Imperial College London

IMPERIAL COLLEGE LONDON

FACULTY OF MEDICINE

Interpretable machine learning models for investigating and pre-screening cardiomyopathies with high-dimensional genomic and imaging data

Author: Melpomeni Kasapi

Supervised by: Dr. Joram M. Posma, Prof. James S. Ware, Prof. Timothy M. D. Ebbels

Submitted in part fulfillment of the requirements for the degree of Doctor of Philosophy of Imperial College London and the Diploma of Imperial College London

November 2023

Declaration of Originality

I, Melpomeni Kasapi, declare that any work presented in this thesis is my own and has not been duplicated from the work of any other authors, journals, or institutions. Any work that does originate from other sources is acknowledged and referenced accordingly.

Copyright Statement

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International License (CC BY-NC).

Under this license, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose.

When reusing or sharing this work, ensure you make the license terms clear to others by naming the license and linking to the license text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes.

Please seek permission from the copyright holder for uses of this work that are not included in this license or permitted under UK Copyright Law.

Abstract

Cardiomyopathies have long been considered as monogenic, Mendelian diseases with rare variant implications; however, as more studies look at the common variation in these low prevalence diseases, additional discoveries of the polygenic interplay are being revealed. Part of these discoveries also associate genomics with phenotype, specifically derived from cardiac magnetic resonance (CMR) imaging, and with the use of machine learning (ML), insights to the complex diversity of these diseases are revealed. The overall purpose of this thesis is to leverage the information hidden in genomic and imaging datasets with the use of ML to either 1) develop new methodologies for handling such complex datatypes and 2) create ML models that can potentially predict or generate risk scores for cardiomyopathies, based on genomic, clinical and imaging data. This is achieved by developing LAVASET, a novel ensemble algorithm that enhances the Random Forest algorithm and is able to mitigate variable correlation issues. In addition, traditional and interpretable ML algorithms are used to pre-screen and stratify cases of cardiomyopathy, while also displaying increased model performance when using LAVASET-derived imaging features. Finally, a deep-learning method for calculating Polygenic Risk Scores (PRS) is described, showing promising associations between high PRS and disease risk as well as high PRS and known cardiomyopathy outcomes and traits.

Acknowledgments

I would first like to start by thanking Mike Sternberg and the administrators from the MSc in Bioinformatics and Systems Biology for their support and liaising with the Wellcome Trust on my behalf. Especially during the COVID lockdown, communications were clear and timely and all necessary extensions were processed in time, so as to help us continue our work seamlessly.

I would then like to thank Sean Zheng, Pantazis Theotokis, and Kathryn McGurk from the Cardiovascular Genetics group for their always helpful suggestions, their willingness to share data, and sit down with me to discuss my methodology. I would specifically like to thank Kathryn, who provided me with valuable insights and genomic knowledge when I was lacking and learning.

My Progress Review Panel, Dr Sarah Filippi and Prof Robert Glen, helped me immensely with their always thoughtful comments and feedback, and I am very grateful to them for serving as my PRP, and making the various milestone processes seamless and helpful.

Then I would like to deeply thank Prof Tim Ebbels, who acted as my second theoretical supervisor on this project. Tim, although a very busy colleague, always had time to discuss, give valuable feedback and provide a higher-level and comprehensive context to the methodologies, without which this project would have not turned out this way.

I'm also extremely thankful to Prof James Ware, my experimental supervisor. Apart from the very logistical aspects of data providing and facilitating my data acquisition by connecting me to the right people, James also provided me with deeply insightful feedback and ideas on the genomics and translational aspects of the project. He was always available to discuss and explain genomics and cardiovascular concepts, while also prompting me to think further and consider new ideas.

Finally, I would like to say a very BIG thank you to my main supervisor, Dr Joram Posma. Joram was there from day 1 to think through, design, and help me succeed through the project; and, needless to say without his continuous guidance and support I wouldn't have reached this final step. Especially during COVID lockdowns and the uncertainty of when we could come back to campus, Joram was always understanding, supportive and went above and beyond to ensure that I had the right resources to deal with the situation and succeed. I'm immensely grateful for meeting him as my supervisor in my MSc final project, and continuing to work with him during the PhD.

Finally, I would like to acknowledge that all 3 of my supervisors have been amazingly supportive throughout this journey. Whether that was highlighting the importance of me seeking guidance from other women in science and understanding the complexities of that, to providing a safe 'virtual' environment throughout the COVID period and lockdowns, which spanned more than half of my PhD duration.

Publications and Contributions

† indicates equal contributions/co-first authorship.

Kasapi M, Xu K, Ebbels TMD, O'Regan DP, Ware JS, Posma JM. LAVASET: Latent Variable Stochastic Ensemble of Trees. An ensemble method for correlated datasets with spatial, spectral, and temporal dependencies. (*requested revision submitted (19 January 2024) to Bioinformatics (Oxford University Press)*) Pre-print DOI: 10.1101/2023.10.20.563223 LAVASET is freely available as a standalone package from https://github.com/melkasapi/LAVASET.

Mujagic Z†, **Kasapi M**†, Jonkers DM, Garcia-Perez I, Vork L, Weerts ZZRM, Serrano-Contreras JI, Zhernakova A, Kurilshikov A, Scotcher J, Holmes E, Wijmenga C, Keszthelyi D, Nicholson JK, Posma JM, Masclee AA. Integrated fecal microbiome-metabolome signatures reflect stress and serotonin metabolism in irritable bowel syndrome. Gut Microbes, 2022 DOI: 10.1080/19490976.2022.2063016

McGurk KA, **Kasapi M**, Ware JS. Effect of taurine administration on symptoms, severity, or clinical outcome of dilated cardiomyopathy and heart failure in humans: a systematic review. Wellcome Open Research, 2022 DOI: 10.12688/wellcomeopenres.17505.3

Kasapi M, McGurk KA, Ebbels TMD, Posma JMP, Ware JS. Interpretable machine learning models for pre-screening cardiomyopathies.

American Society of Human Genetics Annual Meeting 2023. Washington, DC, USA.

https://www.ashg.org/wp-content/uploads/2023/10/ASHG2023-PosterAbstracts.pdf

Kasapi M, Xu K, Ebbels TMD, O'Regan DP, Ware JS, Posma JM. LAVASET: Latent Variable Stochastic Ensemble of Trees. A novel ensemble method for correlated datasets.

