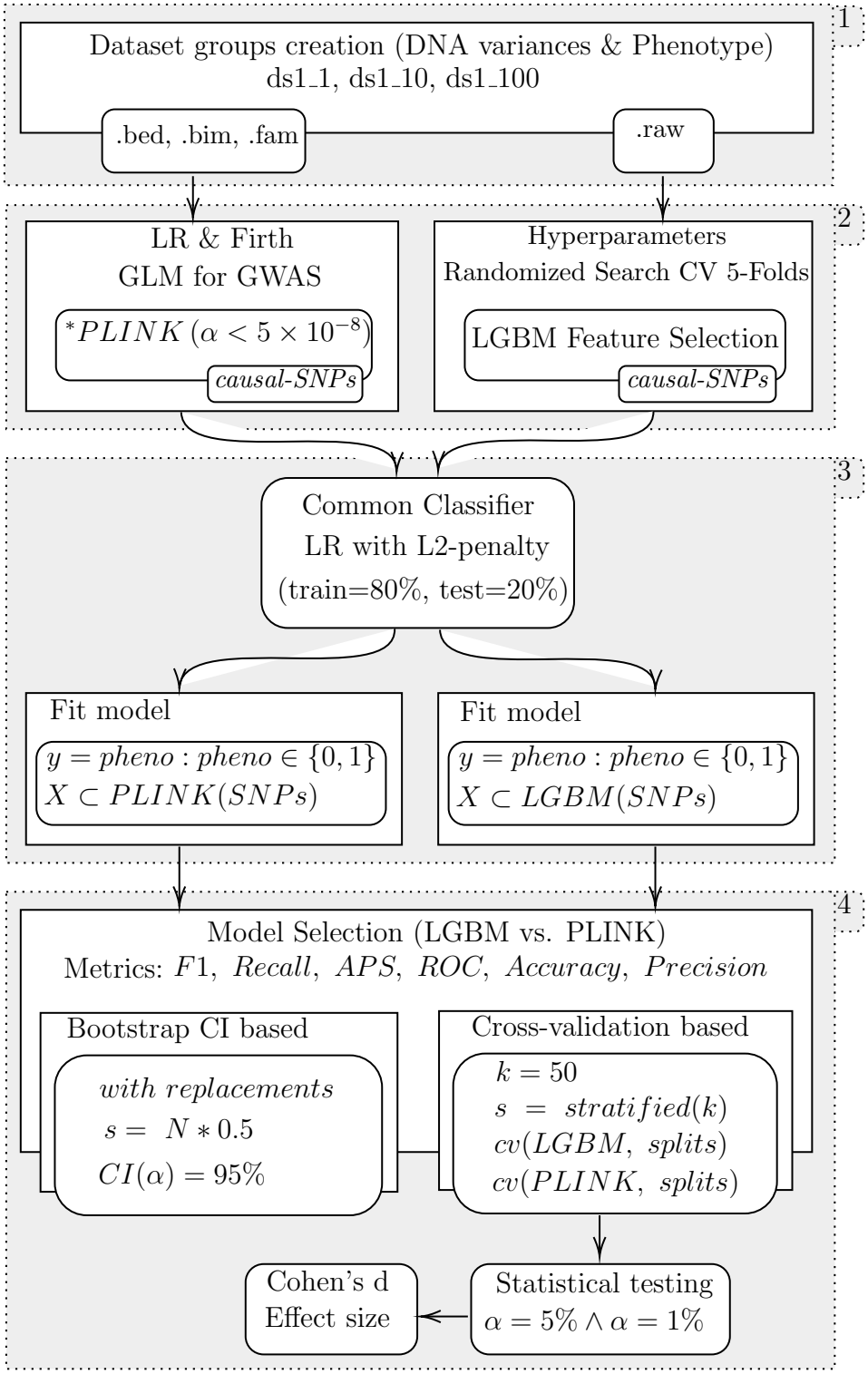

Figure 3.1: Design and implementation workflow diagram. Each step depicted at this image has been detailed below, in section 3.5 (page 41).

3.1 Design Context

A GWAS relies on two different data groups: the genomic data that contains the DNA variances, and the traits to be associated with the SNPs between the *cases* and *controls* cohorts. Usually, the traits to be investigated are human phenotypes, such as diseases status, that can be retrieved from the patients' electronic health records (EHR) (Zhou et al., 2018). This work encompasses the comparison of two GWAS techniques, over three different dataset scenarios, of qualitative traits. Therefore six models have been utilised, combined into three data groups named as *ds1_1*, *ds1_10* and *ds1_100*. The given names represents the class (phenotype status) distribution: $case : control = 1 : 1$, $case : control = 1 : 10$, and $case : control = 1 : 100$ respectively. The compared procedures are the PLINK2 GLM, a GWAS implementation based on GLM that employes Firth when the underlying phenotype is a qualitative type, and the proposed novel of this dissertation, the LightGWAS, a potential GWAS procedure based on LightGBM and CV.

LightGWAS aims to address limitations of the state-of-the-art, such as the problems related to big and sparse data. It increases the statistical power on the causal-SNPs selection, through GOSS and EFB (section 2.2, page 19). CV is also part of the solution. It is employed to select the optimal values of the GBDT hyperparameters. It allows the model to adapt to the underlying genomic data structures, reducing human intervention by dismissing QC steps. Consequently, it favours scalability, and even automation of the entire process. Accordingly, this research aims to accomplish two primary goals, which are listed below, in table 3.1, along with their respective experiments/tasks.

Research objectives	**Data sources**	**Experiments/Tasks**
O_1 - To test if LightGWAS can be used for GWAS.	*ds1_1* *ds1_10* *ds1_100*	**E_1 -** Setup a LightGBM with hyperparameters selected from a CV process.
		E_2 - Fit the model from E_1 with each dataset.
		E_3 - Evaluate if LightGWAS exposes the expected causal-SNPs.
O_2 - To test if LightGWAS outperforms a GWAS based on GLM implementation for each dataset.	*ds1_1*	**E_4 -** Run LightGWAS and a GLM models, against the dataset *ds1_1*.
		E_5 - Test if outcomes from E_4 are statistically significant.
	ds1_10	**E_6 -** Run LightGWAS and a GLM models, against the dataset *ds1_10*.
		E_7 - Test if outcomes from E_6 are statistically significant.
	ds1_10	**E_8 -** Run LightGWAS and a GLM models, against the dataset *ds1_100*.
		E_9 - Test if outcomes from E_8 are statistically significant.

Table 3.1: Research objectives and experiments/tasks.

The equations below are the metrics employed to evaluate the differences between LightGWAS and PLINK2 GLM:

Metrics to evaluate the GWAS models applied to binary phenotype.

$$
\begin{aligned}
Accuracy &= \tfrac{TP+TN}{TP+TN+FP+FN} \\
Precision &= \tfrac{TP}{TP+FP} \\
Recall &= \tfrac{TP}{(TP+FN)} \\
F1 &= 2 \cdot \tfrac{(precision \cdot recall)}{(precision+recall)} \\
APS &= \textstyle\sum_n (R_n - R_{n-1})\ P_n
\end{aligned}
\tag{3.1}
$$

Where *TP, TN, FP*, and FN are the sum of true-positives, true-negatives, false-positives, and false-negatives, respectively. P_n and R_n are the *precision* and *recall* at the nth threshold, respectively. *APS* stands for average precision score, and *F1* is the weighted average of the precision and recall. *Recall* is also known as *sensitivity*, *hit rate*, or *true-positive rate (TPR)*.

The area under the curve (AUC) for the receiver operating characteristic (ROC) has also been used to evaluate the efficiency of the models. It is calculated by plotting TPR against the FPR. The Scikit-learn implementation[1] have been adopted to measure it.

Caveat: Whenever a random data generation/selection was required over this study, the number 13 has been adopted as the initialisation state of the pseudo-random number for the underlying algorithm. It ensures the reproducibility of the experiments. Therefore, to reproduce the results of this research, such seed must be applied.

[1] https://bit.ly/scikit-learn-roc-auc

3.2 Hypotheses

The main challenge of the current state-of-the-art methods for GWAS is controlling the FPR, which means holding sufficient statistical power to avoid Type 1 error. Therefore, this study makes use of metrics surrounding the precision scores of models' result sets. Accuracy metric is also taken into account, however only for the balanced dataset, as such a rate becomes misleading for imbalanced datasets. Hence, the following hypotheses have been set:

- ***Null Hypothesis 1 (H_{01}):*** LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of balanced ($case : control = 1 : 1$) qualitative phenotypes, in terms of accuracy, precision, F1 score, and ROC/AUC.

  Alternative Hypothesis 1 (H_{A1}): LightGWAS outperforms GLM based on LR with Firth regularisation for GWAS, across genomic datasets of balanced ($case : control = 1 : 1$) qualitative phenotypes, in terms of accuracy, precision, F1 score, and ROC/AUC.

- ***Null Hypothesis 2 (H_{02}):*** LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of imbalanced ($case : control = 1 : 10$) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC.

  Alternative Hypothesis 2 (H_{A2}): LightGWAS outperforms GLM based on LR with Firth regularisation for GWAS, across genomic datasets of imbalanced ($case : control = 1 : 10$) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC.

- ***Null Hypothesis 3 (H_{03}):*** LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of high-imbalanced ($case : control = 1 : 100$) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC.

  Alternative Hypothesis 3 (H_{A3}): LightGWAS outperforms GLM based

on LR with Firth regularisation for GWAS, across genomic datasets of high-imbalanced ($case : control = 1 : 100$) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC.

3.3 Participants

A fully synthetic dataset for either genomic data and traits has been simulated. Such an approach was required to track the outcomes based on FPR and statistical power. Otherwise, it would not be possible to distinguish what are the causal-SNPs expected to be exposed by the underlying GWAS models. Therefore, a controlled dataset is vital. The same could be achieved through a phenotype labelled real data. However, the workflow to request and pass by the authorization process of most of the genomic and EHR database institutes would take longer than the expected time to conclude this research. Nonetheless, it is still recommended as a future avenue of research to make sure LightGWAS is as relevant as expected for real-world GWAS challenges.

Three fictitious cohorts have been created. Table 3.2 displays the phenotype (class) distributions of each of them. All the three cohorts had the same variance ratios to determine the (non-)causal-SNPs.

Dataset name	case:control distribution	cases	controls
ds1_1	1 : 1	2500	2500
ds1_10	1 : 10	400	4000
ds1_100	1 : 100	50	5000

Table 3.2: Cohorts' phenotype distribution.

Below is presented the parameters employed to create the synthetic datasets. They have been based on the PLINK SNP simulation tutorial. Naturally, they do not accurately represent realistic genetic data. However, they are relevant enough to test and validate GWAS methods in a controlled environment as per PLINK SNP

simulation tool documentation[2].

Ten thousand one hundred SNPs (features) compose each dataset group, where 100 of them had the population odds ratio set to 2.00, making them the causal-SNPs of the datasets. A prefix name has been employed to facilitate the traceability of each SNP during the analysis: "*d*" for causal-SNPs, and "*n*" for all the others. The MAF ratio was set to variate between 0.00 and 1.00, which ensures a high exposure of SNP among the genetic dataset.

SNPs datasets are usually formed by genes with two different alleles (heterozygous) or two identical alleles (homozygous) within a given ratio dimension. Given that, as per documentation earlier mentioned, the heterozygotes odds ratio has been set to 2.00 for *cases* and 1.00 for *controls*. The homozygotes odds ratio, in turn, received 4.00 and 1.00 respectively. Additional understanding of the mentioned biological terms is recommended, but not required. In the computational context of the models, they are merely adding the needed variances to spread the dataset as expected. Therefore beyond the scope of this dissertation. Table 3.3 consolidates the setup above described.

no. SNPs	**SNP Prefix**	**Lower allele frequency**	**Upper allele frequency range**	**Odds ratio for heterozygotes**	**Odds ratio for homozyygotes**
10000	n	0.00	1.00	1.00	1.00
100	d	0.00	1.00	2.00	4.00

Table 3.3: Phenotype ratios for genetic dataset build-up.

3.4 Datasets and Variables of Interest

The GWAS method based on LR with Firth applied in this research to compare against the LightGWAS is the PLINK2 GLM[3] implementation. PLINK2 accepts the genomic datasets in a set of specific formats, such as variant call format (VCF), or

[2] `http://zzz.bwh.harvard.edu/plink/simulate.shtml`

[3] `https://www.cog-genomics.org/plink/2.0/assoc#glm`

their own designed file set, composed of the file formats: $*.bed$, $*.bim$, and $*.fam$. As PLINK Simulation tool has been used to generate the synthetic datasets, the data was automatically created in its formats, facilitating the process. LightGWAS, in turn, depends on a tab-separated values (TSV) format. Therefore, once the datasets were simulated, the PLINK files have been converted to a tabular view, as described below.

The $*.bed$ file contains the genetic variates (SNPs) in a binary (non-human readable) format. It consists of the primary representation of genotype calls at biallelic variants. Analog to other ordinary ML data, they are the features of a dataset.

The $*.bim$, and $*.fam$ files complement the data required for the GWAS implementations that support PLINK files. Those files contain the phenotype, patients identification (ID's), and SNP list. Once again, analogue to a relational database, the $*.fam$ file is an intermediate table that contains the patient ID and the outcome variable (or *class* in ML classification models). It matches to the SNP table, represented by the $*.bim$ file in this analogy, throughout the binary $*.bed$ file. Tables 3.4 and 3.5 contains the variables of interest in the $*.fam$ and $*.bim$ files, respectively.

Variable	Type	Range	Sample
ID	Nominal	Alphanumeric	$per13$
Phenotype	Numeric	1=control, 2=case	2

Table 3.4: Variables of interest in the *.fam files.

Variable	Type	Range	Sample
SNP	Nominal	Alphanumeric	d_1312
Allele 1	Nominal	Category $[G, C, T$ or $A]$	G
Allele 2	Nominal	Category $[G, C, T$ or $A]$	C

Table 3.5: Variables of interest in the *.bim files.

The $*.raw$ file, in turn, is a tabular representation of all the others together. Among other columns, it contains the patient ID, phenotype status, and the SNPs extracted from the DNA. As mentioned before, the designed GWAS procedure evaluated at this

moment (LightGWAS) does not support the PLINK formats, nor VCFs, and this is why a TSV file format was required.

The $*.raw$ files are composed of $6 + V$ variables each, where V is the number of SNPs, which means 10106 columns form each dataset group. From the six first columns, only the individual ID and the phenotype (*class*) variables are relevant to the experiments, so that, the other four have been discarded. The additional 10100 columns that compose each SNP respects the following pattern: The column name contains the SNP identification (e.g., *d_9935_*), appended to the counted allele (e.g., T), and its alternate allele code in parentheses (e.g., $(/A)$). The value of each column is the SNP allelic dosage, which can be 0, 1, or 2. Once again, the biological understanding is recommended, but not required for the computing science context inferred. As can be seen, this is analogue to a large dataset, composed mostly of categorical variables and a binary class. The table 3.6 describes the file's morphology along with the variables of interest.

Variable	Type	Range	Sample
Individual ID	Nominal	Alphanumeric	*per*13
Phenotype	Numeric	1=control, 2=case	2
...			
n_1351_T(/A)	Numeric	[0, 1 or 2]	2
d_13_G(/T)			
...			

Table 3.6: Variables of interest in the *.raw files.