European Conference on Computational Biology 2022. Sitges, Spain.

https://eccb2022.org/poster-presentations/

Contents

1 Introduction to Cardiomyopathies **1**

1.1 Overview of cardiomyopathies 2

1.1.1 Definition, Diagnosis, Epidemiology 2

1.2 Genetic architecture 6

1.2.1 Rare variants 6

1.2.2 Common variants 9

1.3 Cardiomyopathies and Machine Learning 13

1.3.1 Cardiomyopathies and imaging with ML modeling 13

1.4 Genetics of cardiomyopathies in clinic 21

1.5 Hypothesis and Aims 23

1.6 Scope of this thesis 23

2 Introduction to Machine Learning **25**

2.1 Introduction 26

2.2 Statistical Testing 26

2.3 Unsupervised Learning 28

2.4 Supervised Learning 28

2.4.1 Random Forest 28
2.4.2 Gradient Boosting Trees 30
2.4.3 Support Vector Machines 31
2.4.4 Multi-Layer Feed-Forward Neural Networks 33
2.4.5 Metrics 34

3 Introduction to the Dataset and Assessment of -Omics Sub-groups with Unsupervised Learning 37

3.1 Introduction 38
3.2 Datasets 38
3.2.1 UK Biobank (UKBB) 39
3.2.2 Royal Brompton Hospital Cardiovascular Biobank 41
3.3 Methods 44
3.3.1 Common Variants 44
3.3.2 Rare Variants 45
3.3.3 CMR imaging 45
3.3.4 Unsupervised Pipeline 45
3.4 Results 48
3.4.1 HCM UKBB 48
3.4.2 HCM RBH Cohort 49
3.4.3 DCM UKBB 49
3.4.4 Clusters in common SNPs are attributed to correlation 50
3.4.5 3D-CMR Imaging 51
3.5 Discussion and Future Work 52

3.6 Conclusion 54

4 LAVASET: A Novel Ensemble Method for Correlated Datasets 56

4.1 Introduction 57

4.2 Methods 59

4.2.1 Datasets 59

4.2.2 Approach 60

4.2.3 Feature importance calculation 61

4.2.4 Model evaluation 64

4.3 Results 65

4.3.1 CMR Imaging 65

4.3.2 ECG 68

4.4 Discussion and Future Work 69

4.5 Conclusion 72

5 Pre-Screening Cardiomyopathies with Machine Learning 74

5.1 Introduction 75

5.2 Methods 76

5.2.1 Dataset and basic model architecture 76

5.2.2 Support Vector Machines 77

5.2.3 Random Forest 79

5.2.4 Gradient-Boosted Trees (GB-Trees) 80

5.2.5 Genomic Variables - Rare Variants 81

5.2.6 HCM Models and LAVASET-derived Imaging Features 84

5.3 Results 86
5.3.1 UKBB Feature Evaluation 86
5.3.2 SVM-linear models can distinguish best between HCM and Controls . . 89
5.3.3 Important features in HCM models include genetic and clinical characteristics 91
5.3.4 SVM models can distinguish best between DCM and Controls 94
5.3.5 Important features in DCM models include genetic and clinical characteristics 95
5.3.6 HCM and DCM P/LP variants and common clinical factors show importances in both phenotypes 99
5.3.7 HCM models including LAVASET-derived CMR imaging features perform better 101
5.4 Discussion and Future Work 103
5.4.1 ML models can differentiate between CMs and Controls 103
5.4.2 Models are interpretable and important features can be evaluated . . . 106
5.4.3 HCM models including LAVASET-derived CMR features perform better 111
5.5 Conclusion 112

6 Deep Learning Models for Evaluating PRS in Cardiomyopathies 114

6.1 Introduction 115
6.2 Methods 116
6.2.1 SNP Filtering 117
6.2.2 Neural Network Architecture 118
6.2.3 Dataset and Training/Validation/Test Splits 120

6.2.4 Statistical Analysis 122
6.3 Results 123
6.3.1 PRS can separate between HCM cases and Controls 124
6.3.2 HCM PRS groups show differences in left ventricular phenotypes . . . 125
6.3.3 PRS can separate between DCM cases and Controls 127
6.3.4 DCM PRS groups show differences in imaging and clinical features . . 128
6.4 Discussion and Future Work 134
6.5 Conclusion 140

7 Concluding Remarks **142**
7.1 Concluding Remarks 143
7.2 Appendix 147

Chapter 1

Introduction to Cardiomyopathies

1.1 Overview of cardiomyopathies

1.1.1 Definition, Diagnosis, Epidemiology

Definition and diagnosis

Inherited cardiac conditions (ICCs) affect more than 600,000 patients in the UK ($\tilde{1}$% of the population). Cardiomyopathy— a group of primary heart muscle diseases characterized by abnormal function, and subclassified according to morphological appearance — is the most common cause of sudden death in young adults, and the leading cause of heart transplantation (Braunwald, 2017). By ESC guidelines (Arbelo et al., 2023), a cardiomyopathy is defined as 'a myocardial disorder in which the heart muscle is structurally and functionally abnormal, in the absence of coronary artery disease (CAD), hypertension, valvular disease, and congenital heart disease (CHD) sufficient to cause the observed myocardial abnormality'. Cardiomyopathies can simultaneously occur with ischemic, valvular, and hypertensive diseases and the existence of one does not negate the potential presence of the others.

The two most common inherited cardiomyopathies are Hypertrophic Cardiomyopathy (HCM) and Dilated Cardiomyopathy (DCM). HCM is characterized by unexplained thickening of the muscle wall of the heart's left ventricle. It is often called the disease of the sarcomere— the part of the muscle responsible for providing the molecular force for contraction (Sabater-Molina et al., 2018) — due to genetic variants in sarcomere genes that are associated with HCM. DCM is primarily diagnosed by the left ventricular dilation and impaired contraction of the heart muscle that cannot be attributed to other conditions like hypertension (increased blood pressure) (Schultheiss et al., 2019). The most common cause of the DCM phenotype is ischemia, derived from volume overload often due to valve incompetence.

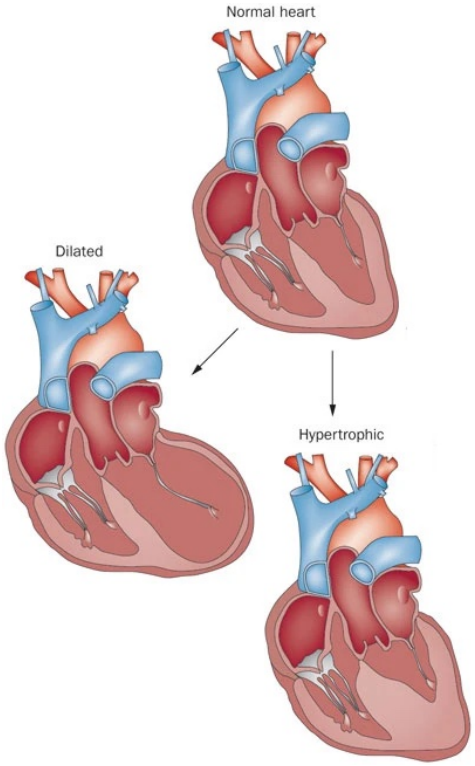


Figure 1.1: Graphical schematic of the left ventricular phenotypes for HCM and DCM. Left ventricular wall thickness for HCM and dilation for DCM are evident in comparison to the normal heart. Adapted from Hershberger et al. (2013). Copyright license found in Appendix 7.2.

Cardiomyopathies intervene with the normal function of the cardiac muscle; systole, the state of contraction of the muscle, and diastole, the state of relaxation, when affected can impair the heart's function to pump blood and cause a variety of symptoms. However, both HCM and DCM have high phenotypic variability that is still poorly understood (Verdonschot et al., 2019). Disease severity can vary hugely even in a single family, where the same genetic cause is shared by all affected; it is also possible that some of the members carrying the genetic predisposition might not manifest the disease at all. This presents a clinical challenge since we need to target costly and burdensome advance therapies to the appropriate subgroups. To tackle such a variability in phenotypes additional tests have been introduced in clinical practice which aim to personalize patient profiles for tailoring their diagnosis and treatment. Genetic testing has been one of the recommended routes for assessment of ICCs, with medical imaging and specifically cardiac magnetic resonance (CMR) being a second one (Kramer, 2015).

Epidemiology

Present estimates for the prevalence of HCM (1/500) and DCM (1/250) are derived from screening studies, primarily among healthy young adults. These figures likely under-represent the actual prevalence of the disease and the potential for disease-related issues (McKenna and Judge, 2021). This is due to the wide phenotypic variability found across cases which makes it difficult to give exact diagnoses. HCM can be characterized by specific diagnostic markers, like left ventricular hypertrophy (LVH) of $\geq$ 13mm, with changes present in genes linked to sarcomeric proteins. However, this same marker can appear in various genetic and non-genetic scenarios. Moreover, post-mortem examinations of sudden cardiac deaths in patients with

unexplained LVH often present diagnostic puzzles (Arbelo et al., 2023). In the case of DCM, research indicates that most individuals meeting the diagnostic criteria show no symptoms and usually do not seek medical care (Redfield et al., 2003, Wang et al., 2003, Yeboah et al., 2012). Another common issue is that DCM diagnosis can be often missed in patients with end stage heart failure (Seidelmann et al., 2016).

Initial studies indicated that HCM predominantly affected young males and older females. This perception aligns with demographic information from patient group studies, increased instances of sudden cardiac death in males with HCM, and screening study results that highlight age-related disease manifestation and more pronounced disease in males (Finocchiaro et al., 2019, Ho et al., 2018, Maron et al., 2000). These findings appear uniform across individuals of European, Native American, African, or Asian descent (McKenna and Judge, 2021).

Initial estimates of DCM prevalence and incidence originate from a study in Olmstead County, Minnesota, USA, spanning 1975-1984. This study found a DCM prevalence of around 1 in 2,700 people and an incidence rate of 6 per 100,000 person-years (Codd et al., 1989). However, this same study significantly underestimated HCM's prevalence, a fact revealed a decade later (Maron et al., 1995). Using this data and additional findings, Hershberger et al. (2013) and colleagues adjusted the estimated prevalence of DCM to about 1 in 250 people. They also showed that DCM accounted for nearly 1 in 6 heart failure cases in the USA. Although DCM can impact anyone, it appears more frequently in individuals of African descent as opposed to European (Coughlin et al., 1993), is more prevalent in men than women (3:1 ratio) (Fairweather et al., 2023), and typically shows symptoms in adults aged between 30 and 60 (Hershberger et al., 2013).

1.2 Genetic architecture

Cardiomyopathies are treated as monogenic diseases in clinical practice, however recent findings described below suggest that part of the cases can be attributed to a more complex genetic architecture that involves multiple variants. Some phenotypes still behave as monogenic and would have a single variant as the main risk factor, while for these cases polygenic determinants can contribute to penetrance and expressivity. Others do not carry a single rare variant and are completely polygenic. Efforts in genomic variant interpretation have identified clinically actionable disease-causing variants.

1.2.1 Rare variants

Based on the ClinGen gene curation clinical validity framework (Strande et al., 2017), out of 33 genes previously reported as associated with HCM, 8 were classified as definitive, 3 as moderate, 16 as limited, and 6 showed no evidence. Definitive genes like *MYBPC3*, *MYH7*, *TNNT2*, *TNNI3*, *TPM1*, *ACTC1*, *MYL3*, and *MYL2* have been recognized in diagnostic panels for years and were identified primarily due to genetic evidence indicating their link to HCM from linkage studies in families. Pathogenic variants in HCM are mostly found in these definitive sarcomere related genes. The majority of HCM cases exhibit rare variants in two primary HCM genes (*MYBPC3* in 19% of cases and *MYH7* in 14.2% of cases) (Ho et al., 2015). These two genes encode the myosin-binding protein C and the myosin heavy chain beta (MHC-β) isoform protein, respectively (Sabater-Molina et al., 2018). Both these proteins play a role in the structure and function of the sarcomere. Other well-studied HCM genes (*TNNI3*, *TNNT2*, *TPM1*, *MYL2*, *MYL3*, *ACTC1*, *PLN*) and genes that mimic the phenotype (*GLA*,

LAMP2, *PRKAG2*) exhibited rare variants in 2% or fewer cases per gene. For each gene, there is a significant difference ($p < 0.05$ after adjusting for multiple comparisons) in variation between cases and the Genome Aggregation Database (gnomAD, Gudmundsson et al. (2022)), thereby reinforcing their correlation with the disease (Walsh et al., 2017).

Variants in at least 50 genes have been reported to play a role in DCM (Tayal et al., 2017). Evaluation of genetic and experimental data identified 19 genes with significant evidence, indicating a role in monogenic DCM. Interestingly, over half of the analyzed genes were found to have limited evidence, no disease relationship or only animal-model evidence (Jordan et al., 2021). Specifically, 12 genes spanning 8 gene ontologies, were categorized as having definitive (*BAG3*, *DES*, *FLNC*, *LMNA*, *MYH7*, *PLN*, *RBM20*, *SCN5A*, *TNNC1*, *TNNT2*, *TTN*) or strong (*DSP*) evidence. Seven genes (*ACTC1*, *ACTN2*, *JPH2*, *NEXN*, *TNNI3*, *TPM1*, *VCL*) covering 2 additional ontologies, were identified as having moderate evidence. These genes are anticipated to upgrade to either strong or definitive categories as more evidence becomes available. About six of these strong and moderate evidence genes also show similar associations to HCM (Jordan et al., 2021), however when genes are linked with both types of cardiomyopathies, the molecular mechanisms underlying the pathogenesis are almost always distinct (Kimura, 2015).