3.5 Procedure

Figure 3.1 (page 33) depicts the procedure executed to test the alternative hypotheses earlier stated. The procedure is composed of four layers, as outlined below:

1. Dataset creation: Details about the dataset has already been provided over the section 3.4 (page 39).

2. GWAS models:

   (a) PLINK2 GLM was used to conduct GWAS through LR with support to Firth. Firth aims to minimize the fitting errors through sparse regularisation terms. PLINK2 ensures that Firth regularisation is applied whenever LR algorithm fails due to low-count variants ($MAC < 400$). Once the association is completed, the causal-SNPs are extracted by filtering out those with standard cut-off $p \leq 5 \times 10^{-8}$ (Fadista et al., 2016). There is a caveat here: the dataset $ds1\_100$ demanded a cut-off of $p \leq 5 \times 10^{-4}$ because, within the first setup, no SNP was found by the model. Such a decision has been grounded on Ma et al. (2013). They also adopted a higher threshold on their tests due to a similar situation regarding the dataset balance.

   (b) The LightGWAS model, in turn, selects the causal-SNPs on model's fitting-time, which is the instant when the features of importance became available by the underlying GBDT algorithm. Therefore, finding out the best parameters to fit the model is the crucial step of the proposed solution. With that, to ensure the most relevant SNPs are correctly selected, a CV to pick the hyperparameters was proposed and employed. It makes sure the arguments to shape the model will be tested over multiple combinations until it finds the best-fit, as much as the most relevant features (the causal-SNPs) that better predicts the underlying phenotype (class of the model) for the given dataset group. In this experiment, a 5-folds through 200 iterations has been adopted. The *RandomizedSearchCV*[4] implementation has been used to test each light gradient boosting machine (aka lightgbm) (LGBM) parameter's rages. The chosen ranges have been set arbitrarily surrounding default-values available in the LightGBM documentation[5], and they can be found in the supplementary material, at the code-block A.1, of section A (page 83). Table 3.7 below contains the elected optimal hyperparameters for each dataset group. The pre-processing for the CV execution embraced

[4] `https://bit.ly/Sciskit-learn-RandomizedSearchCV`

[5] `https://lightgbm.readthedocs.io/en/latest/Parameters.html`

three stages: Firstly the unnecessary variables have been dropped, remaining only the ones outlined over the section 3.4. Secondly, the categorical variables have been factorized. Thirdly, the whole dataset has been split between a train (80%) and test (20%) subsample. Stratification has been adopted to ensure proportional distribution of *cases* and *controls* over each subsample.

	ds1_1	**ds1_10**	**ds1_100**
colsample_bytree	0.47328041	0.47328041	0.866621446
learning_rate	0.03	0.03	0.01
max_depth	1	1	6
min_child_samples	147	147	454
min_child_weight	1.0	1.0	1.0
min_split_gain	0	0	0
n_estimators	2000	2000	2000
num_leaves	35	35	41
reg_alpha	0.1	0.1	5
reg_lambda	0.1	0.1	50
subsample	0.995930118	0.995930118	0.820421212
subsample_for_bin	200000	200000	200000

Table 3.7: LightGBM's hyperparameters selected through 5-folds cross-validation.

Where the *colsample_bytree* represents the feature fraction. For example, a *colsample* of 0.7 means LightGBM will select 70% of feature before training each tree. *learning_rate* is the bias rate employed to restrict the influence of each tree on the final result. It controls the magnitude of the variance in the estimates. The *max_depth* controls overfitting by determining the depth of the tree splits. *min_child_samples* is the minimal number of information in one leaf of a tree. *min_child_weight* is the tree leaf minimal sum hessian. *min_split_gain* is the minimal number of splits/gain of a tree. *n_estimators* represents the number

of boosting iterations. *reg_alpha* and *reg_lambda* is the ratio applied to a L1 and L2 regularization, respectively. *subsample* is a randomly selection without resampling. *subsample_for_bin* is the number of data that sampled to construct histogram bins.

3. Common classifier.

   (a) Once both employed GWAS models were concluded, the selected SNPs could be used as input features for a common classification model, allowing a comparison between them. The Scikit Learn *LogisticRegression*[6] implementation has been chosen for such a common model. It is a LR with a L2-Penalty model. No customizations have been applied to such a common model, as the main goal is to verify how well a simple classification model could predict the phenotypes throughout the selected causal-SNPs of each evaluated model. The default parameters applied are listed below, in table 3.8.

LR parameters	**Values**
C	1
fit_intercept	TRUE
max_iter	200
multi_class	'auto'
penalty	L2'
solver	'lbfgs'
tol	0.0001
warm_start	FALSE

Table 3.8: Parameters for the logistic regression common classifier.

Where C represents the inverse of regularization strength. *fit_intercept* specify whether bias or intercept should be added to the decision function.

[6] `https://bit.ly/scikit-learn-LogisticRegression`

max_iter is the maximum number of iterations until the solvers converge. *multi_class* automatically identify if this is a multinominal problem. The current scenario is not a multinominal. Therefore, it is interpreted as a binary classification problem. *penalty* is the employed regularization method. *solver* is the optimization algorithm. *tol* accounts for the tolerance for stopping criteria. *warm_start* For learning transfer purposes. When set to *True*, it reuses the solution of the previous call to fit as initialization.

(b) The common classifier has been executed six times (each time means 50-fold CV and 5000 bootstraps to calculate the CI of each metric): Three times with the features (SNPs) selected by the LightGWAS and another three times with the PLINK2 picked SNPs. Three iterations each because the experiment is composed of three data groups as previously detailed in section 3.4 (page 39). An important caveat is, as per table 3.3 in section 3.3 (page 38), 100 SNPs were simulated as causal-SNPs. Therefore, as a preparation for the common classifier (model's pre-processing), only the top-100 most relevant selected features from each model have been used, avoiding unfair bias in the classification results. For LightGWAS model, the GBDT algorithm selects the features of importance based on *gain* scores. With that, the causal-SNPs extraction has been made by sorting the tree split gain scores descending, and then picked up the top-100 features. For PLINK2, in turn, the lower-100 SNPs from the cut-off SNP's p-value $\leq 5 \times 10^{-8}$ have been selected. Except for the dataset *ds1_100*, where a cut-off of p-value $\leq 5 \times 10^{-4}$ was required, as above explained. Once parameters and features were set, the models were trained with 80% of the datasets and tested with the other 20%. It was ensured that the train sub-sample does not contain the test sub-sample, avoiding overfitting in this matter.

4. The model selection layer generates quantitative data results that provide evidence to support or reject the null hypotheses of this research. They have been originated from a 50-folds CV for ML model selection, which has been

applied against each GWAS models. Moreover, to ensure the scores evaluated had representativeness, 5000 bootstraps have also been employed to calculate the confidence interval of each score. Each bootstrap iteration had 50% of the data, which in turn has been randomly selected through subsampling with replacements. Stratification of the phenotype (class) has been applied for either the 50-folds CV and in the subsamples of the 5k bootstraps. A CI of $95\%(LL = 0.025, UL = 0.975)$ has been take into consideration ($\alpha = 0.05$). Also, dependent (paired) sample Student's t-test and Wilcoxon signed-rank test have been implemented to test how significant were the differences between each measured metric against each model. The decision about which one to report depends on the distribution normality analysis that, in turn, relies on the skewness and kurtosis scores from each metric. According to Darren George (2011), the distribution can be assumed as normal whether the relevant standardised scores for skewness and kurtosis fall within the range ± 2. Also, Andy Field (2012) advises a distribution is normal if 95% of the scores fall within the bounds of ± 3.29, for datasets larger than 80 cases. Analysis of histogram plots along with D'Agostino-Pearson normality test (D'Agostino, 1986) has been applied for such an end. Whenever the data distribution did not comply with the thresholds of normality, Box-Cox transformation (Box & Cox, 1964) has been applied, in an attempt to set the data into a Gaussian form, before opting for a nonparametric approach. Therefore, t-test has been considered when the results held a normal distribution, and Wilcoxon otherwise. Either $\alpha = 1\%$ and $\alpha = 5\%$ cut-offs were evaluated, along with Cohen's d test (Cohen & Press, 1977) and Wilcoxon r score to measure the effect differences among each compared metric. The evaluated metrics were: *weighted average of the precision and recall (F1)*, *recall*, *average precision score (APS)*, *ROC/AUC*, *accuracy*, and *precision*.

(a) 50**-folds CV procedure:** Firstly two instances of the common classifiers were setup. One with the causal-SNPs selected by the LightGWAS model, and the other with the causal-SNPs selected by PLINK2 model. Secondly, the dataset was split into 50 stratified folds. It is important to mention

that the same folds have been used for both common classifiers, making sure they were tested with the same samples across each iteration. Thirdly, the CV has been executed for each common model, resulting in two new datasets, containing 50 records each, whose each column was one of the earlier mentioned statistical metrics. Finally, each metric pair (one from the LightGWAS common model, and the other from the PLINK common model) has been submitted to statistical tests, to evaluate how significantly was the measured differences. Also, tests of normality, based on histogram plotting analysis, kurtosis score, and skewness score have been executed to evaluate whether the distributions were in a Gaussian shape.

5k Bootstraps 95% CI: For each of the 5000 bootstraps, a resample (with replacement) containing 50% of the dataset has been applied. From this amount, 10% has been separated for test and the rest for the training. It was ensured that no test sample was included in the training sample ($test \notin training$). Over each iteration, both the common models have been refit with the same subsamples above mentioned, and the outcome trained models have been utilised to predict the test sample data. The predictions result sets were composed of five thousand records each, having the metrics as the variables of the dataset. Then they have been used to calculate the confidence interval. The lower limit has been fixed to 0.025 and upper limit to 0.975 in order to satisfy the 95% of CI ($\alpha = 0.05$).

Chapter 4

Results, Evaluation and Discussion

In this chapter, the research's results and evaluations are disclosed (section 4.1). It also counts with a discussion section (4.2), where the outcomes are debated concerning the approaches taken, the statistical significance of the results, and the relevance in scientifical perspectives.

As outlined in section 3.5 (page 41), the comparison between PLINK2 GLM[1] (Hill et al., 2017) and LightGWAS models have been done through 50-folds CV, and the CI of the results calculated through $5,000$ bootstraps (Wilcox, 2010; Hair et al., 2017). Tests of normality (Darren George, 2011; Andy Field, 2012) based on histogram analysis, kurtosis and skewness (D'Agostino-Pearson normality test (D'Agostino, 1986)) have been employed, followed by Box-Cox transformation (Box & Cox, 1964) whenever the data did not comply with normality thresholds. They aim to decide when a metric result distribution is Gaussian. Whenever a metric fits into the normality ranges, paired student's t-test is used. Wilcoxon signed-rank test otherwise. Also, Cohen's d and Wilcoxon r score have been applied to measure the models' differences effect, assisting on the evaluation analysis regarding scientifical relevance of the results (Greenland et al., 2016), besides their statistical significance.

[1]For the readability sake of this chapter, the model PLINK2 GLM will be called simply PLINK.

4.1 Results and Evaluation

The executed experiments for each dataset are disclosed and evaluated over this section. The raw results have been consolidated and appended to the supplementary material of this research (appendix A, from page 84).

4.1.1 Dataset ds1_1: Normality Test

Analysis of normality through histogram plots (see figure 4.1 below) suggests that *APS* and *ROC/AUC* hold a relevant degree of negative skewness for both PLINK and LightGWAS results. *accuracy*, *F1*, *precision*, and *recall* seems to have the distribuitions closer to the normality, however some of the scores have clustered the majority frequency of the values.

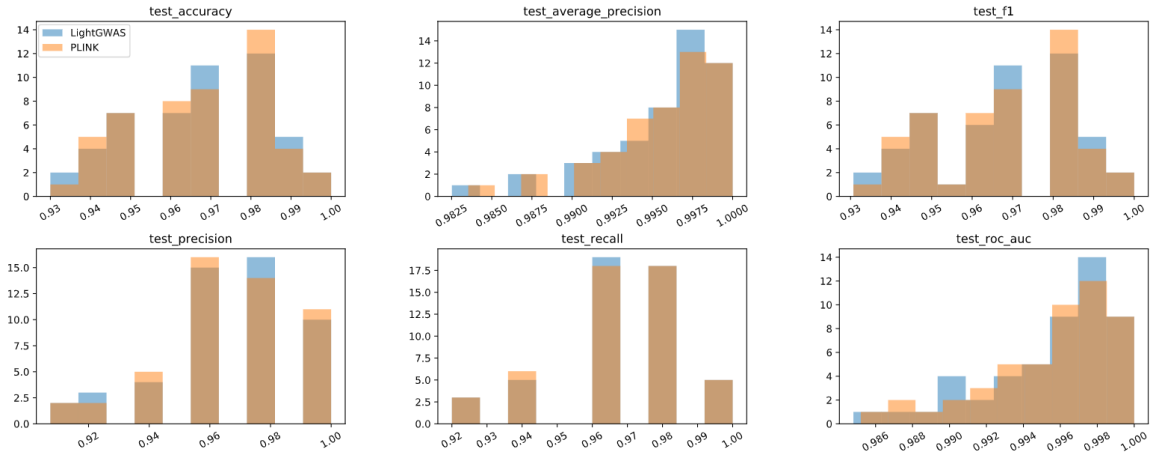

Figure 4.1: ds1_1 histograms.

D'Agostino's K^2 Normality Test has been conducted for a more accurate evaluation of the score's distribution. It tests the null hypothesis that states the data is in normal distribuition. The results are listed below:

- **F1**: LightGWAS ($skewness = -0.255$, $kurtosis = -0.648$, $SE = 0.002$, $z = 1.651$, $p = .438$). Plink ($skewness = -0.19$, $kurtosis = -0.754$, $SE = 0.002$, $z = 2.134$, $p = .344$).

- **recall**: LightGWAS ($skewness = -0.474$, $kurtosis = 0.048$, $SE = 0.003$, $z = 2.425$, $p = .298$). Plink ($skewness = -0.438$, $kurtosis = -0.122$, $SE = 0.003$, $z = 1.916$, $p = .384$).

- **APS**: LightGWAS ($skewness = -1.431$, $kurtosis = 2.208$, $SE = 0.001$, $z = 20.182$, $p < .001$). Plink ($skewness = -1.387$, $kurtosis = 2.036$, $SE = 0.0$, $z = 19.078$, $p < .001$).

- **ROC / AUC**: LightGWAS ($skewness = -1.11$, $kurtosis = 0.671$, $SE = 0.001$, $z = 11.21$, $p = .004$). Plink ($skewness = -1.107$, $kurtosis = 0.726$, $SE = 0.0$, $z = 11.335$, $p = .003$).

- **accuracy**: LightGWAS ($skewness = -0.291$, $kurtosis = -0.611$, $SE = 0.002$, $z = 1.649$, $p = .439$). Plink ($skewness = -0.217$, $kurtosis = -0.732$, $SE = 0.002$, $z = 2.064$, $p = .356$).

- **precision**: LightGWAS ($skewness = -0.7$, $kurtosis = 0.034$, $SE = 0.003$, $z = 4.638$, $p = .098$). Plink ($skewness = -0.611$, $kurtosis = -0.007$, $SE = 0.003$, $z = 3.626$, $p = .163$).

According to the results above, all of the metrics satisfied the Gaussian distribution, but *APS* and *ROC / AUC*. The results from the D'Agostino's normality test for both of them returned evidence on $\alpha = 0.01$ to reject the null hypothesis that states the data is normally distributed. Therefore, Box-Cox transformation has been applied to them, and a new iteration of D'Agostino's K^2 Normality Test resulted in a successful normalization of the data:

- **APS**: LightGWAS ($skewness = -0.172$, $kurtosis = -0.738$, $SE = 0.0$, $z = 1.933$, $p = .380$). Plink ($skewness = -0.144$, $kurtosis = -0.674$, $SE = 0.0$, $z = 1.371$, $p = .504$).

- **ROC / AUC**: LightGWAS ($skewness = -0.17$, $kurtosis = -0.794$, $SE = 0.0$, $z = 2.45$, $p = .294$). Plink ($skewness = -0.153$, $kurtosis = -0.759$, $SE = 0.0$, $z = 2.055$, $p = .358$).