In contrast to HCM, DCM genetic architecture is attributed and spanning more than 10 gene ontologies. Many of these genes produce proteins for the sarcomere. Genes producing Z-disk proteins are also predominant in DCM which is explained by the crucial role they play in the sarcomere's mechanosensation. Other genes are linked to proteins associated with calcium or potassium channels, the nuclear envelope, heat-shock chaperones, and mitochondria (Hershberger et al., 2013). The precise reason for this varied genetic structure in the onset of

DCM has not been fully explained. However, DCM can be seen as a terminal phenotype that arises when the cellular mechanisms ensuring contraction force or ventricular stability are disturbed due to pathogenic variations in genes producing essential proteins in these pathways (Hershberger et al., 2010).

The percentage of cases with a certain genetic diagnosis, meaning the identification of a probable causal Mendelian genetic variant, is quite low, at about 40% for HCM and approximately 30% for DCM. In HCM, two multi-center studies have shown that HCM manifestation related to sarcomeric rare variants results in more severe outcomes than non-sarcomeric-related HCM. Patients with proven pathogenic sarcomere mutations showed symptoms at a younger age and experienced more and earlier HCM-related issues. The likelihood of encountering the combined outcome by age 50 was 29% for patients with sarcomere mutations (SARC+) compared to 14% for those without (SARC−). Even when accounting for a younger age at initial presentation, the presence of sarcomere mutations significantly elevated the risk (Ho et al., 2015). These findings were corroborated by the Hypertrophic Cardiomyopathy Registry, by associating genotype with phenotypic features via cardiac magnetic resonance imaging. SARC+ individuals often displayed a reverse septal curvature morphology, had a higher presence of late gadolinium enhancement (LGE), but showed less resting left ventricular outflow (LVOT) obstruction. SARC− individuals typically exhibited isolated basal septal hypertrophy, had less LGE, but more LVOT obstruction. The former group represents the inherited familial form of HCM, while the latter is believed to be the result of a combination of factors, as indicated by the increased prevalence of secondary LV hypertrophy causes such as hypertension, high BMI, being male, advanced age, among others (Neubauer et al., 2019).

In DCM, genotype positive individuals (those that carry a rare Mendelian variant) display

a higher occurrence of major cardiovascular incidents, advanced heart failure, and significant ventricular arrhythmias. They also show reduced left ventricle reverse remodeling compared to their non-carrier peers (Escobar-Lopez et al., 2021). Ten-year survival between the two groups does not show major differences, indicating that comorbidities and aging can alter the effect of the genotype (Gigli et al., 2019).

Finally, phenotypic variability in cardiomyopathies can arise from incomplete penetrance of rare pathogenic variants. This explains how not every individual with a causative variation will display the disease symptoms, and among those who do, the age of onset and disease intensity can vary widely. So, while some may experience severe conditions requiring heart transplants early in life, others might not show any symptoms their entire life or only be slightly affected (Lorenzini et al., 2020, McGurk et al., 2023). This variability can arise from differences in causative variations, the influence of non-genetic factors, and inheriting additional genetic elements that either intensify or mitigate the impact of the primary Mendelian genetic variant on the phenotype.

1.2.2 Common variants

Complex inheritance is based on the simultaneous inheritance of multiple predisposing variants. Genome wide association studies (GWAS) have been traditionally used as a means to understand the associations of single nucleotide polymorphisms (SNPs) with disease or phenotypes. These SNPs are commonly found in the genome, they usually have small effect sizes and their aggregate effect can be calculated by the use of a Polygenic Risk Score (PRS).

Polygenic Risk Score (PRS)

The traditional PRS is derived using the total of independent risk variants associated to a condition, as they are discovered and defined by large-scale GWAS for the given condition. For every individual, the count of risk alleles at each variant (0, 1, or 2) is summed, adjusted by its impact (e.g., log Odds-Ratio for binary characteristics or the beta coefficient for continuous ones) (Choi et al., 2020). This results in a singular score indicating each individual's genetic predisposition for a disease or continuous trait (Lewis and Vassos, 2020). Calculating PRS relies on two datasets: base data, derived from GWAS, which provide summary statistics like genotype-phenotype correlations at SNPs across the genome. These are often publicly available or cataloged in resources with large-cohort GWAS efforts. The target data also consists of genotypes and often phenotypes. These are independent samples from the base data, used solely to calculate the effects of the previously GWAS-known SNPs of interest and generate the PRS. The main analysis takes place in this dataset of unseen samples and are stored as `plink` binary files (Chang et al., 2015).

A number of different methods are available for calculating risk scores. One set of method involves 'clumping/pruning and thresholding', in which a filtered group of genetic variants is selected by pruning based on linkage disequilibrium and considering their association with the studied trait (clumping). PRSs are then derived by summing over all SNPs that meet a specific *p*-value limit or multiple thresholds, as seen in tools like PRSice (Choi and O'Reilly, 2019, Euesden et al., 2015). Alternatively, some techniques focus on achieving the best prediction across the genome by modeling the correlation between variants without trying to pinpoint a minimal SNP set for forecasting; the Bayesian LDpred approach (Vilhjálmsson et al., 2015) is the most common choice in this category. Another Bayesian approach includes SBayesR

(Lloyd-Jones et al., 2019).

HCM and DCM polygenic evidence

Harper et al. (2021) reported results from a GWAS involving 2,780 HCM cases and 47,486 controls, where 12 significant genetic risk factors for HCM were identified. Data on SNPs revealed a notable polygenic effect, particularly in HCM cases without sarcomere mutations (64% of cases; $h^2g = 0.34 \pm 0.02$). A genetic risk evaluation displayed a pronounced impact on HCM likelihood in a validation study, cutting the risk by half in the lowest quantile and increasing it twofold in the highest quantile. This evaluation also affected the severity of symptoms in those with sarcomere variants.

Following up on this, Zheng et al. (2023b) showed the clinical utility of PRS in HCM by incorporating SNPs from the largest HCM GWAS, comprising of 5900 cases and 68359 controls. They demonstrated that PRS effectively differentiates penetrance among carriers of rare pathogenic variants in sarcomeric genes, while also stratifying the risk of developing HCM in the general population. PRS also correlates with HCM risk in family members of probands and serves as a new risk indicator for survival and adverse cardiovascular events in cases (Tadros et al., 2023, Zheng et al., 2023b).

Another GWAS and multi-trait analysis was done in HCM (n=1733), DCM (n=5521), and 9 LV traits, where 16 loci were found to be associated with HCM, 13 with DCM and 23 with the imaging traits. The study presented pronounced genetic associations of the cardiomyopathies with the LV traits, that have opposing outcomes between the two types of cardiomyopathy. The authors explored whether common variants and polygenic effects could explain the variability in disease severity of HCM cases attributed to rare pathogenic

sarcomeric variants. It was shown that PRS is indeed associated with maximum LV wall thickness, indexed to body surface area and adverse clinical events; indicating that an increase in PRS increased the risk of these (each SD increase results to 28% higher risk of clinical event) (Tadros et al., 2021).

Another two new SNPs were identified with GWAS in DCM, in a 2021 study of 2,719 cases and 4,440 controls. The same study confirmed another two loci from previous work and generated a genetic risk score based on the count of risk alleles from all four DCM loci. This showed that individuals with 8 risk alleles had a risk of DCM that was three times higher than those with five risk alleles (the median in the referral group). They also did *in silico* functional analyses of these loci and found connections with taurine transporters, which are shown to have associations with myocardial dysfunction and DCM (McGurk et al., 2022). Interestingly, certain SNPs from the two newly identified loci linked to DCM were previously associated with cardiac structure and function in the general population (which had a normal average ejection fraction) in the UK Biobank (Pirruccello et al., 2020).

Finally, DCM's polygenic architecture was shown in the largest DCM GWAS meta analysis of 7 cohorts by the HERMES consortium (Lumbers et al., 2021), consisting of 14,255 DCM cases and 1,199,156 controls. Eighty genetic susceptibility loci were identified and 61 potential effector genes. Single-nuclei transcriptomics data from a small subset of DCM cases and controls uncovered cellular conditions, biological processes, and intercellular interactions central to DCM pathogenesis. PRS was shown to predict DCM in the broader population and influence the expression of rare pathogenic and likely pathogenic variants in genes causing DCM. Specifically, there is a 4-fold increased risk for DCM in the top PRS centile compared to the median, and 7-fold compared to the bottom centile. In over 25,000 individuals with specific

cardiac traits measured, the PRS was associated with reduced contractility and larger volumes, typical of DCM. In carriers of rare DCM-causing variants, the PRS effectively differentiated DCM prevalence, especially among those carrying rare protein-truncating variants in the *TTN* gene (Zheng et al., 2023a).

1.3 Cardiomyopathies and Machine Learning

The majority of research combining machine learning models and some prediction, stratification or risk assessment of cardiomyopathies involves basic genomic data information and outputs of imaging diagnostic tests such as cardiac magnetic resonance imaging (CMR) or electrocardiogram (ECG) data. In Chapter 2 there is a detailed summary of main machine learning algorithms used extensively in biomedical settings and the most common neural network models, also used in this thesis. In this section fundamental studies in imaging and genomic data in relation to cardiomyopathies will be introduced.

1.3.1 Cardiomyopathies and imaging with ML modeling

Left ventricular (LV) hypertrophy of the heart muscle is a strong predictor of cardiovascular outcomes; recent work (Biffi et al., 2018) has shed light on its genetic regulation through the use of a framework for statistical parametric modeling of 3D cardiac structure. This robust framework encodes multiple phenotypic traits and when validated on both synthetic and clinical datasets it shows suitability for detecting genotype-phenotype interactions on LV shape and geometry.

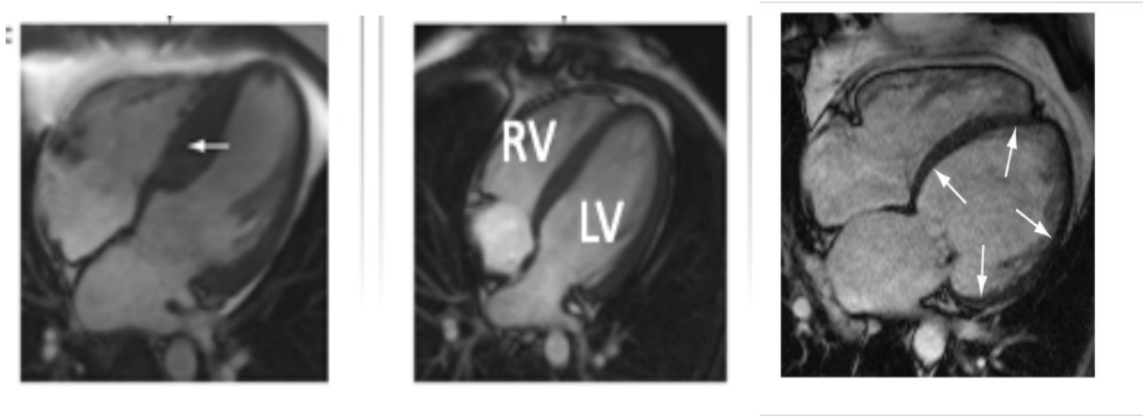


Figure 1.2: From left to right: HCM, normal, and DCM heart muscle cardiac magnetic resonance images. The arrow in the HCM image shows the enlargement of the wall. Arrows in the DCM image show the dilation of the ventricle and subsequent thinking of the wall. RV: right ventricle, LV: left ventricle (Reproduced with permission from (Watkins et al., 2011), Copyright Massachusetts Medical Society.)

CMR-based studies

The Biffi framework utilizes CMR image inputs that have been segmented by a method developed in the O'Regan lab (Bai et al., 2018, Bello et al., 2019). Briefly, the left and right ventricles from the CMR images are segmented using previous knowledge of the cardiac anatomy from a set of manually annotated atlases. Each of these segments are then co-registered to ensure the anatomical consistency between subjects, creating a template space of the heart structure. This template space is referred to as the atlas. The atlas acts as a reference for the subsequent transformation of the anatomical and functional information of each subject image. For example, when an image goes through transformation to the template space, the algorithm ensures that the position and orientation of the points across all subjects are identical but it still keeps the differences in size and shape (Bai et al., 2015, Marvao et al., 2015). This results in a 3D shape cardiac model with 46,808 points/vertices that can be visualized in a standard coordinate space. Slight differences in the spatial coordinates across all subjects account for the size and shape variations.

These 3D structures of the left ventricle were further used by Curran et al. (2023) to build a tree-based classification of HCM patients, based on a machine learning dimensionality reduction method. Four distinguishable phenotypic patterns were identified: (a) hypertrophy not linked to sarcomeric variants but associated with hypertension; (b) widespread and lower asymmetrical hypertrophy linked to outflow tract blockage; (c) isolated basal hypertrophy; (d) milder hypertrophy without obstruction, primarily found in familial cases with sarcomeric HCM. Additionally, polygenic risk related to HCM showed varied disease expression patterns. The model's reliability was confirmed in a separate cohort.