4.1.2 Dataset ds1_1: Mean/Median Test

LightGWAS outperformed PLINK on metrics *F1*, *recall*, and *ROC/AUC*, while PLINK outperformed LightGWAS on *APS*, and *precision*. Both models reached out the same mean value for *accuracy* so that zero mean absolute difference (MD) in this metric. Once every metric held into the thresholds of a Gaussian distribution, a parametric test to measure how significant the observed differences between PLINK and LightGWAS could be applied. The paired t-test has been employed for such an end, and the results are listed below.

- **F1**: LightGWAS ($M = 0.967$, $SD = 0.017$, 95% CI $[0.962, 0.982]$) vs. PLINK ($M = 0.967$, $SD = 0.017$, 95% CI $[0.962, 0.984]$). $t(49) = 0.029$, $p = .977$, $MD < .001$, $d = 0.001$ (small effect).

- **recall**: LightGWAS ($M = 0.967$, $SD = 0.02$, 95% CI $[0.952, 0.984]$) vs. PLINK ($M = 0.966$, $SD = 0.02$, 95% CI $[0.952, 0.984]$). $t(49) = 0.375$, $p = .709$, $MD < .001$, $d = 0.02$ (small effect).

- **APS**: LightGWAS ($M = -0.003$, $SD = 0.002$, 95% CI $[-0.003, -0.001]$) vs. PLINK ($M = -0.003$, $SD = 0.002$, 95% CI $[-0.002, -0.001]$). $t(49) = 0.56$, $p = .578$, $MD < .001$, $d = 0.021$ (small effect).

- **ROC/AUC**: LightGWAS ($M = -0.003$, $SD = 0.002$, 95% CI $[-0.003, -0.001]$) vs. PLINK ($M = -0.003$, $SD = 0.002$, 95% CI $[-0.003, -0.001]$). $t(49) = 0.745$, $p = .460$, $MD < .001$, $d = 0.027$ (small effect).

- **accuracy**: LightGWAS ($M = 0.967$, $SD = 0.017$, 95% CI $[0.962, 0.982]$) vs. PLINK ($M = 0.967$, $SD = 0.017$, 95% CI $[0.962, 0.984]$). $t(49) = 0.0$, $p = 1.000$, $MD = 0$, $d < .001$ (small effect).

- **precision**: LightGWAS ($M = 0.969$, $SD = 0.025$, 95% CI $[0.964, 0.988]$) vs. PLINK ($M = 0.969$, $SD = 0.024$, 95% CI $[0.964, 0.992]$). $t(49) = -0.45$, $p = .655$, $MD < .001$, $d = 0.016$ (small effect).

The t-tests indicated no statistical significance on $\alpha = 0.05$ for any of the measured metrics. The standardized difference between the means resulted in a small effect for all of the metrics ($d < 0.5$). Also, there is 95% of a likelihood the reported LL and UL represent the confidence intervals of the true metrics' performance. The kernel density estimation (KDE) plot originated from the 5000 bootstraps to calculate such a CI has been appended to the supplementary material. It can be consulted at figure A.1 (page 81).

4.1.3 Dataset ds1_1: Discovery of Causal-SNPs

In terms of causal-SNP selection, LightGWAS selected 86 SNPs, while PLINK selected 90 SNPs. PLINK managed to pick all SNPs selected by LightGWAS, plus other four causal-SNPs.

4.1.4 Dataset ds1_10: Normality Test

Analysis of normality through histogram plots (see figure 4.2 below) suggests that all of the metrics have a relevant degree of negative skewness for both PLINK and LightGWAS results. The *accuracy* and *F1* seems closer to a normal distribution, but the image does not allow a conclusive judgment as the scores have clustered the majority frequency of the values.

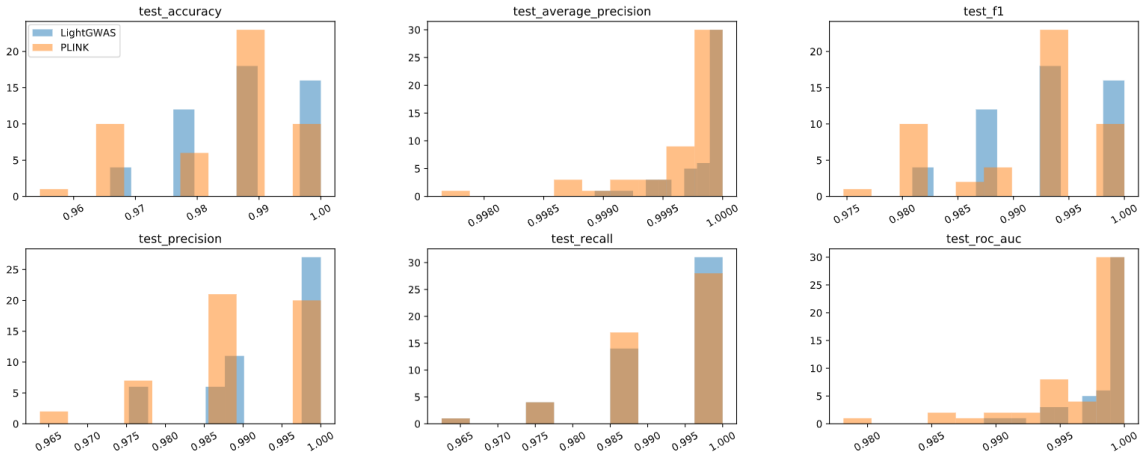

Figure 4.2: ds1_10 histograms.

D'Agostino's K^2 Normality Test has been conducted for a more accurate evaluation of the score's distribution. It tests the null hypothesis that states the data is in normal distribuition. The results are listed below:

- **F1**: LightGWAS ($skewness = -0.437$, $kurtosis = -0.762$, $SE = 0.001$, $z = 3.703$, $p = .157$). Plink ($skewness = -0.56$, $kurtosis = -0.661$, $SE = 0.001$, $z = 4.03$, $p = .133$).

- **recall**: LightGWAS ($skewness = -1.4$, $kurtosis = 1.423$, $SE = 0.001$, $z = 17.398$, $p < .001$). Plink ($skewness = -1.205$, $kurtosis = 1.044$, $SE = 0.001$, $z = 13.586$, $p = .001$).

- **APS**: LightGWAS ($skewness = -1.691$, $kurtosis = 2.118$, $SE = 0.0$, $z = 23.472$, $p < .001$). Plink ($skewness = -2.064$, $kurtosis = 4.797$, $SE = 0.0$, $z = 35.251$, $p < .001$).

- **ROC/AUC**: LightGWAS ($skewness = -1.684$, $kurtosis = 2.117$, $SE = 0.0$, $z = 23.383$, $p < .001$). Plink ($skewness = -1.892$, $kurtosis = 3.654$, $SE = 0.001$, $z = 30.331$, $p < .001$).

- **accuracy**: LightGWAS ($skewness = -0.428$, $kurtosis = -0.78$, $SE = 0.002$, $z = 3.804$, $p = .149$). Plink ($skewness = -0.554$, $kurtosis = -0.699$, $SE = 0.002$, $z = 4.223$, $p = .121$).

- **precision**: LightGWAS ($skewness = -0.762$, $kurtosis = -0.659$, $SE = 0.001$, $z = 6.169$, $p = .046$). Plink ($skewness = -0.729$, $kurtosis = -0.109$, $SE = 0.001$, $z = 4.787$, $p = .091$).

According to the results above, only the *F1* and *accuracy* satisfied the Gaussian distribuition. PLINK's *precision* also satisfied the thresholds of normality. The results from the D'Agostino's normality test for all the others returned evidence on $\alpha = 0.05$ to reject the null hypothesis that states the data is normally distributed. Therefore, Box-Cox transformation has been attempted to them, but a new iteration of D'Agostino's K^2 Normality Test also resulted in a non-Gaussian distribution. The

PLINK's *accuracy* was the only distribution normalized by Box-Cox transformation. In conclusion, only the *F1* and *accuracy* (non-transformed) can be validated through parametric tests. All the others will demand a nonparametric approach.

4.1.5 Dataset ds1_10: Mean/Median Test

LightGWAS outperformed PLINK for every measured metrics. Given that the metrics *recall*, *APS*, *ROC/AUC*, and *precision* did not comply with the thresholds of a normal distribution, a nonparametric test to measure how significant the observed differences between PLINK and LightGWAS had to be applied. The Wilcoxon signed-rank test has been employed for such an end. The metrics *F1* and *accuracy*, in turn, held the Gaussian distribution, so that paired t-test was applied for them. The results are listed below.

- **F1**: LightGWAS ($M = 0.993$, $SD = 0.006$, 95% CI $[0.988, 0.996]$) vs. PLINK ($M = 0.991$, $SD = 0.007$, 95% CI $[0.985, 0.994]$). $t(49) = 2.365$, $p = .022$, $MD = 0.002$, $d = 0.292$ (small effect).

- **recall**: LightGWAS ($Mdn = 1.0$, $M = 0.994$, $SD = 0.009$, 95% CI $[0.99, 1.0]$) vs. PLINK ($Mdn = 1.0$, $M = 0.993$, $SD = 0.009$, 95% CI $[0.985, 0.998]$). $z = 113.5$, $p = .662$, $MdnD = 0$, $MD = 0.001$, $r = 16.051$ (large effect), $d = 0.082$ (small effect).

- **APS**: LightGWAS ($Mdn = 1.0$, $M = 1.0$, $SD = 0.0$, 95% CI $[0.999, 1.0]$) vs. PLINK ($Mdn = 1.0$, $M = 1.0$, $SD = 0.0$, 95% CI $[0.999, 1.0]$). $z = 54.0$, $p = .002$, $MdnD < .001$, $MD < .001$, $r = 7.637$ (large effect), $d = 0.403$ (small effect).

- **ROC/AUC**: LightGWAS ($Mdn = 1.0$, $M = 0.998$, $SD = 0.003$, 95% CI $[0.995, 0.999]$) vs. PLINK ($Mdn = 0.998$, $M = 0.997$, $SD = 0.005$, 95% CI $[0.992, 0.999]$). $z = 48.5$, $p = .006$, $MdnD = 0.002$, $MD = 0.002$, $r = 6.859$ (large effect), $d = 0.403$ (small effect).

- **accuracy**: LightGWAS ($M = 0.988$, $SD = 0.011$, 95% CI $[0.977, 0.993]$) vs. PLINK ($M = 0.984$, $SD = 0.012$, 95% CI $[0.973, 0.989]$). $t(49) = 2.393$, $p = .021$, $MD = 0.003$, $d = 0.295$ (small effect).

- **precision**: LightGWAS ($Mdn = 1.0$, $M = 0.993$, $SD = 0.009$, 95% CI $[0.98, 0.995]$) vs. PLINK ($Mdn = 0.988$, $M = 0.99$, $SD = 0.01$, 95% CI $[0.98, 0.993]$). $z = 37.5$, $p = .007$, $MdnD = 0.012$, $MD = 0.003$, $r = 5.303$ (large effect), $d = 0.315$ (small effect).

The t-tests indicated statistical significance on $\alpha = 0.05$ for both *F1* and *accuracy*. The Wilcoxon test indicated statistical significance on $\alpha = 0.01$ for *APS*, *ROC/AUC* and *precision*. No statistical significance on $\alpha = 0.05$ has been observed for *recall*. The standardized difference between the means resulted in a small effect for all of the metrics ($d < 0.5$). Also, there is 95% of a likelihood the reported LL and UL represent the confidence intervals of the true metrics' performance. The KDE plot originated from the 5000 bootstraps to calculate such a CI has been appended to the supplementary material. It can be consulted at figure A.2 (page 82).

4.1.6 Dataset ds1_10: Discovery of Causal-SNPs

In terms of causal-SNP selection, LightGWAS selected 80 SNPs, while PLINK selected 76 SNPs. LightGWAS managed to pick all SNPs selected by PLINK, plus other four causal-SNPs.

4.1.7 Dataset ds1_100: Normality Test

Analysis of normality through histogram plots (see figure 4.3 below) are not conclusive. The scores have clustered the majority frequency of the values. No decision regarding normality can be made without a statistical test.

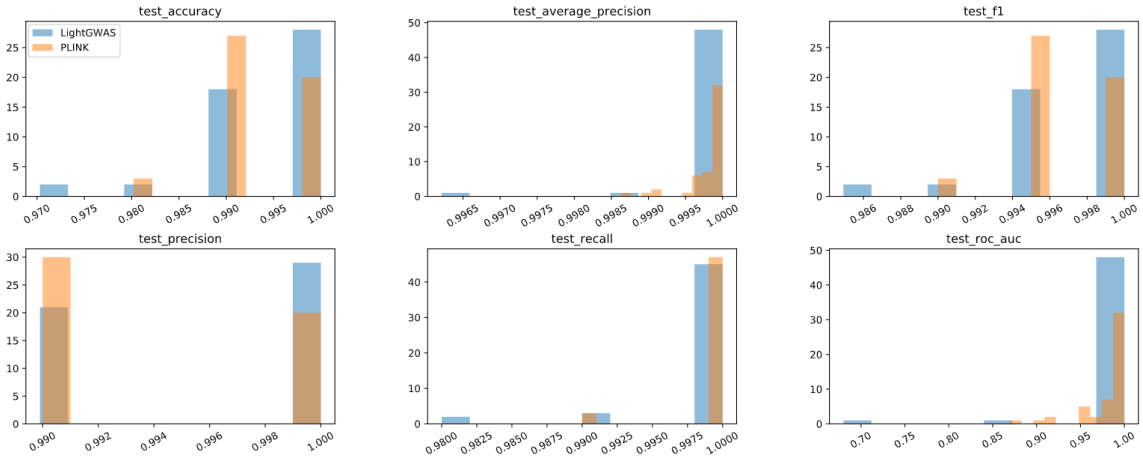

Figure 4.3: ds1_100 histograms.

D'Agostino's K^2 Normality Test has been conducted for a more accurate evaluation of the score's distribution. It tests the null hypothesis that states the data is in normal distribuition. The results are listed below:

- **F1**: LightGWAS ($skewness = -1.506$, $kurtosis = 2.257$, $SE = 0.001$, $z = 21.348$, $p < .001$). Plink ($skewness = -0.262$, $kurtosis = -0.642$, $SE = 0.0$, $z = 1.655$, $p = .437$).

- **recall**: LightGWAS ($skewness = -3.267$, $kurtosis = 9.747$, $SE = 0.001$, $z = 58.331$, $p < .001$). Plink ($skewness = -3.705$, $kurtosis = 11.73$, $SE = 0.0$, $z = 65.602$, $p < .001$).

- **APS**: LightGWAS ($skewness = -5.735$, $kurtosis = 33.536$, $SE = 0.0$, $z = 99.223$, $p < .001$). Plink ($skewness = -2.296$, $kurtosis = 4.998$, $SE = 0.0$, $z = 38.701$, $p < .001$).

- **ROC/AUC**: LightGWAS ($skewness = -5.506$, $kurtosis = 31.148$, $SE = 0.007$, $z = 96.246$, $p < .001$). Plink ($skewness = -2.219$, $kurtosis = 4.598$, $SE = 0.004$, $z = 36.85$, $p < .001$).

- **accuracy**: LightGWAS ($skewness = -1.485$, $kurtosis = 2.169$, $SE = 0.001$, $z = 20.807$, $p < .001$). Plink ($skewness = -0.25$, $kurtosis = -0.668$, $SE = 0.001$, $z = 1.756$, $p = .416$).

- **precision**: LightGWAS ($skewness = -0.324$, $kurtosis = -1.894$, $SE = 0.001$, $z = 1461.396$, $p < .001$). Plink ($skewness = 0.408$, $kurtosis = -1.833$, $SE = 0.001$, $z = 4711.908$, $p < .001$).