Another study looked at more precise phenotype extraction from the 3D images that leads to better genetic discovery of related loci. Phenotyping is carried out using autoencoders and the latent variables capture the geometry of the biological structure in question. Training of this network happens in two phases: the initial phase is independent of genotype, while the second phase associates with genetic markers chosen through GWAS based on the intermediate representation. This genotype-based optimization helps refine the phenotypes from the autoencoder, enhancing our comprehension of the genetic markers' impacts. The method was evaluated using UKBB LV meshes, unveiling new genetic associations previously not found in cardiac phenotype literature (Bonazzola et al., 2021).

Given the strong link of DCM with heart failure, the majority of studies using ML and imaging data focus on heart failure prediction or trying to explain DCM's progression to heart failure. Thanaj et al. (2022) evaluated CMR data from UKBB participants, utilizing the same segmentation framework described above but also evaluating motion tracking of the cine images using non-rigid image registration between successive frames (Puyol-Anton et al., 2018, Rueckert et al., 1999). Three established diastolic function traits were measured:

radial and longitudinal peak early diastolic strain rate and the maximum body surface area-indexed left atrial volume. GWAS was performed on these diastolic function traits, so as identify actionable causative mechanisms. Diastolic function was found to be hereditary, with associated loci linked to myofilament mechanics, protein synthesis under mechanical stress, and the regulation of cardiac contractility. Mendelian randomization identified a causal link between genetically defined diastolic function and heart failure outcomes. Examination of CMR images has also shown a greater trabecular fractal dimension in DCM patients compared to controls, mostly at the base and apex of the left ventricle (Meyer et al., 2020). Cardiac trabeculation is the network of muscle fibers or ridges (called trabeculae) that form on the inner surfaces of the ventricles. Excessive trabeculation has been shown to be associated with adverse cardiac outcomes (Petersen et al., 2023) while an inverse relationship can be observed with DCM and heart failure where loci linked to reduced trabeculation are correlated with higher risk of disease (Meyer et al., 2020).

Two-dimensional cardiac traits can be derived from the short axis cine image stacks of the CMR, that relate to either hypertrophic or dilated cardiomyopathy. Standard measurements of the right and left ventricles (RV and LV) like chamber volumes, ejection fraction, and LV mass can be obtained from the short-axis cine images. Tagging sequences enable the assessment of strain, providing a more detailed view of myocardial contraction than traditional measures like ejection fraction. Strain measurements can also be achieved using feature tracking methods on both long- and short-axis cine images (Raisi-Estabragh et al., 2021). A full list of the 80 2D cine image derived traits can be found in Chapter 3.1. These derived traits are not as high-dimensional as the meshes and have been used more extensively in studies.

Li et al. (2023) investigated the predictive value of different segments of wall thickness

for mortality in DCM. The left ventricular posterior wall thickness (PWT) independently predicted overall mortality. Relative interventricular septum wall thickness (RWT) showcased a higher predictive capacity compared to PWT. Patients with a PWT of less than 5.7 mm or an RWT of less than 0.19 faced a notably higher mortality rate. However, there was no evident link between the thickness of the left ventricular septum wall and DCM patient outcomes. These results indicated that both PWT and RWTPW are valuable markers for forecasting DCM outcomes and different parts of the left ventricular wall might be more responsible for the phenotype.

In addition to discovery of new factors or traits explaining cardiomyopathies, a few studies also use imaging for diagnosis between the different types of cardiomyopathies and cardiovascular diseases. A number of studies have developed neural network algorithms for diagnosis prediction between HCM, DCM, controls and another two types of cardiomyopathy from the Automated Cardiac Diagnosis Challenge (ACDC) dataset. Findings demonstrate that using multi-task training leads to a convergence rate that is 2.5 times quicker and cuts the diagnostic error from 32% down to 22% when compared to a foundational method trained without segmentation (Snaauw et al., 2019). Diagnostic models where segmentation is separate include pipelines where the segmentation is done by a neural network architecture, and then a number of features are selected as input to an additional ML model where the disease classification is performed. Random Forests have been used as models for multi-class classification using these derived CMR features, with accuracies at 90%-91% (Khened et al., 2018, Wolterink et al., 2018).

Finally, de Marvao et al. (2021) demonstrated the relationships between LV traits and HCM sarcomeric P/LP variants. Notably, these variants were linked to a higher lifetime risk of

adverse cardiovascular outcomes, mainly due to heart failure and arrhythmia. This risk was somewhat independent of LV wall thickness and was not limited to individuals with HCM or those meeting cardiomyopathy criteria. Therefore, it is evident that the risk of heart failure is not fully captured by visible phenotypic expression in the LV wall; which, aligns with the profibrotic state that arises before LVH development in cases with HCM variants (Ho et al., 2010). This emphasizes the potential of genomic testing for sarcomeric variants to adjust cardiovascular risk evaluations, even without CMR-based phenotyping.

CMR/ECG-based studies

An ECG is a diagnostic tool that measures the electrical activity of the heart. The test provides insights into the heart's rhythm and electrical conduction. Specifically, the conventional 12-lead ECG captures voltage variations between specific points on the body's surface throughout the cardiac rhythm cycle. These changes mirror the shifts in transmembrane voltages in myocardial cells during each cycle's depolarization and repolarization phases. Disruptions or abnormalities in this activity can indicate various cardiac conditions or disorders. The ECG can reveal patterns corresponding to arrhythmias, myocardial infarction, and other cardiovascular disease, including cardiomyopathies (Kligfield et al., 2007).

A study looking at penetrance of HCM and associating ECG and CMR phenotypes showed that an abnormal ECG is associated with a greater risk of developing HCM; while around 50% of carriers of sarcomeric P/LP variants manifested the disease even after initial negative screening. The abnormal ECG association with higher risk was consistent irrespective of gender or genotype. Presenting a normal ECG does not necessarily rule out HCM, as over a quarter of patients lacked significant ECG irregularities at the time of their diagnosis (Lorenzini

et al., 2020). For DCM, using the 12-lead ECG and CMR screening in UKBB for individuals with suspected pathogenic variants, 16.2% were found to have early, subclinical signs of DCM, including cardiac arrhythmias, isolated ventricular enlargement, and non-dilated hypokinetic DCM (Shah et al., 2022). Identifying cardiomyopathy cases with the use of imaging and specifically ECGs, given the evident source of information hidden in them, can be effectively empowered by the use of ML.

A convolutional neural network (CNN) was designed to detect HCM from 12-lead ECGs. The model's performance reached an AUC of 0.96, with low rates of false positives and a high negative predictive value at various HCM probability thresholds. The model implemented on the younger age group of the study (less than 40 yo) achieved the highest metrics values; however, sample sizes in young adults were low and pediatric patients were not considered. Moreover, for participants meeting the ECG criteria for LVH, the model's performance is comparable to the general population with an AUC of 0.95 (95% CI: 0.94 to 0.97). This indicates the model's efficiency in differentiating HCM from other LVH causes (Ko et al., 2020). HCM patients where clustered in phenogroups based on high-resolution ECG signals and analyzing the QRS patterns and T wave indicators. These phenogroups exhibit variations in both HCM Risk-SCD scores and the localization of LV hypertrophy, as visualized through CMR (Lyon et al., 2018). The use of AI-ECG in HCM also showed promise in observing disease progression, cardiac hemodynamics, and treatment effects in a universally available and non-intrusive way, as shown in a study monitoring the response to mavacamten (a cardiac myosin inhibitor) via longitudinal ECG (Siontis et al., 2023). Studies assessing cardiac contractile activity or hypertrophy as phenotypes, and not classifying and differentiating HCM from non-HCM controls, also yield additional information on the structural intricacies of the

heart (Attia et al., 2019). Unlike existing methods that simply detect cardiac structural issues through ECG, the ecgAI method measures the severity of these abnormalities on a scale. Using echocardiography as a benchmark, the accuracy of ecgAI varied, with the best predictions for left ventricular mass and the least accuracy for left atrial volume. When measurements were categorized, the model effectively distinguished between LVH and diastolic dysfunction, but struggled with estimating larger atrial volumes. The ecgAI method can also highlight specific ECG components that most influence the classification of each cardiac anomaly (Tison et al., 2019).

Finally, the ability to predict all-cause mortality and future cardiovascular events via features from Neural Network (NN) processed ECGs can create significant prognostic value. A model developed by Sau et al. (2023) defined subgroups that leverage existing ECG markers; their results indicate that these groups offer added prognostic insights beyond traditional markers. When tested on various diverse cohorts, including the UKBB and SaMi-Trop, the identified phenogroups consistently showed their prognostic relevance, even in different demographic contexts and specific diseases like Chagas Cardiomyopathy. This suggests that these NN-based ECG features can be beneficial for a variety of cardiac conditions and clinical scenarios.

1.4 Genetics of cardiomyopathies in clinic

The development of next generation sequencing has allowed for clinical genetic testing panels with genes related to a specific phenotype (Hershberger et al., 2018). Double-stranded DNA probes are used as a mutation capture technology in the clinical setting. This technology shows high efficacy in detecting mutations in cardiac genes and therefore is commonly used for clinical testing (Wilson et al., 2015). Approaches like the ACMG (American College of Medical Genetics and Genomics) framework (Richards et al., 2015) and tools like CardioClassifier (Whiffin et al., 2018), an automated and interactive web tool for clinical variant interpretation, provide a way to discriminate benign and pathogenic variants for clinical application across a wide range of ICCs.

In research settings, whole exome sequencing (WES) is more commonly used as it provides more information than targeted panels. Specifically, WES is the process of identifying the sequences of all protein-coding regions of the genome, the exome (Ng et al., 2009). This technique can capture variants that alter protein structure and lead to disease from a more inclusive set of regions compared to targeted panels. As WES is being introduced in clinical settings, more and more labs are now conducting WES pipelines where the analysis includes a virtual panel of all exomes generated, but only genes of interest analyzed (Seaby et al., 2016). This WES clinical implementation makes the need to overcome issues in clinical variant annotation and other hindering utility factors more imperative.

WES proves effective in diagnosing patients with DCM. A recent study showed that WES picked up a pathogenic or likely pathogenic (P/LP) variant in 12% of the DCM patients in a typical clinical environment. An additional 10% had truncating *TTN* variants labeled as variants of unknown significance (VUS). No variants in the expanded gene set data were

assigned as pathogenic. Alarmingly, 5 participants (6%) had variants deemed as unknown based on strict American College of Medical Genetics (ACMG) guidelines, yet other clinical labs classified them as either P/LP. The study's findings suggest that when adhering closely to current guidelines for classification, the identification rate of pathogenic variants is less than what was earlier reported (Ramchand et al., 2020).

The HYPERGEN study verified the effectiveness of WES in detecting HCM. Beyond the variants identified by targeted sequencing, WES found additional variations, indicating a broader genetic heterogeneity and possibly oligogenism. While it reinforced the significant role of *MYBPC3* and *MYH7*, it also highlighted an unexpected number of mutations in genes like *TTN*, *FLNC*, *SCN5A*, and *RYR2*, suggesting a link between HCM and arrhythmia genes (Nguyen et al., 2021).

1.5 Hypothesis and Aims

My hypothesis for this project lies on the above evidence that both genomic and CMR data can be utilized to predict cardiomyopathy clinical outcomes and therefore enhance the effectiveness of diagnosis, prognosis, risk assessment, and stratification of patients.

The overall aim of this thesis is to leverage novel ML methods to either investigate important CMR imaging features and the polygenic nature of cardiomyopathies as well as to pre-screen and/or stratify patients.

My specific objectives are:

1) Develop new ML methodology to mitigate the issues of correlation in general '-omics' datasets, encompassing the genomics and imaging data utilized in this thesis.

2) Employ interpretable ML methods to train models for pre-screening and/or stratification of HCM and DCM patients, with the use of genomic and clinical information readily available in the health record.

3) Develop a novel deep learning method for calculating Polygenic Risk Score (PRS) and assess its utility on known cardiomyopathy outcomes and traits.

1.6 Scope of this thesis

This thesis will combine the efforts of developing new ML methodology to address some long-standing issues with -omics datasets (including genomics and imaging) while also leveraging already established ML techniques to assess their utility in stratifying, pre-screening or creating means (such as PRS) for evaluating cardiomyopathies.

This Chapter 1 provided an introduction on cardiomyopathies and Chapter 2 provides a

detailed methodological description of ML methods utilized throughout the thesis. Chapter 3 will introduce the datasets available for the analyses and preliminary unsupervised exploration of the data types, which includes the genomic and the highly-correlated CMR imaging data. Chapter 4 describes the novel method developed to address correlated datasets (LAVASET) which investigates more efficiently the important features found in the CMR imaging data. Chapter 5 introduces interpretable ML models for HCM and DCM patient stratification, with the use of clinically readily available information, and how these models' performance is enhanced when incorporating important CMR features produced by LAVASET in Chapter 4. Finally, Chapter 6 showcases how deep learning can be used to generate PRS and evaluate its utility in cardiomyopathies, given their complex genetic architecture. Chapter 7 rounds off the thesis with general conclusions and future perspectives.