According to the results above, only PLINK's *F1* and *accuracy* satisfied the Gaussian distribuition. The results from the D'Agostino's normality test for all the others returned evidence on $\alpha = 0.05$ to reject the null hypothesis that states the data is normally distributed. Therefore, Box-Cox transformation has been attempted to them, but a new iteration of D'Agostino's K^2 Normality Test also resulted in a non-Gaussian distribution. Consequently, nonparametric method must be used to validate all of the metrics, as no one satisfied the thresholds of normality.

4.1.8 Dataset ds1_100: Mean/Median Test

LightGWAS outperformed PLINK for every measured metrics. Given that no metrics complied with the thresholds of a normal distribution, a nonparametric test to measure how significant the observed differences between PLINK and LightGWAS had to be applied. The Wilcoxon signed-rank test has been employed for such an end. The results are listed below.

- **F1**: LightGWAS ($Mdn = 1.0$, $M = 0.997$, $SD = 0.004$, 95% CI $[0.994, 0.997]$) vs. PLINK ($Mdn = 0.995$, $M = 0.997$, $SD = 0.003$, 95% CI $[0.994, 0.997]$). $z = 183.0$, $p = .431$, $MdnD = 0.005$, $MD < .001$, $r = 25.88$ (large effect), $d = 0.144$ (small effect).

- **recall**: LightGWAS ($Mdn = 1.0$, $M = 0.999$, $SD = 0.005$, 95% CI $[0.996, 1.0]$) vs. PLINK ($Mdn = 1.0$, $M = 0.999$, $SD = 0.002$, 95% CI $[0.994, 1.0]$). $z = 5.0$, $p = .234$, $MdnD = 0$, $MD = 0.001$, $r = 0.707$ (medium effect), $d = 0.221$ (small effect).

- **APS**: LightGWAS ($Mdn = 1.0$, $M = 1.0$, $SD = 0.001$, 95% CI $[1.0, 1.0]$) vs. PLINK ($Mdn = 1.0$, $M = 1.0$, $SD = 0.0$, 95% CI $[0.999, 1.0]$). $z = 163.5$,

$p = .096$, $MdnD = 0$, $MD < .001$, $r = 23.122$ (large effect), $d = 0.076$ (small effect).

- **ROC/AUC**: LightGWAS ($Mdn = 1.0$, $M = 0.987$, $SD = 0.048$, 95% CI $[0.964, 0.998]$) vs. PLINK ($Mdn = 1.0$, $M = 0.983$, $SD = 0.029$, 95% CI $[0.938, 0.985]$). $z = 166.5$, $p = .107$, $MdnD = 0$, $MD = 0.004$, $r = 23.547$ (large effect), $d = 0.11$ (small effect).

- **accuracy**: LightGWAS ($Mdn = 1.0$, $M = 0.994$, $SD = 0.008$, 95% CI $[0.988, 0.994]$) vs. PLINK ($Mdn = 0.99$, $M = 0.993$, $SD = 0.006$, 95% CI $[0.988, 0.994]$). $z = 180.0$, $p = .388$, $MdnD = 0.01$, $MD = 0.001$, $r = 25.456$ (large effect), $d = 0.147$ (small effect).

- **precision**: LightGWAS ($Mdn = 1.0$, $M = 0.996$, $SD = 0.005$, 95% CI $[0.99, 0.996]$) vs. PLINK ($Mdn = 0.99$, $M = 0.994$, $SD = 0.005$, 95% CI $[0.99, 0.994]$). $z = 176.0$, $p = .343$, $MdnD = 0.01$, $MD = 0.002$, $r = 24.89$ (large effect), $d = 0.36$ (small effect).

The Wilcoxon test indicated no statistical significance on $\alpha = 0.05$ for any of the measured metrics. The standardized difference between the means resulted in a small effect for all of the metrics ($d < 0.5$). Also, there is 95% of a likelihood the reported LL and UL represent the confidence intervals of the true metrics' performance. The KDE plot originated from the 5000 bootstraps to calculate such a CI has been appended to the supplementary material. It can be consulted at figure A.3 (page 83).

4.1.9 Dataset ds1_100: Discovery of Causal-SNPs

In terms of causal-SNP selection, LightGWAS selected 28 out of 100 SNPs, while PLINK selected 19 SNPs out of 100. LightGWAS managed to pick 14 SNPs missed by PLINK, and PLINK, in turn, managed to select 5 SNPs missed by LightGWAS. As mentioned in section 3.5 (page 41), PLINK cut-off for causal-SNPs in this dataset demanded $p \leq 5 \times 10^{-4}$ cut-off, as the cut-off $p \leq 5 \times 10^{-8}$ did not selected any SNP.

4.2 Discussion

LightGWAS performed as good as PLINK for GWAS over balanced datasets (*case* : *control* = 1 : 1). Although the slightly better performance of LightGWAS over PLINK, the statistical tests disclosed in section 4.1.2 (page 51) showed that none of the measured differences are statistically significant on cut-off $\alpha = 0.01$. Also, the measured effects through Cohen's d presented a small standardised effect between all the metrics' means. In terms of causal-SNP selection, PLINK outperformed LightGWAS in four units. LightGWAS managed to identify 86 causal-SNPs, while PLINK managed to identify 90 causal-SNPs. Both out of 100 causal-SNPs. Although the small difference, it is believed that a few more iterations in the CV employed to select the model's hyperparameter could address it better for LightGWAS. Such results demonstrate the effectiveness of LightGWAS as a GWAS method for balanced datasets, of qualitative phenotypes.

The experiments involving an imbalanced dataset (*case* : *control* = 1 : 10) as outlined in section 4.1.5 (page 54) brought evidences that support the alternative hypothesis H_{A2}. LightGWAS outperforms PLINK for such a scenario. Although recall did not reach statistical significance on $\alpha = 0.01$ (therefore as good as PLINK), all the other metrics had relevant results on $\alpha = 0.01$ (accuracy on $\alpha = 0.05$). Furthermore, the metrics measured through non-parametric tests (*recall*, APS, ROC/AUC and *precision*) resulted in a large effect ($r \geq 0.8$). LightGWAS also selected four more causal-SNPs than PLINK. LightGWAS identified 80 causal-SNPs while PLINK 76 SNPs.

When the models were submitted to a high-imbalanced dataset (*case* : *control* = 1 : 100), LightGWAS outperformed PLINK over every metric. Although the differences did not reach statistical significance ($\alpha = 0.05$), LightGWAS opened a medium effect margin for *recall* ($r \geq 0.5 \wedge r < 0.8$) and a large effect to the others ($r \geq 0.8$) against PLINK. Moreover, LightGBM selected fourteen causal-SNPs missed by PLINK, and also outperformed it in nine units. LightGBM discovered 28 causal-SNPs and plink 19 causal-SNPs.

Given the analysed results above, the proposed hypotheses of this dissertation are concluded as follows:

The null hypothesis H_{01} that states: *"LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of balanced (case : control = 1 : 1) qualitative phenotypes, in terms of accuracy, precision, F1 score, and ROC/AUC."* was maintained. LightGWAS performed as good as PLINK. There is no sufficient evidence to reject such a null hypothesis based on the performed statistical tests.

The null hypothesis H_{02} that states: *"LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of imbalanced (case : control = 1 : 10) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC."* was rejected. LightGWAS outperformed PLINK for every measured metric, having F1 and accuracy statistically significant on $\alpha = 0.05$, APS, ROC/AUC and *precision* statistically significant on $\alpha = 0.01$.

The null hypothesis H_{03} that states: *"LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of high-imbalanced (case : control = 1 : 100) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC."* was maintained. Although LightGWAS outperformed PLINK for every measured metric, the results are not statistically significant on $\alpha = 0.05$.

An important caveat is that statistical significance should not be the exclusive approach to evaluate how relevant is a GWAS model. As per Greenland et al. (2016), the scientific perspective (or significance) of the problem should also be taken into consideration. Although some metrics did not reach statistical significance, it is notorious how satisfactory did LightGWAS to select causal-SNPs, while controlling by false-positives and statistical power. Such an interpretation is especially correct within the tests involving (high-)imbalanced datasets. The results from the experiments over the $ds1\_100$ dataset, for example (aimed to test the null hypothesis H_{03}), had a medium effect ($r \geq 0.5 \wedge d < 0.8$) observed in the *recall* and and a large effect ($r \geq 0.8$) for the other measured metrics. Moreover, LightGWAS managed to select nine extra causal-SNPs over PLINK. Therefore, although the results were not statistically

significant (which maintained the exemplified null hypothesis), they are scientifically meaningful. By all the means, as earlier mentioned, the section 5.5 (page 69) present recommendations for future studies. It is possible that increasing the CV iterations of the LightGWAS design, to select the hyperparameters more accurately, will potentially improve the results, and the outcomes may also become statistically significant.

To conclude this chapter, the research question "*Can LightGWAS be an alternative method to the state-of-the-art for genome-wide association studies, by increasing statistical power on causal-SNP detection, and reduction of manual quality control steps?*" can be answered positively. The evidence collected from the tested hypotheses supports the theory that LightGWAS is a potential GWAS method, which performs as good as the current state-of-the-art for balanced datasets, and relatively better for imbalanced datasets.

Chapter 5

Conclusion

This dissertation has proposed the LightGWAS, a novel machine learning (ML) procedure for genome-wide association study (GWAS) based on LightGBM and k-fold cross-validation (CV). Its effectiveness has been assessed throughout a comparison with one of the available state-of-the-art implementations for GWAS, the PLINK2 (Hill et al., 2017). The experiments were designed and executed upon three different datasets of qualitative phenotypes: (1) balanced ($case : control = 1 : 1$), (2) imbalanced ($case : control = 1 : 10$) and (3) high-imbalanced ($case : control = 1 : 100$) class distribuitions. The results from statistical tests denoted that LightGWAS performs equivalently to PLINK2 method for balanced dataset scenarios, and outperforms for imbalanced and high-imbalanced datasets. The conducted literature review identified that the currently available GWAS implementations rely on massive manual steps to address statistical problems, such as controlling for false-positive inflation and power reduction (challenges increase as the data grows or become imbalanced). It also showed they demand a particular GWAS method for each type of genomic data structure, which increases human dependency. This research, thereupon, has presented evidence that LightGWAS is a potential single, resilient, autonomous and scalable solution to address such concerns.

5.1 Research Overview

GWAS is a crucial method to identify genetic risk factors of living beings (Bush & Moore, 2012). It aims to expose single-nucleotide polymorphisms (SNPs) (genetic variants) correlated to phenotypes (genetic traits). Analogue to an ordinary model's dataset, the SNPs represent the features (or dependent variables), and the phenotype the class (or independent variable). Therefore, GWAS allows the identification of genetic variants associated with specific traits. Such variants are the causal-SNPs of a phenotype (e.g., a disease). In this dissertation, LightGWAS has been presented as an alternative to the state-of-the-art for GWAS implementations. LightGWAS is a potential autonomous and self-contained GWAS method based on LightGBM: a leaf-wise growth gradient boosted decision trees (GBDT) implementation, with gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB). K-fold CV is also part of the proposed architecture, which is employed to find the best hyperparameters for the gradient boosting machine (GBM) model, ensuring it will adapt to different phenotype and genotype datasets' morphology. LightGWAS is a potential single-solution and scalable GWAS implementation that reduces human intervention. In this dissertation, LightGWAS has been confronted against PLINK2 general linear model (GLM) (Hill et al., 2017) implementation, one of the available state-of-the-art for GWAS. The comparisons have been conducted with three different genomic data structures of qualitative phenotype: balanced ($case : control = 1 : 1$), imbalanced ($case : control = 1 : 10$), and high-imbalanced ($case : control = 1 : 100$) datasets. Each model returned a set of causal-SNPs for each dataset. The models' effectiveness has been evaluated throughout the amounts of discovered causal-SNPs, as well as the overlaps between them. A common ML classifier (based on logistic regression (LR) with L1 regularization) was implemented to evaluate which GWAS method discovered the most relevant causal-SNPs. It has been fit once with the SNPs selected by LightGWAS, and another time with the SNPs returned by PLINK. Such a model selection process was conducted through 50-folds CV. The measured metrics' confidence interval were calculated through 5000 bootstraps 95% CI technique.

5.2 Problem Definition

The main problems surrounding the current state-of-the-art for GWAS implementations are related to maintaining statistical power by controlling false-positives through quality control (QC) steps. The QC actions rely on multiple manual approaches, such as choosing the correct statistical model depending on the phenotype and genomic data structure, tunning the features through principal component (PC) extracted from principal component analysis (PCA), and handling scaliness issues through external genomic data sources for data imputation purposes. The available GWAS methods also demands manual intervention whenever the minor allele frequency (MAF) differs between cases and controls due to regular ancestry deviations. Such a scenario denominates population stratification, and manual adjustments are required. Another issue is the exponential growth of genomic datasets due to cheaper technologies for deoxyribonucleic acid (DNA) sequencing. Besides the size of the datasets, false-positives tend to increase due to datasets too sparse, caused by the imbalanced distribution of case-control phenotypes, and a large number of features (SNPs) with a few samples. According to Bush & Moore (2012), SNPs datasets tend to be replaced by whole-genome in the near term. This is a realistic prediction considering that computational processing and storage capacity has increased at the same time their costs have reduced. It means that, instead of datasets with a few millions of SNPs, GWAS methods will have to handle about 3 billion nucleotides (features in the datasets). The available GWAS models, based exclusively on linear algorithms, are becoming obsolete for all these reasons. The manual tasks nowadays applied to address the problems mentioned above will be challenging across such scenarios, whether not impossible of execution.

5.3 Design/Experiments, Evaluation & Results

This research aimed to accomplish two main objectives: (a) To test if LightGWAS can be used for GWAS. (b) To test if LightGWAS outperforms one of the GWAS implementations that compose the current state-of-the-art. Since this research has been conducted with genomic datasets of qualitative phenotypes, with about five thousand

samples, the chosen GWAS method to compare with LightGWAS had to be an implementation based on GLM, which supports Firth regularisation, as per state-of-the-art definitions. The chosen method, thereupon, was PLINK2 GLM (Hill et al., 2017). Both LightGWAS and PLINK2 have been employed to discover the causal-SNPs. The extracted SNPs have been utilized as independent variables over the common classifier. The design has been grounded on literature reviews conducted through a secondary research method, and the hypotheses examined through deductive reasoning.

Firstly, the involved datasets had to be gathered. Three qualitative phenotype datasets have been simulated: ds1_1 ($N = 5000$, $cases = 2500$, $controls = 2500$), ds1_10 ($N = 4400$, $cases = 4000$, $controls = 400$), and ds1_100 ($N = 5050, cases = 5000, controls = 50$). They represent, respectively, the following examined scenarios: balance ($case : control = 1 : 1$), imbalanced ($case : control = 1 : 10$), and high-imbalanced ($case : control = 1 : 100$) genomic datasets. Secondly, LightGWAS was assembled. It is composed of a GBDT implementation called LightGBM framework, along with k-fold CV for optimal parameters selection. Such a framework is the state-of-the-art for GBM based on decision trees. The CV, in turn, ensures the framework adapts to any genomic data structure, reducing human dependency and allowing resilience and scalability of the model. Five folds over two hundred iterations have been employed for such an end. Thirdly, both LightGWAS and PLINK were utilised to perform GWAS across the early mentioned datasets. PLINK's outcome is a file with every SNPs accompanied by a p-value. The causal-SNPs filtering is reached by assuming a cut-off (α) for such a p-value. For the datasets ds1_1 and ds1_10, the cut-off $p \leq \alpha | \alpha = 5 \times 10^{-8}$ was assumed. In turn, for the dataset ds1_100, the cut-off had to be $\alpha =\leq 5 \times 10^{-4}$ because no SNP was selected with the first one. Those thresholds are the baseline for GWAS (Fadista et al., 2016; Mills & Rahal, 2019). The LightGWAS, on the other hand, scores each SNP with the *gain* score of the decision tree (DT) splits. Therefore, the extraction of the causal-SNPs selected by LightGWAS was, in fact, the list of features of importance generated right after the model's training.

Up to here, GWAS has been accomplished (causal-SNPs have been discovered).

However, it is also in the scope of this research to compare how effective is LightGWAS in comparison to PLINK. For this purpose, a third model has been employed (the earlier mentioned common classifier): A ML model based on LR with support to L1 regularisation. It was fit upon two conditions: once the features were the causal-SNPs collected by LightGWAS, and another with causal-SNPs selected by PLINK. The class (or target), in turn, was the phenotype variable. As a result, a set of comparable metrics was generated, allowing a contrast between LightGWAS and PLINK outcomes, in statistical terms. The employed metrics were: *weighted average of the precision and recall (F1)*, *recall*, *average precision score (APS)*, *receiver operating characteristic (ROC)/area under the curve (AUC)*, *accuracy*, and *precision*. The model selection was done through 50-folds CV technique, which generated a separated dataset with 50 result samples of each metric. Also, each metric result had its confidence interval (CI) set to 95% and validated through 5000 bootstraps' samples. The evaluation was performed through statistical tests. Dependent (paired) sample Student's t-test has been used for the metric result sets that held a normal distribution, and Wilcoxon signed-rank test otherwise. Tests to assess whether the metrics were in a Gaussian distribution were conducted with D'Agostino's K2 Normality Test, analysing skewness scores, kurtosis scores, and histogram plots. The effect of the observed mean differences was calculated through Cohen's d test, and Wilcoxon r score.

It has been observed that LightGWAS performed as good as PLINK for GWAS applied upon balanced datasets. LightGWAS had the best performance when measured by *F1*, *recall*, and *ROC/AUC*, and PLINK outperformed on *APS* and *precision*. Although such results, none of them reached statistical significance on $\alpha = 0.05$, and the measured differences in the means returned a small effect ($d < 0.5$) for every evaluated metric. Also, LightGWAS discovered 86 causal-SNPs, while PLINK selected 90 causal-SNPs. With that, the null hypothesis H_{01} that states "*LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of balanced (case : control = 1 : 1) qualitative phenotypes, in terms of accuracy, precision, F1 score, and ROC/AUC.*" was maintained. There was no sufficient statistical evidence to support that LightGWAS outperforms PLINK on balanced datasets of

qualitative phenotypes.

In turn, the experiments over the imbalanced dataset resulted that LightGWAS outperforms PLINK for every measured metric. Furtheremore, *F1* and *accuracy* were statistically significant on $\alpha = 0.05$, and *APS* and *ROC/AUC* on $\alpha = 0.01$. *Recall* was the only one with no statistical significance ($\alpha = 0.05$). However, it did reach a large effect ($r \geq 0.8$), along with *APS*, *ROC/AUC*, and *precision*. LightGWAS also outperformed PLINK on the causal-SNPs selection. It discovered 80 causal-SNPs, while PLINK did 76. Given these results, the null hypothesis H_{02} that states "*LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of imbalanced (case : control = 1 : 10) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC.*" was rejected. It was found statistical evidence that LightGWAS outperforms PLINK for GWAS with imbalanced datasets of qualitative phenotypes.

Last but not least, the results from the experiments with a high-imbalanced dataset disclosed that LightGWAS outperformed PLINK with a large effect ($r \geq 0.8$) across all the measured metrics. However, with no statistical significance on $\alpha = 0.5$ for any of them. LightGWAS also selected more causal-SNPs (28 units) than plink (19 units). Hence, the H_{03} that states "*LightGWAS do not outperform GLM based on LR with Firth regularisation for GWAS, across genomic datasets of high-imbalanced (case : control = 1 : 100) qualitative phenotypes, in terms of precision, F1 score, and ROC/AUC.*" was maintained. Although the large effect over the measured metrics, they did not reach statistical significance, thus, there is no enough statistical evidence that LightGWAS outperform PLINK for such a scenario.

There is a crucial point about the results aforementioned: As advised by Greenland et al. (2016), statistical significance should not be taken as the unique and final way to assess the relevance of a model. The scientific perspective (or significance) should also be considered. For example, although the results on high-imbalanced dataset did not reach statistical significance, the large effect observed for every metric are reasons of concern. The nonparametric approach provided by LightGWAS demonstrated to be more efficient than PLINK when the datasets are (high-)imbalanced. Since

LightGWAS outperformed PLINK consistently for such scenarios (95% of confidence interval has been ensured for every measured metric), it is reasonable to conclude that, although some outcomes did not reach statistical significance, they are scientifically meaningful for GWAS context.

5.4 Contributions and Impact

The proposed GWAS technique in this dissertation, LightGWAS, is a potential single-solution for every GWAS scenario. It reduces human dependency by employing ML techniques to eliminate manual QC steps often required by current state-of-the-art implementations. It has been validated with genomic datasets composed by qualitative phenotypes, with about 5,000 samples, in a balanced, imbalanced, and high-imbalanced trait distribution.

LightGWAS is based on LightGBM, the state-of-the-art for GBDT implementations. LighGBM is a framework designed to be highly efficient, with low memory consumption, capable of handling large and high-sparse data. It also supports graphics processing unit (GPU) technology, which reduces the models' training time significantly. Hence, this dissertation shows originality by taking a specific technique and applying it in a new domain. K-fold CV is also part of LightGWAS architecture. It is used for optimal parameters selection, which, in turn, ensures the underlying GBM model will adapt to any genomic data structure. It helps to reduce (possibly dismiss) human intervention, as QC steps to address statistical power is automatically handled through the model's adaptability. Therefore, LightGWAS can scale, following the growth of genomic datasets. As it eliminates manual interactions, it could even be automated through a computational pipeline.

This research has presented statistical evidence from empirical tests that LightGWAS can be used as a GWAS alternative procedure to the current state-of-the-art available methods. For all these reasons, LightGWAS is a new contribution from data science towards the evolvement of molecular biology science.

5.5 Future Work

This research has compared LightGWAS with PLINK2 GLM (Hill et al., 2017) association model. However, the current state-of-the-art for GWAS is composed of three methods: GLM with a regularisation function (such as the mentioned PLINK2 GLM), linear mixed model (LMM) (eg., BOLT-LMM), and scalable and accurate implementation of generalized mixed model (SAIGE). Depending on the phenotype representation (quantitative or qualitative), the size of the data, and the distribution, one of them should be employed. Therefore, for future studies, it is recommended to test LightGWAS against LMM, and SAIGE as well.

This research has been conducted with synthetic qualitative phenotype datasets. Hence, it should also be replicated with real data, of either qualitative and quantitative traits, so that LightGWAS can be evaluated against different data structures of real-world genomic datasets.

The causal-SNPs selected by LightGWAS are the products of a randomised search on hyperparameters with 5-fold CV 200 iterations. It is suggested for future studies to explore different numbers of iterations as much as the CV techniques, such as using grid search instead of randomised search CV.

Last but not least, it is recommended a research to develop a mechanism to identify causal-SNPs from DT's *gain* score, as no p-values exist in such a context for GWAS. It is crucial to develop a system analogue to the cut-off employed by the current state-of-the-art regression models for GWAS to filter causal-SNPs ($p \leq \alpha$).

References

Andy Field, Z. F., Jeremy Miles. (2012). *Discovering statistics using r*. SAGE Publications Ltd. Retrieved from https://www.xarg.org/ref/a/1446200450

Auton, A., Abecasis, G. R., Altshuler, D. M., et al. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. Retrieved from https://doi.org/10.1038/nature15393 doi: 10.1038/nature15393

Bengio, Y., & Grandvalet, Y. (2003, 11). No unbiased estimator of the variance of k-fold cross-validation. *CIRANO, CIRANO Working Papers*, *5*.

Berrar, D. (2019). Cross-validation. In *Encyclopedia of bioinformatics and computational biology* (pp. 542–545). Elsevier. Retrieved from https://doi.org/10.1016/b978-0-12-809633-8.20349-x doi: 10.1016/b978-0-12-809633-8.20349-x

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, *26*(2), 211–252. Retrieved from http://www.jstor.org/stable/2984418

Bush, W. S., & Moore, J. H. (2012, December). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822. Retrieved from https://doi.org/10.1371/journal.pcbi.1002822 doi: 10.1371/journal.pcbi.1002822

Cameron, A., Trivedi, P., Trivedi, P., Trivedi, P., Press, C. U., Library, E., & Corporation, E. (2005). *Microeconometrics: Methods and applications*. Cambridge University Press. Retrieved from https://books.google.ie/books?id=Zf0gCwxC9ocC

Cawley, G., & Talbot, N. (2010, 07). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079-2107.

Chen, X., & Ishwaran, H. (2012, June). Random forests for genomic data analysis. *Genomics*, *99*(6), 323–329. Retrieved from https://doi.org/10.1016/j.ygeno.2012.04.003 doi: 10.1016/j.ygeno.2012.04.003

Cohen, J., & Press, A. (1977). *Statistical power analysis for the behavioral sciences*. Academic Press. Retrieved from https://books.google.ie/books?id=H1iCAAAAIAAJ

Cox, D. R. (2006). *Principles of statistical inference*. Cambridge University Press. Retrieved from https://doi.org/10.1017/cbo9780511813559 doi: 10.1017/cbo9780511813559

D'Agostino, R. (1986). *Goodness-of-fit-techniques*. Taylor & Francis. Retrieved from https://books.google.ie/books?id=1BSEaGVBj5QC

Darren George, P. M. (2011). *Ibm spss statistics 19 step by step: A simple guide and reference*. Pearson; 12 edition. Retrieved from https://www.xarg.org/ref/a/0205255884

Davidson, R., & MacKinnon, J. G. (2000, January). Bootstrap tests: how many bootstraps? *Econometric Reviews*, *19*(1), 55–68. Retrieved from https://doi.org/10.1080/07474930008800459 doi: 10.1080/07474930008800459

Dekking, F., Kraaikamp, C., Lopuhaä, H., & Meester, L. (2006). *A modern introduction to probability and statistics: Understanding why and how*. Springer London. Retrieved from https://books.google.ie/books?id=TEcmHJX67coC

Duda, R., Hart, P., & Stork, D. (2012). *Pattern classification*. Wiley. Retrieved from https://books.google.ie/books?id=Br33IRC3PkQC

EMBL-EBI, European Bioinformatics Institute. (2020). *What are genome wide association studies (gwas)?* Retrieved from https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/Fig3.png ([Online; accessed May 1st, 2020])

Fadista, J., Manning, A. K., Florez, J. C., & Groop, L. (2016, January). The (in)famous GWAS p-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, *24*(8), 1202–1205. Retrieved from https://doi.org/10.1038/ejhg.2015.269 doi: 10.1038/ejhg.2015.269

Farrell, R. E. (2017). Functional genomics and transcript profiling. In *RNA methodologies* (pp. 685–695). Elsevier. Retrieved from https://doi.org/10.1016/b978-0-12-804678-4.00024-5 doi: 10.1016/b978-0-12-804678-4.00024-5

FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika*, *80*(1), 27–38. Retrieved from https://doi.org/10.1093/biomet/80.1.27 doi: 10.1093/biomet/80.1.27

Fitzmaurice, G. M., & Laird, N. M. (2015). Linear mixed models. In *International encyclopedia of the social & behavioral sciences* (pp. 162–168). Elsevier. Retrieved from https://doi.org/10.1016/b978-0-08-097086-8.42016-7 doi: 10.1016/b978-0-08-097086-8.42016-7

Gauthier, T. D., & Hawley, M. E. (2015). Statistical methods. In *Introduction to environmental forensics* (pp. 99–148). Elsevier. Retrieved from https://doi.org/10.1016/b978-0-12-404696-2.00005-9 doi: 10.1016/b978-0-12-404696-2.00005-9

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016, April). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. Retrieved from https://doi.org/10.1007/s10654-016-0149-3 doi: 10.1007/s10654-016-0149-3

Grinberg, N. F., Orhobor, O. I., & King, R. D. (2019, October). An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*. Retrieved from https://doi.org/10.1007/s10994-019-05848-5 doi: 10.1007/s10994-019-05848-5

Guo, Q., Wu, W., Massart, D., Boucon, C., & de Jong, S. (2002, February). Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, *61*(1-2), 123–132. Retrieved from https://doi.org/10.1016/s0169-7439(01)00203-9 doi: 10.1016/s0169-7439(01)00203-9

Hair, J., Sarstedt, M., Ringle, C., & Gudergan, S. (2017). *Advanced issues in partial least squares structural equation modeling*. SAGE Publications. Retrieved from https://books.google.ie/books?id=-f1rDgAAQBAJ

A haplotype map of the human genome. (2005, October). *Nature*, *437*(7063), 1299–1320. Retrieved from https://doi.org/10.1038/nature04226 doi: 10.1038/nature04226

Heinze, G. (2006). A comparative investigation of methods for logistic regression with separated or nearly separated data. *Statistics in Medicine*, *25*(24), 4216–4226. Retrieved from https://doi.org/10.1002/sim.2687 doi: 10.1002/sim.2687

Hill, A., Loh, P.-R., Bharadwaj, R. B., Pons, P., Shang, J., Guinan, E., ... Jelinsky, S. A. (2017, February). Stepwise distributed open innovation contests for software development: Acceleration of genome-wide association analysis. *GigaScience*, *6*(5). Retrieved from https://doi.org/10.1093/gigascience/gix009 doi: 10.1093/gigascience/gix009

Jiang, L., Zheng, Z., Qi, T., Kemper, K. E., Wray, N. R., Visscher, P. M., & Yang, J. (2019, November). A resource-efficient tool for mixed model association analysis of large-scale data. *Nature Genetics*, *51*(12), 1749–1755. Retrieved from https://doi.org/10.1038/s41588-019-0530-8 doi: 10.1038/s41588-019-0530-8

Ke, G., Meng, Q., Finely, T., Wang, T., Chen, W., Ma, W., ... Liu, T.-Y. (2017, December). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems 30 (nip 2017).* Retrieved from https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. , *14*.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling.* Springer New York. Retrieved from https://books.google.ie/books?id=xYRDAAAAQBAJ

Lee, S., Wright, F. A., & Zou, F. (2010, December). Control of population stratification by correlation-selected principal components. *Biometrics*, *67*(3), 967–974. Retrieved from https://doi.org/10.1111/j.1541-0420.2010.01520.x doi: 10.1111/j.1541-0420.2010.01520.x

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2018, January). Feature selection. *ACM Computing Surveys*, *50*(6), 1–45. Retrieved from https://doi.org/10.1145/3136625 doi: 10.1145/3136625

Liu, H., & Setiono, R. (1995). Chi2: feature selection and discretization of numeric attributes. In Anon (Ed.), *Proceedings of the international conference on tools with artificial intelligence* (pp. 388–391). IEEE. (Proceedings of the 1995 IEEE 7th International Conference on Tools with Artificial Intelligence ; Conference date: 05-11-1995 Through 08-11-1995) doi: 10.1109/tai.1995.479783

Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P., & Price, A. L. (2018, June). Mixed-model association for biobank-scale datasets. *Nature Genetics*, *50*(7), 906–908. Retrieved from https://doi.org/10.1038/s41588-018-0144-6 doi: 10.1038/s41588-018-0144-6

Lubke, G., Laurin, C., Walters, R., Eriksson, N., Hysi, P., Spector, T., ... Boomsma, D. (2013). Gradient boosting as a SNP filter: an evaluation using simulated and

hair morphology data. *Journal of Data Mining in Genomics & Proteomics*, *04*(04). Retrieved from https://doi.org/10.4172/2153-0602.1000143 doi: 10.4172/2153-0602.1000143

Ma, C., Blackwell, T., Boehnke, M., & and, L. J. S. (2013, June). Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genetic Epidemiology*, *37*(6), 539–550. Retrieved from https://doi.org/10.1002/gepi.21742 doi: 10.1002/gepi.21742

Mills, M. C., & Rahal, C. (2019, January). A scientometric review of genome-wide association studies. *Communications Biology*, *2*(1). Retrieved from https://doi.org/10.1038/s42003-018-0261-x doi: 10.1038/s42003-018-0261-x

Mo, K., & Li, J. (2019). A deep auto-encoder based LightGBM approach for network intrusion detection system. In *Proceedings of the international conference on advances in computer technology, information science and communications.* SCITEPRESS - Science and Technology Publications. Retrieved from https://doi.org/10.5220/0008098401420147 doi: 10.5220/0008098401420147

Pearson, T. A. (2008, March). How to interpret a genome-wide association study. *JAMA*, *299*(11), 1335. Retrieved from https://doi.org/10.1001/jama.299.11.1335 doi: 10.1001/jama.299.11.1335

Pérez-Enciso, & Zingaretti. (2019, July). A guide for using deep learning for complex trait genomic prediction. *Genes*, *10*(7), 553. Retrieved from https://doi.org/10.3390/genes10070553 doi: 10.3390/genes10070553

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., & Reich, D. (2006, July). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, *38*(8), 904–909. Retrieved from https://doi.org/10.1038/ng1847 doi: 10.1038/ng1847

Reed, E., Nunez, S., Kulp, D., Qian, J., Reilly, M. P., & Foulkes, A. S. (2015, September). A guide to genome-wide association analysis and post-analytic in-

terrogation. *Statistics in Medicine*, *34*(28), 3769–3792. Retrieved from https://doi.org/10.1002/sim.6605 doi: 10.1002/sim.6605

Ruppert, D. (2001). Statistical analysis, special problems of: Transformations of data. In *International encyclopedia of the social & behavioral sciences* (pp. 15007–15014). Elsevier. Retrieved from https://doi.org/10.1016/b0-08-043076-7/00513-1 doi: 10.1016/b0-08-043076-7/00513-1

Saunders, M., Lewis, P., & Thornhill, A. (2009). *Research methods for business students*. Prentice Hall. Retrieved from https://books.google.ie/books?id=u-txtfaCFiEC

Schratz, P., Muenchow, J., Iturritxa, E., Richter, J., & Brenning, A. (2019, August). Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data. *Ecological Modelling*, *406*, 109–120. Retrieved from https://doi.org/10.1016/j.ecolmodel.2019.06.002 doi: 10.1016/j.ecolmodel.2019.06.002

Sebastiani, P., Timofeev, N., Dworkis, D. A., Perls, T. T., & Steinberg, M. H. (2009, August). Genome-wide association studies and the genetic dissection of complex traits. *American Journal of Hematology*, *84*(8), 504–515. Retrieved from https://doi.org/10.1002/ajh.21440 doi: 10.1002/ajh.21440

Shah, S. C., & Kusiak, A. (2004, July). Data mining and genetic algorithm based gene/SNP selection. *Artificial Intelligence in Medicine*, *31*(3), 183–196. Retrieved from https://doi.org/10.1016/j.artmed.2004.04.002 doi: 10.1016/j.artmed.2004.04.002

Shao, J. (1993, June). Linear model selection by cross-validation. *Journal of the American Statistical Association*, *88*(422), 486–494. Retrieved from https://doi.org/10.1080/01621459.1993.10476299 doi: 10.1080/01621459.1993.10476299

Singh, S. P., Singh, P., & Mishra, A. (2020, February). Predicting potential applicants for any private college using LightGBM. In *2020 international confer-*

ence on innovative trends in information technology (ICITIIT). IEEE. Retrieved from https://doi.org/10.1109/icitiit49094.2020.9071525 doi: 10.1109/icitiit49094.2020.9071525

Snijders, T. (2001). Hypothesis testing: Methodology and limitations. In *International encyclopedia of the social & behavioral sciences* (pp. 7121–7127). Elsevier. Retrieved from https://doi.org/10.1016/b0-08-043076-7/00737-3 doi: 10.1016/b0-08-043076-7/00737-3

Song, Y., Jiao, X., Qiao, Y., Liu, X., Qiang, Y., & Liu, Z. (2019). Prediction of double-high biochemical indicators based on LightGBM and XGBoost. In *Proceedings of the 2019 international conference on artificial intelligence and computer science - AICS 2019*. ACM Press. Retrieved from https://doi.org/10.1145/3349341.3349400 doi: 10.1145/3349341.3349400

Spencer, C. C. A., Su, Z., Donnelly, P., & Marchini, J. (2009, May). Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. *PLoS Genetics*, *5*(5), e1000477. Retrieved from https://doi.org/10.1371/journal.pgen.1000477 doi: 10.1371/journal.pgen.1000477

Sukhumsirichart, W. (2018, 10). Polymorphisms.. doi: 10.5772/intechopen.76728

Urso, A., Fiannaca, A., Rosa, M. L., Ravì, V., & Rizzo, R. (2019). Data mining: Prediction methods. In *Encyclopedia of bioinformatics and computational biology* (pp. 413–430). Elsevier. Retrieved from https://doi.org/10.1016/b978-0-12-809633-8.20462-7 doi: 10.1016/b978-0-12-809633-8.20462-7

Wang, R., Liu, Y., Ye, X., Tang, Q., Gou, J., Huang, M., & Wen, Y. (2019, November). Power system transient stability assessment based on bayesian optimized LightGBM. In *2019 IEEE 3rd conference on energy internet and energy system integration (EI2)*. IEEE. Retrieved from https://doi.org/10.1109/ei247390.2019.9062027 doi: 10.1109/ei247390.2019.9062027

White, D., & Rabago-Smith, M. (2010, October). Genotype–phenotype associations and human eye color. *Journal of Human Genetics*, *56*(1), 5–7. Retrieved from https://doi.org/10.1038/jhg.2010.126 doi: 10.1038/jhg.2010.126

Wilcox, R. (2010). *Fundamentals of modern statistical methods: Substantially improving power and accuracy*. Springer New York. Retrieved from https://books.google.ie/books?id=PUkOBwAAQBAJ

Wilcoxon, F. (1945, December). Individual comparisons by ranking methods. *Biometrics Bulletin*, *1*(6), 80. Retrieved from https://doi.org/10.2307/3001968 doi: 10.2307/3001968

Yang, J., Lee, S. H., Goddard, M. E., & Visscher, P. M. (2011, January). GCTA: A tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, *88*(1), 76–82. Retrieved from https://doi.org/10.1016/j.ajhg.2010.11.011 doi: 10.1016/j.ajhg.2010.11.011

Yap, B. W., & Sim, C. H. (2011, December). Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, *81*(12), 2141–2155. Retrieved from https://doi.org/10.1080/00949655.2010.520163 doi: 10.1080/00949655.2010.520163

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

Zhou, W., Nielsen, J. B., Fritsche, L. G., Dey, R., Gabrielsen, M. E., Wolford, B. N., ... Lee, S. (2018, August). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, *50*(9), 1335–1341. Retrieved from https://doi.org/10.1038/s41588-018-0184-y doi: 10.1038/s41588-018-0184-y

Appendices

Appendix A

Suplementary Material

Parameters	Ranges
learning_rate	$[0.01, 0.02, 0.03, 0.04, 0.05, 0.08, 0.1, 0.2, 0.3, 0.4]$
n_estimators	$[100, 200, 300, 400, 500, 600, 800, 1000, 1500, 2000, 3000, 5000]$
num_leaves	$sp\_randint(6, 50)$
min_child_samples	$sp\_randint(100, 500)$
min_child_weight	$[1e-5, 1e-3, 1e-2, 1e-1, 1, 1e1, 1e2, 1e3, 1e4]$
subsample	$sp\_uniform(loc = 0.2, scale = 0.8)$
max_depth	$[-1, 1, 2, 3, 4, 5, 6, 7]$
colsample_bytree	$sp\_uniform(loc = 0.4, scale = 0.6)$
reg_alpha	$[0, 1e-1, 1, 2, 5, 7, 10, 50, 100]$
reg_lambda	$[0, 1e-1, 1, 5, 10, 20, 50, 100]$

Table A.1: LightGBM hyperparameters' ranges for the 5-fold CV.

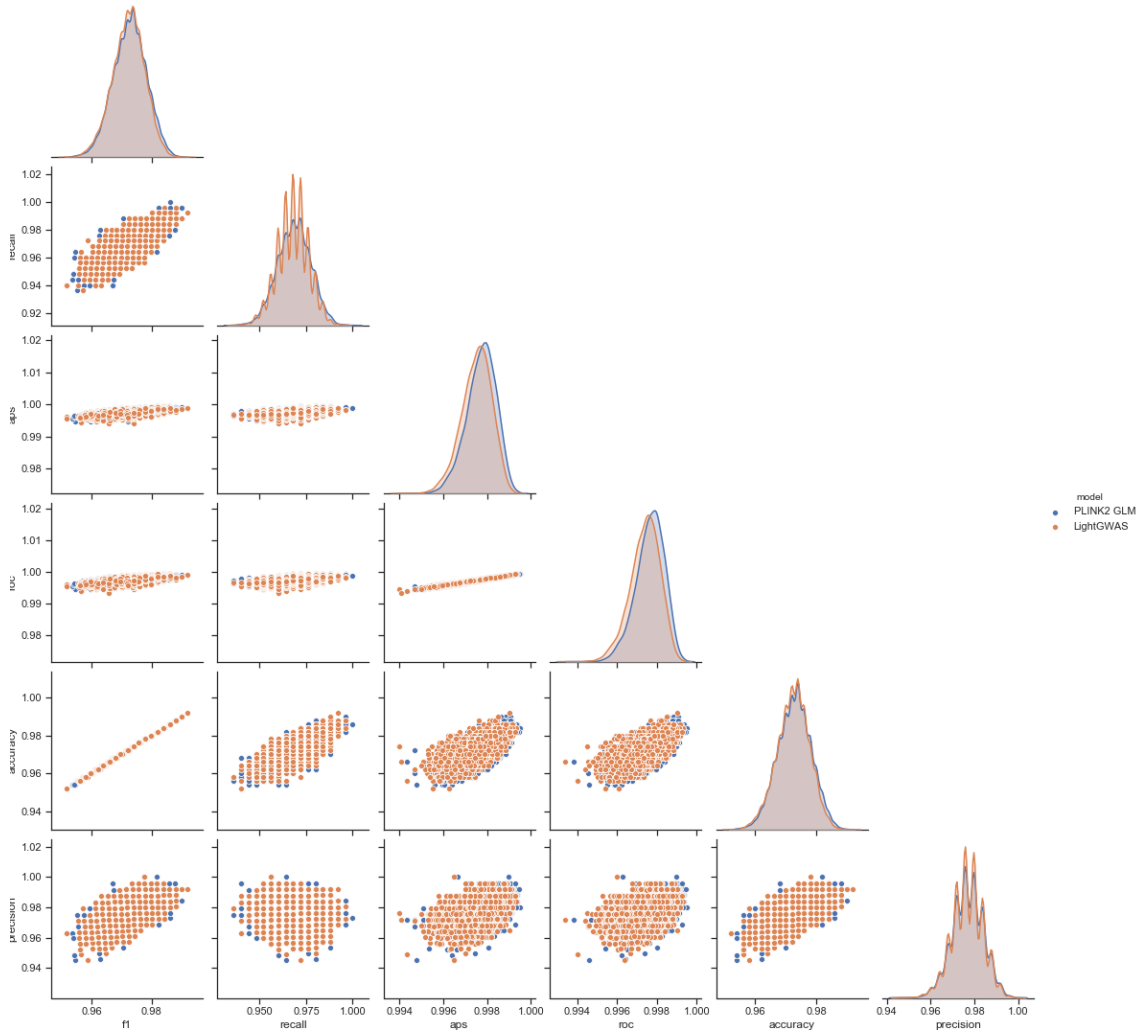

Figure A.1: ds1_1 5k bootstraps: pairwise KDE relationships.

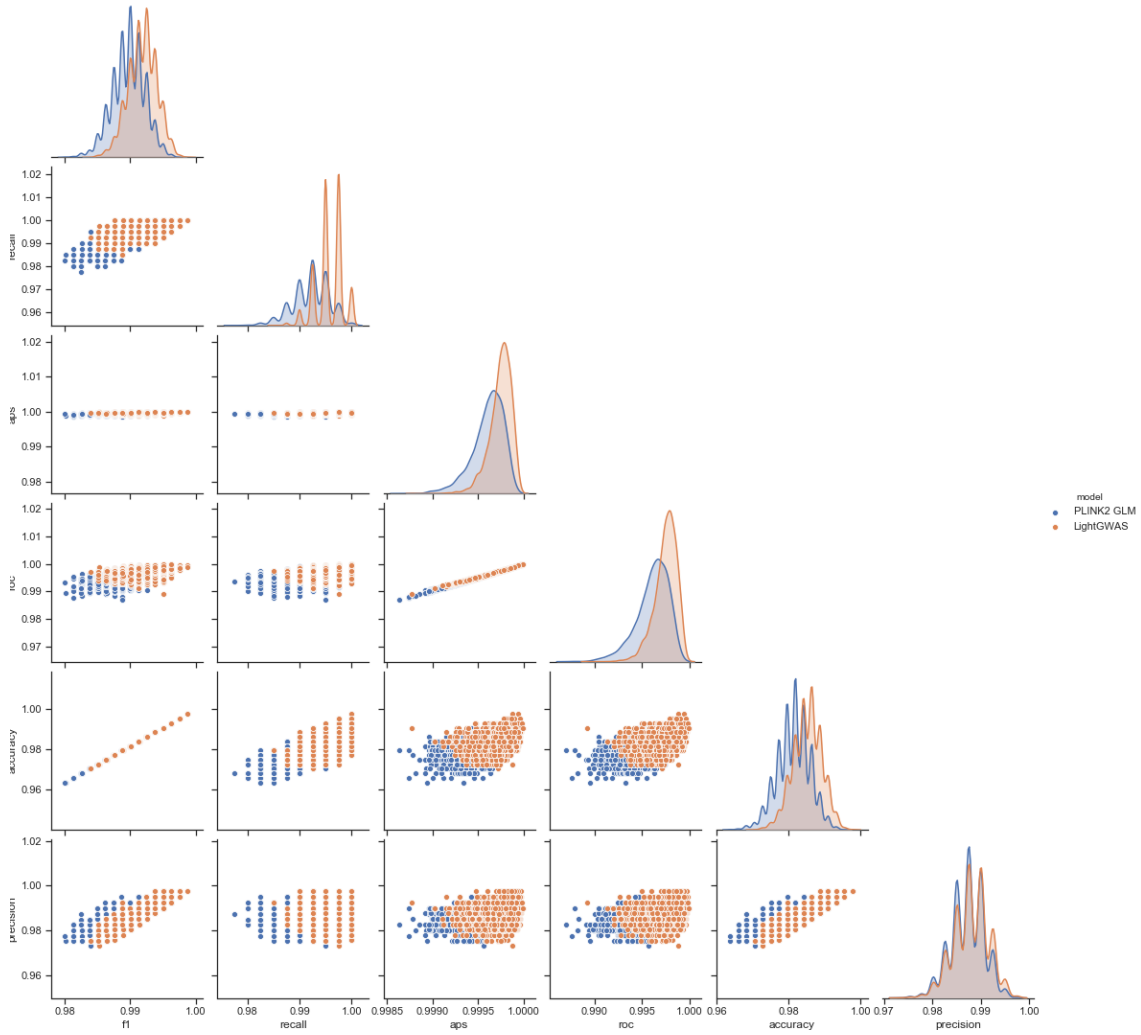

Figure A.2: ds1_10 5k bootstraps: pairwise KDE relationships.

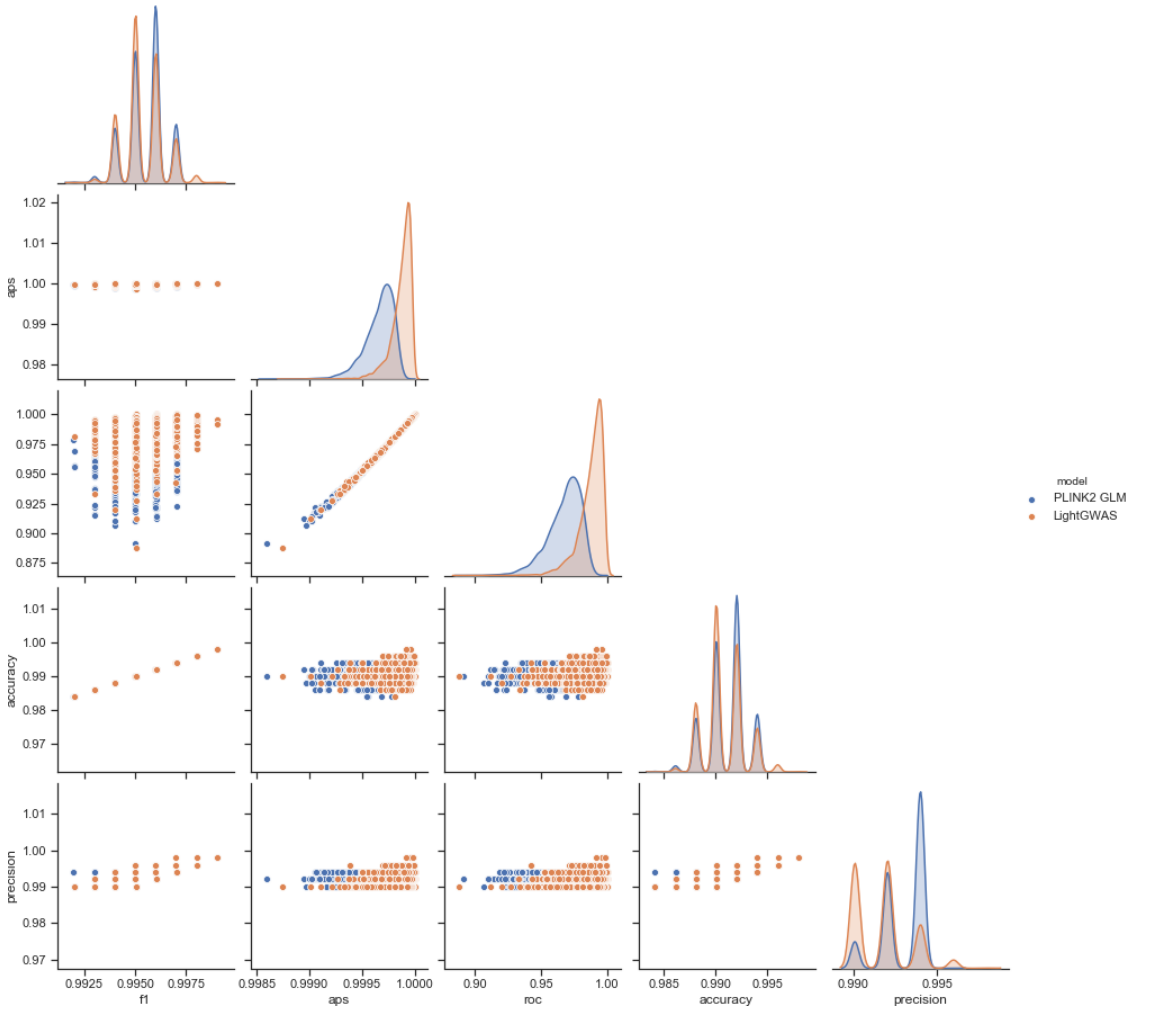

Figure A.3: ds1_100 5k bootstraps: pairwise KDE relationships. Note: Omitted *recall* because its data did not satisfy a KDE plot.

50-folds Cross-Validation	**f1**	**recall**	**average precision**	**roc auc**	**accuracy**	**precision**
LightGWAS mean	0.967 436	0.9668	0.995 725	0.995 664	0.9674	0.968 505
LightGWAS std	0.017 298	0.020 045	0.003 669	0.003 572	0.017 474 18	0.024 702
LightGWAS Mdn	0.969 697	0.96	0.996 973	0.9968	0.97	0.978 945
LightGWAS skew	−0.254 781	−0.474 451	−1.430 597	−1.109 939	−0.291 389 1	−0.700 194
LightGWAS kurtosis	−0.647 515	0.048 439	2.2084	0.671 16	−0.610 907 6	0.033 728
LightGWAS SE	0.002 446	0.002 835	0.000 519	0.000 505	0.002 471 222	0.003 493
PLINK mean	0.967 416	0.9664	0.995 748	0.995 648	0.9674	0.968 896
PLINK std	0.016 862	0.020 38	0.003 506	0.003 49	0.017 000 6	0.024 434
PLINK Mdn	0.969 997	0.96	0.996 767	0.9968	0.97	0.970 131
PLINK skew	−0.189 643	−0.438 049	−1.386 734	−1.107 475	−0.216 736 6	−0.610 631
PLINK kurtosis	−0.753 602	−0.122 315	2.036 149	0.725 61	−0.732 358 5	−0.006 701
PLINK SE	0.002 385	0.002 882	0.000 496	0.000 494	0.002 404 248	0.003 456
LightGWAS Agostino stats	1.650 727	2.424 599	20.182 116	11.210 152	1.648 631	4.638 44
LightGWAS Agostino p-val	0.438 076	0.297 512	0.000 041	0.003 679	0.438 535 2	0.098 35
LightGWAS Agostino sig 5%?	0	0	1	1	0	0
LightGWAS Agostino sig 1%?	0	0	1	1	0	0
PLINK Agostino stats	2.133 753	1.915 882	19.077 826	11.335 082	2.063 66	3.626 191
PLINK Agostino p-val	0.344 082	0.383 682	0.000 072	0.003 456	0.356 354 2	0.163 148
PLINK Agostino sig 5%?	0	0	1	1	0	0
PLINK Agostino sig 1%?	0	0	1	1	0	0
LightGWAS ≥ PLINK (M)?	1	1	0	1	1	0
MD	0.000 02	0.0004	0.000 022	0.000 016	0	0.000 39
LightGWAS ≥ PLINK (Mdn)?	0	1	1	1	1	1
MdnD	0.0003	0	0.000 206	0	0	0.008 814
LightGWAS 5k BS 95% (0.025)	0.961 616	0.952	0.996 011	0.995 76	0.962	0.963 563
LightGWAS 5k BS 95% (0.975)	0.981 966	0.984	0.998 711	0.998 672	0.982	0.987 904
LightGWAS M 95% CI?	1	1	0	0	1	1
LightGWAS Mdn 95% CI?	1	1	1	1	1	1
PLINK 5k BS 95% (0.025)	0.961 767	0.952	0.996 256	0.996 08	0.962	0.963 71
PLINK 5k BS 95% (0.975)	0.983 936	0.984	0.998 87	0.998 848	0.984	0.991 701
PLINK M 95% CI?	1	1	0	0	1	1
PLINK Mdn 95% CI?	1	1	1	1	1	1
t test paired stats	0.028 797	0.374 701	−0.211 063	0.156 564	$3.172\,73 \times 10^{-15}$	−0.449 793
t test pvalues	0.977 144	0.709 499	0.833 713	0.876 232	1	0.654 843
t test sig 5%?	0	0	0	0	0	0
t test sig 1%?	0	0	0	0	0	0
wilcoxon paired stats	33.5	11.5	312	306.5	36	28.5
wilcoxon pvalues	0.665 775	0.660 39	0.741 455	0.888 696	0.796 253 4	0.688 646
wilcoxon sig 5%?	0	0	0	0	0	0
wilcoxon sig 1%?	0	0	0	0	0	0
wilcoxon effect (r)	4.737 615	1.626 346	44.123 463	43.345 646	5.091 169	4.030 509
Cohen's d	0.001 191	0.019 789	0.006 192	0.004 531	0	0.015 893

Table A.2: 50-folds CV: Raw results from dataset ds1_1.

50-folds Cross-Validation	**f1**	**recall**	**average precision**	**roc auc**	**accuracy**	**precision**
LightGWAS mean	−0.028 533	−0.027 714	−0.002 709	−0.002 796	−0.028 102	−0.024 036
LightGWAS std	0.013 84	0.014 812	0.001 61	0.001 666	0.013 58	0.016 216
LightGWAS Mdn	−0.027 483	−0.033 957	−0.002 428	−0.002 561	−0.026 918	−0.018 685
LightGWAS skew	−0.031 128	−0.018 228	−0.171 534	−0.169 555	−0.033 926	−0.076 465
LightGWAS kurtosis	−0.673 704	−0.263 189	−0.737 99	−0.794 309	−0.649 081	−0.704 379
LightGWAS SE	0.001 957	0.002 095	0.000 228	0.000 236	0.001 92	0.002 293
PLINK mean	−0.029 364	−0.028 182	−0.002 743	−0.002 841	−0.028 986	−0.024 584
PLINK std	0.014 162	0.015 22	0.001 577	0.001 65	0.013 948	0.017 135
PLINK Mdn	−0.027 791	−0.034 186	−0.002 564	−0.002 571	−0.027 518	−0.025 516
PLINK skew	−0.026 962	−0.024 677	−0.144 426	−0.152 787	−0.029 652	−0.065 971
PLINK kurtosis	−0.737 852	−0.368 035	−0.673 915	−0.759 156	−0.721 024	−0.745 655
PLINK SE	0.002 003	0.002 152	0.000 223	0.000 233	0.001 973	0.002 423
LightGWAS Agostino stats	1.164 581	0.006 975	1.932 952	2.450 383	1.013 517	1.427 66
LightGWAS Agostino p-val	0.558 617	0.996 518	0.380 421	0.293 701	0.602 445	0.489 765
LightGWAS Agostino sig 5%?	0	0	0	0	0	0
LightGWAS Agostino sig 1%?	0	0	0	0	0	0
PLINK Agostino stats	1.637 456	0.085 853	1.370 528	2.055 321	1.502 753	1.741 151
PLINK Agostino p-val	0.440 992	0.957 982	0.503 957	0.357 843	0.471 717	0.418 711
PLINK Agostino sig 5%?	0	0	0	0	0	0
PLINK Agostino sig 1%?	0	0	0	0	0	0
LightGWAS $\geq$ PLINK (M)	1	1	1	1	1	1
MD	0.000 831	0.000 468	0.000 034	0.000 045	0.000 884	0.000 548
LightGWAS $\geq$ PLINK (Mdn)	1	1	1	1	1	1
MdnD	0.000 308	0.000 229	0.000 136	0.000 009	0.0006	0.006 831
LightGWAS 5k BS 95% (0.025)	−0.030 646	−0.043 222	−0.002 617	−0.002 733	−0.030 386	−0.030 964
LightGWAS 5k BS 95% (0.975)	−0.016 198	−0.015 45	−0.001 117	−0.001 148	−0.016 166	−0.011 452
LightGWAS M 95% CI?	1	1	0	0	1	1
LightGWAS Mdn 95% CI?	1	1	1	1	1	1
PLINK 5k BS 95% (0.025)	−0.033 055	−0.043 506	−0.002 457	−0.002 526	−0.032 977	−0.031 637
PLINK 5k BS 95% (0.975)	−0.015 104	−0.015 484	−0.000 988	−0.001 005	−0.015 066	−0.008 04
PLINK M 95% CI?	1	1	0	0	1	1
PLINK Mdn 95% CI?	1	1	0	0	1	1
t test paired stats	1.498 461	0.618 888	0.559 679	0.744 993	1.656 281	0.915 925
t test pvalues	0.140 429	0.538 857	0.578 248	0.459 835	0.104 056	0.364 192
t test sig 5%?	0	0	0	0	0	0
t test sig 1%?	0	0	0	0	0	0
wilcoxon paired stats	237	124	511	486	237	114
wilcoxon pvalues	0.000 316	0.000 007	0.574 894	0.409 13	0.000 307	0.000 068
wilcoxon sig 5%?	1	1	0	0	1	1
wilcoxon sig 1%?	1	1	0	0	1	1
wilcoxon sig 1%?	33.516 861	17.536 248	72.266 313	68.730 779	33.516 861	16.122 035
Cohen's d	0.059 381	0.031 137	0.021 258	0.027 091	0.064 197	0.032 854

Table A.3: 50-folds CV: Transformed (Box-Cox) raw results from dataset ds1_1.

50-folds Cross-Validation	**f1**	**recall**	**average precision**	**roc auc**	**accuracy**	**precision**
LightGWAS mean	0.993 251	0.993 75	0.999 829 6	0.998 281 3	0.987 727	0.992 842
LightGWAS std	0.005 909	0.009 193	0.000 272 074	0.002 738 358	0.010 729	0.008 637
LightGWAS Mdn	0.993 789	1	1	1	0.988 636	1
LightGWAS skew	−0.437 164	−1.399 523	−1.690 604	−1.684 362	−0.428 431	−0.761 914
LightGWAS kurtosis	−0.761 655	1.423 282	2.118 442	2.117 041	−0.779 717	−0.658 681
LightGWAS SE	0.000 836	0.0013	0.000 038 477	0.000 387 262	0.001 517	0.001 221
PLINK mean	0.991 394	0.993	0.999 671	0.996 718 8	0.984 318	0.989 887
PLINK std	0.006 772	0.009 161	0.000 485 632	0.004 748 225	0.012 34	0.010 093
PLINK Mdn	0.993 75	1	0.999 845 7	0.998 437 5	0.988 636	0.987 654
PLINK skew	−0.560 334	−1.204 521	−2.064 277	−1.891 769	−0.553 625	−0.728 877
PLINK kurtosis	−0.660 631	1.044 466	4.797 388	3.653 708	−0.699 161	−0.109 119
PLINK SE	0.000 958	0.001 296	$6.867\,87 \times 10^{-5}$	0.000 671 5	0.001 745	0.001 427
LightGWAS Agostino stats	3.702 605	17.397 629	23.4721	23.382 98	3.803 618	6.169 216
LightGWAS Agostino p-val	0.157 032	0.000 167	$8.000\,17 \times 10^{-6}$	8.3647×10^{-6}	0.149 298	0.045 748
LightGWAS Agostino sig 5%?	0	1	1	1	0	1
LightGWAS Agostino sig 1%?	0	1	1	1	0	0
PLINK Agostino stats	4.029 561	13.585 975	35.251 48	30.331 43	4.222 832	4.786 879
PLINK Agostino p-val	0.133 35	0.001 122	$2.214\,31 \times 10^{-8}$	$2.591\,88 \times 10^{-7}$	0.121 066	0.091 315
PLINK Agostino sig 5%?	0	1	1	1	0	0
PLINK Agostino sig 1%?	0	1	1	1	0	0
LightGWAS ≥ PLINK (M)?	1	1	1	1	1	1
MD	0.001 857	0.000 75	0.000 158 663	0.001 562 5	0.003 409	0.002 955
LightGWAS ≥ PLINK (Mdn)?	1	1	1	1	1	1
MdnD	0.000 039	0	0.000 154 321	0.001 562 5	0	0.012 346
LightGWAS 5k BS 95% (0.025)	0.987 562	0.99	0.999 461 7	0.994 75	0.977 273	0.980 344
LightGWAS 5k BS 95% (0.975)	0.996 255	1	0.999 925 5	0.999 25	0.993 182	0.995
LightGWAS M 95% CI?	1	1	1	1	1	1
LightGWAS Mdn 95% CI?	1	1	0	0	1	0
PLINK 5k BS 95% (0.025)	0.985	0.985	0.999 182 7	0.991 937 5	0.972 727	0.980 247
PLINK 5k BS 95% (0.975)	0.993 789	0.9975	0.999 863	0.998 625	0.988 636	0.992 537
PLINK M 95% CI?	1	1	1	1	1	1
PLINK Mdn 95% CI?	1	0	1	1	1	1
t test paired stats	2.364 684	0.596 12	2.786 933	2.829 582	2.393 172	3.060 335
t test pvalues	0.022 051	0.553 839	0.007 549 595	0.006 737 659	0.020 579	0.003 581
t test sig 5%?	1	0	1	1	1	1
t test sig 1%?	0	0	1	1	0	1
wilcoxon paired stats	183.5	113.5	54	48.5	141.5	37.5
wilcoxon pvalues	0.131 427	0.662 096	0.002 023 744	0.006 190 166	0.022 961	0.006 574
wilcoxon sig 5%?	0	0	1	1	1	1
wilcoxon sig 1%?	0	0	1	1	0	1
wilcoxon effect (r)	25.950 819	16.051 324	7.636 753	6.858 936	20.011 122	5.303 301
Cohen's d	0.292 229	0.081 727	0.403 093 2	0.403 138 6	0.294 84	0.314 576

Table A.4: 50-folds CV: Raw results from dataset ds1_10.

50-folds Cross-Validation	**f1**	**recall**	**average precision**	**roc auc**	**accuracy**	**precision**
LightGWAS mean	−0.005 377	−0.002 948 878	−0.000 071	−0.000 721	−0.009 824	−0.004 344 802
LightGWAS std	0.004 365	0.003 852 838	0.000 093	0.000 942	0.007 978	0.004 881 851
LightGWAS Mdn	−0.005 485	0	0	0	−0.010 055	0
LightGWAS skew	−0.093 928	−0.576 890 2	−0.690 782	−0.689 184	−0.094 276	−0.328 988 1
LightGWAS kurtosis	−1.157 434	−1.540 471	−1.255 568	−1.260 038	−1.160 249	−1.643 616
LightGWAS SE	0.000 617	0.000 544 874	0.000 013	0.000 133	0.001 128	0.000 690 398
PLINK mean	−0.006 505	−0.003 917 629	−0.000 159	−0.001 601	−0.011 875	−0.007 297 474
PLINK std	0.004 438	0.004 569 232	0.000 174	0.001 741	0.008 109	0.006 578 759
PLINK Mdn	−0.005 463	0	−0.000 133	−0.001 342	−0.009 947	−0.009 968 196
PLINK skew	−0.100 787	−0.395 724	−0.531 144	−0.522 888	−0.103 597	−0.147 752 6
PLINK kurtosis	−0.912 839	−1.600 168	−1.230 784	−1.242 435	−0.918 638	−1.233 779
PLINK SE	0.000 628	0.000 646 187	0.000 025	0.000 246	0.001 147	0.000 930 377
LightGWAS Agostino stats	10.038 052	54.890 58	18.980 926	19.228 235	10.149 779	96.399 34
LightGWAS Agostino p-val	0.006 611	1.2041×10^{-12}	0.000 076	0.000 067	0.006 252	$1.167\,21 \times 10^{-21}$
LightGWAS Agostino sig 5%?	1	1	1	1	1	1
LightGWAS Agostino sig 1%?	1	1	1	1	1	1
PLINK Agostino stats	3.787 901	73.781 37	15.974 62	16.529 008	3.886 607	13.676 95
PLINK Agostino p-val	0.150 476	$9.518\,72 \times 10^{-17}$	0.000 34	0.000 257	0.143 23	0.001 071 735
PLINK Agostino sig 5%?	0	1	1	1	0	1
PLINK Agostino sig 1%?	0	1	1	1	0	1
LightGWAS $\geq$ PLINK (M)	1	1	1	1	1	1
MD	0.001 128	0.000 968 751	0.000 088	0.000 88	0.002 05	0.002 952 672
LightGWAS $\geq$ PLINK (Mdn)	0	1	1	1	0	1
MdnD	0.000 022	0	0.000 133	0.001 342	0.000 108	0.009 968 196
LightGWAS 5k BS 95% (0.025)	−0.010 322	−0.007 261 454	−0.000 277	−0.002 834	−0.018 81	−0.015 558 13
LightGWAS 5k BS 95% (0.975)	−0.003 537	0	−0.000 067	−0.000 68	−0.006 435	−0.004 703 822
LightGWAS M 95% CI?	1	1	1	1	1	0
LightGWAS Mdn 95% CI?	1	1	0	0	1	0
PLINK 5k BS 95% (0.025)	−0.012 207	−0.011 131 11	−0.000 405	−0.004 112	−0.022 255	−0.015 367 45
PLINK 5k BS 95% (0.975)	−0.005 694	−0.002 373 796	−0.000 12	−0.001 209	−0.010 426	−0.006 771 662
PLINK M 95% CI?	1	1	1	1	1	1
PLINK Mdn 95% CI?	0	0	1	1	0	1
t test paired stats	2.078 324	1.536 095	4.483 247	4.467 655	2.073 969	4.600 788
t test pvalues	0.042 939	0.130 947 6	0.000 044	0.000 047	0.043 359	$2.995\,36 \times 10^{-5}$
t test sig 5%?	1	0	1	1	1	1
t test sig 1%?	0	0	1	1	0	1
wilcoxon paired stats	355	129	38	38	360	37
wilcoxon pvalues	0.153 842	0.090 138 26	0.000 103	0.000 103	0.171 386	$3.440\,88 \times 10^{-5}$
wilcoxon sig 5%?	0	0	1	1	0	1
wilcoxon sig 1%?	0	0	1	1	0	1
wilcoxon effect (r)	50.204 581	18.243 355	5.374 012	5.374 012	50.911 688	5.232 59
Cohen's d	0.256 334	0.229 222 8	0.633 073	0.628 972	0.254 902	0.509 716 2

Table A.5: 50-folds CV: Transformed (Box-Cox) raw results from dataset ds1_10.

50-folds Cross-Validation	**f1**	**recall**	**average precision**	**roc auc**	**accuracy**	**precision**
LightGWAS mean	0.997 205	0.9986	0.999 857	0.987	0.994 455	0.995 83
LightGWAS std	0.003 806	0.004 522 055	0.000 564 904	0.048 497 84	0.007 527	0.004 951
LightGWAS Mdn	1	1	1	1	1	1
LightGWAS skew	−1.505 589	−3.267 018	−5.734 615	−5.505 747	−1.484 58	−0.324 387
LightGWAS kurtosis	2.256 867	9.746 949	33.536 16	31.148 27	2.169 348	−1.894 484
LightGWAS SE	0.000 538	0.000 639 515	$7.988\,94 \times 10^{-5}$	0.006 858 631	0.001 064	0.0007
PLINK mean	0.996 713	0.9994	0.999 822 7	0.9826	0.993 465	0.994 053
PLINK std	0.002 956	0.002 398 979	0.000 303 524	0.029 263 75	0.005 869	0.004 905
PLINK Mdn	0.995 025	1	1	1	0.990 099	0.990 099
PLINK skew	−0.261 505	−3.705 468	−2.295 627	−2.218 764	−0.250 315	0.408 179
PLINK kurtosis	−0.642 297	11.7305	4.997 979	4.598 109	−0.668 302	−1.833 295
PLINK SE	0.000 418	0.000 339 267	$4.292\,48 \times 10^{-5}$	0.004 138 52	0.000 83	0.000 694
LightGWAS Agostino stats	21.348 364	58.330 86	99.222 67	96.246 02	20.806 61	1461.396 217
LightGWAS Agostino p-val	0.000 023	$2.155\,83 \times 10^{-13}$	$2.844\,93 \times 10^{-22}$	$1.260\,21 \times 10^{-21}$	0.000 03	0
LightGWAS Agostino sig 5%?	1	1	1	1	1	1
LightGWAS Agostino sig 1%?	1	1	1	1	1	1
PLINK Agostino stats	1.654 855	65.6021	38.701 19	36.850 01	1.755 615	4711.908 093
PLINK Agostino p-val	0.437 172	5.6844×10^{-15}	$3.945\,87 \times 10^{-9}$	$9.956\,84 \times 10^{-9}$	0.415 693	0
PLINK Agostino sig 5%?	0	1	1	1	0	1
PLINK Agostino sig 1%?	0	1	1	1	0	1
LightGWAS ≥ PLINK (M)?	1	0	1	1	1	1
MD	0.000 492	0.0008	$3.432\,68 \times 10^{-5}$	0.0044	0.000 99	0.001 776
LightGWAS ≥ PLINK (Mdn)?	1	1	1	1	1	1
MdnD	0.004 975	0	0	0	0.009 901	0.009 901
LightGWAS 5k BS 95% (0.025)	0.994 024	0.996	0.999 624	0.964	0.988 119	0.990 079
LightGWAS 5k BS 95% (0.975)	0.997 009	1	0.999 984	0.9984	0.994 059	0.996 008
LightGWAS M 95% CI?	0	1	1	1	0	1
LightGWAS Mdn 95% CI?	0	1	0	0	0	0
PLINK 5k BS 95% (0.025)	0.994	0.994	0.999 329 8	0.9376	0.988 119	0.990 079
PLINK 5k BS 95% (0.975)	0.997 009	1	0.999 851 4	0.9852	0.994 059	0.994 036
PLINK M 95% CI?	1	1	1	1	1	0
PLINK Mdn 95% CI?	1	1	0	0	1	1
t test paired stats	0.647 493	−1.158 648	0.378 819 2	0.549 943 9	0.658 505	1.691
t test pvalues	0.520 335	0.252 215 4	0.706 458 6	0.584 856	0.513 296	0.097 189
t test sig 5%?	0	0	0	0	0	0
t test sig 1%?	0	0	0	0	0	0
wilcoxon paired stats	183	5	163.5	166.5	180	176
wilcoxon pvalues	0.430 596	0.234 194 3	0.095 637 79	0.107 380 9	0.387 66	0.342 925
wilcoxon sig 5%?	0	0	0	0	0	0
wilcoxon sig 1%?	0	0	0	0	0	0
wilcoxon effect (r)	25.880 108	0.707 107	23.122 392	23.546 656	25.455 844	24.890 159
Cohen's d	0.144 224	0.221 014 4	0.075 700 46	0.109 855 8	0.146 695	0.360 441

Table A.6: 50-folds CV: Raw results from dataset ds1_100.

50-folds Cross-Validation	**f1**	**recall**	**average precision**	**roc auc**	**accuracy**	**precision**
LightGWAS mean	−0.001 527 229	−0.000 141 542	−0.000 034	−0.003 485	−0.003 044 371	−0.002 676 829
LightGWAS std	0.001 782 568	0.000 428 937	0.000 065	0.006 621	0.003 552 989	0.003 177 622
LightGWAS Mdn	0	0	0	0	0	0
LightGWAS skew	−0.406 215 6	−2.666 668	−1.592 403	−1.592 973	−0.405 472 7	−0.324 249 4
LightGWAS kurtosis	−1.558 453	5.111 118	1.109 876	1.112 696	−1.560 344	−1.894 763
LightGWAS SE	0.000 252 093	$6.066\,09 \times 10^{-5}$	0.000 009	0.000 936	0.000 502 469	0.000 449 384
PLINK mean	−0.002 896 239	$-3.618\,12 \times 10^{-5}$	−0.000 076	−0.007 666	−0.005 787 792	−0.011 760 04
PLINK std	0.002 534 869	0.000 144 663	0.000 092	0.009 325	0.005 065 214	0.009 699 853
PLINK Mdn	−0.004 465 761	0	0	0	−0.008 924 837	−0.019 565 7
PLINK skew	−0.045 691 42	−3.705 468	−0.680 763	−0.680 856	−0.045 028 59	0.408 034 6
PLINK kurtosis	−1.116 007	11.7305	−1.072 171	−1.071 976	−1.117 365	−1.833 212
PLINK SE	0.000 358 485	$2.045\,85 \times 10^{-5}$	0.000 013	0.001 319	0.000 716 329	0.001 371 766
LightGWAS Agostino stats	58.598 32	43.626 26	19.070 126	19.086 572	59.181 44	1457.907
LightGWAS Agostino p-val	$1.885\,98 \times 10^{-13}$	$3.362\,62 \times 10^{-10}$	0.000 072	0.000 072	$1.409\,01 \times 10^{-13}$	0
LightGWAS Agostino sig 5%?	1	1	1	1	1	1
LightGWAS Agostino sig 1%?	1	1	1	1	1	1
PLINK Agostino stats	8.467 788	65.6021	11.295 258	11.290 75	8.512 637	4611.49
PLINK Agostino p-val	0.014 495 83	5.6844×10^{-15}	0.003 526	0.003 534	0.014 174 39	0
PLINK Agostino sig 5%?	1	1	1	1	1	1
PLINK Agostino sig 1%?	0	1	1	1	0	1
LightGWAS ≥ PLINK (M)	1	0	1	1	1	1
MD	0.001 369 009	0.000 105 361	0.000 041	0.004 181	0.002 743 421	0.009 083 211
LightGWAS ≥ PLINK (Mdn)	1	1	1	1	1	1
MdnD	0.004 465 761	0	0	0	0.008 924 837	0.019 565 7
LightGWAS 5k BS 95% (0.025)	−0.008 252 685	−0.001 627 458	−0.000 144	−0.014 33	−0.016 492 93	−0.035 706 38
LightGWAS 5k BS 95% (0.975)	−0.003 501 117	0	−0.000 015	−0.001 523	−0.006 968 625	−0.006 382 508
LightGWAS M 95% CI?	0	1	1	1	0	0
LightGWAS Mdn 95% CI?	0	1	0	0	0	0
PLINK 5k BS 95% (0.025)	−0.003 949 765	−0.004 090 314	−0.000 293	−0.029 099	−0.007 831 213	−0.001 817 311
PLINK 5k BS 95% (0.975)	−0.002 408 406	0	−0.000 12	−0.012 098	−0.004 784 587	−0.001 756 223
PLINK M 95% CI?	1	1	0	0	1	0
PLINK Mdn 95% CI?	0	1	0	0	0	0
t test paired stats	2.899 631	−1.753 387	2.517 852	2.516 754	2.910 195	6.054 518
t test pvalues	0.005 577 558	0.085 789 48	0.015 125	0.015 167	0.005 419 658	$1.928\,75 \times 10^{-7}$
t test sig 5%?	1	0	1	1	1	1
t test sig 1%?	1	0	0	0	1	1
wilcoxon paired stats	199	3	123	123	199	55
wilcoxon pvalues	0.004 148 562	0.061 569 84	0.014 121	0.014 121	0.004 148 562	$1.412\,84 \times 10^{-6}$
wilcoxon sig 5%?	1	0	1	1	1	1
wilcoxon sig 1%?	1	0	0	0	1	1
wilcoxon effect (r)	28.142 85	0.424 264	17.394 827	17.394 827	28.142 85	7.778 175
Cohen's d	0.624 763 4	0.329 160 4	0.517 258	0.517 021	0.627 076 9	1.258 499

Table A.7: 50-folds CV: Transformed (Box-Cox) raw results from dataset ds1_100.