Chapter 2

Introduction to Machine Learning

2.1 Introduction

This chapter will introduce and explain in detail a variety of classical machine learning methods, along with some neural network algorithms, that are used extensively in this thesis. Both supervised and unsupervised learning techniques have been employed for the analyses. Supervised learning is a type of ML where the algorithm learns from labeled data. In this approach, the dataset used for training contains both input features and their corresponding output labels or target values. The algorithm learns to map the input features to the correct output labels by minimizing the error between the predicted and actual labels. Supervised learning algorithms can be used for classification (predicting binary/categorical labels) and regression (predicting continuous values). Unsupervised learning deals with unlabeled data, where the algorithm learns patterns or structures in the data without any explicit target labels. The goal of unsupervised learning is to discover hidden patterns, relationships, or clusters in the data. The algorithm aims to represent the data in a structured way by trying to group similar data points together (Bishop, 2006).

2.2 Statistical Testing

Mann-Whitney U-test: The Mann-Whitney U-test is a non-parametric statistical test that is used to determine whether there are significant differences between two independent groups on a continuous or binary outcome. It does not assume a normal distribution of the outcome variable. The test works by ranking all the values from low to high across both groups and then assessing whether one group tends to have higher or lower ranks compared to the other group (Mann and Whitney, 1947).

Kruskal-Wallis H-test: The Kruskal-Wallis H-test is also a non-parametric test and an extension of the Mann-Whitney U-test for more than two groups. It is used to determine if there are significant differences between three or more independent groups on a continuous or ordinal outcome. Like the Mann-Whitney U test, it ranks all the data but across multiple groups (Kruskal and Wallis, 1952).

Independent T-test: The independent t-test, also known as the two-sample t-test, is used to determine whether there is a statistically significant difference between the means of two independent or unrelated groups. It is a parametric test that assumes the data are normally distributed within each group and that the variances between the two groups are equal (Pfanzagl and Sheynin, 1996).

Multiple testing correction: For multiple testing correction, the Hommel method was used as a method designed to control the family-wise error rate (FWER) when conducting multiple comparisons. The family-wise error rate is the probability of making one or more false discoveries, or Type I errors, when performing multiple hypotheses tests. Hommel's method is an improvement over the more conservative Bonferroni correction, which can be overly stringent and reduce statistical power. The Bonferroni correction simply divides the significance level (e.g., 0.05) by the number of tests to ensure that the overall probability of a Type I error across all tests does not exceed the desired level. Hommel's methodology is more complex and adjusts *p*-values while taking into account the interdependence of the tests. It is less stringent by applying a smaller adjustment to each individual test, thus increasing the chances of detecting true effects while still controlling the FWER (Hommel, 1988).

2.3 Unsupervised Learning

1. **Dimensionality Reduction** UMAP (Uniform Manifold Approximation and Projection) was used to reduce the high-dimensional number of features (common SNPs or rare variants). UMAP is a nonlinear method that preserves both local and global structure in the data. It creates a fuzzy topological representation of the high-dimensional data and then finds a low-dimensional embedding that best preserves this representation. UMAP achieves this by optimizing an objective function that balances the preservation of pairwise distances and local neighborhood structures in the data (Mcinnes et al., 2020).

2. **Clustering** K-means clustering is a partition algorithm which aims to separate n points in k clusters by assigning each point to the cluster with the nearest mean. This process is done a number of times so as to minimize the variance within each cluster by calculating the squared Euclidean distances of the points. To find the optimal number of clusters, the within-cluster-sum of squared errors (WSS) for 2 to 10 number of clusters was calculated. When plotting the WSS against the number of clusters we pick the point where WSS starts to flatten also visible as an elbow on the graph. This point is the value with the maximum second-order difference in WSS (Arthur and Vassilvitskii, 2006).

2.4 Supervised Learning

2.4.1 Random Forest

Random Forest (RF) is an ensemble learning method that was developed as an extension to the decision tree algorithm by Breiman (2001).

The RF algorithm starts by constructing multiple decision trees in parallel. Each tree is built independently and using a different subset of the data which are created by randomly sampling the dataset with replacement (bootstrapping). The user can define the proportion of samples to be used for each tree, and whether samples should be chosen with or without replacement. To increase robustness, for each decision tree, RF also considers a random subset of features for each node prior to determining the split, rather than considering all features. Here, the user can again define the proportion of features to be selected for each tree (Breiman, 1984, 2001).

In default settings, trees are built to their maximum depth and are not pruned. Unless specified otherwise by the maximum depth parameter, individual trees will grow until all samples are being separated in the purest leaf nodes. Splits are being decided based on different metrics: for classification, Gini impurity or entropy can be used and for regression, variance. Once all the trees in the ensemble have been built, the RF makes predictions by aggregating the prediction outputs of all the decision trees (bagging). For classification the mode of the classes (i.e., the class that gets the highest number of votes) predicted by each tree is taken as the final prediction. For regression it takes the average of the outputs by different trees. The final prediction aggregated from all the trees is returned as the output of the Random Forest (Breiman, 1984, 2001).

A useful feature of RF is its ability to perform cross-validation during training, due to the fact that each tree is constructed using a different bootstrap sample subset. The user can specify the proportion of samples left out of the bootstrap sample, and use these as the out-of-bag (OOB) sample. For each tree, the respective OOB sample can be used to validate the training performance on an unseen set.

Advantages of RFs include robustness due to the number of trees participating in the process. RFs can also handle missing values and large data sets with higher dimensionality. Its major limitation lies on the ability to disregard correlation between variables.

2.4.2 Gradient Boosting Trees

Gradient Boosting Trees (GB-Trees) is an ensemble learning method used for regression and classification problems. It is based on the idea of boosting, which involves combining several weak learners to create a stronger ensemble. In the case of Gradient Boosting Trees, similar to Random Forest, the weak learners are decision trees.

The algorithm starts by predicting a constant value for all samples in the dataset. This constant is usually chosen to be the average target value for regression, or the majority class for classification. After making predictions with the base learner (first tree trained), the algorithm computes the loss, defined by the difference between the predicted values and the real values. It then calculates the negative gradients of the loss with respect to the predictions. Construction of the next trees is guided by these gradients, also known as 'pseudo-residuals', which dictate how each tree should correct and minimize the errors of the previous one. After fitting a decision tree to these gradients, a weight is assigned to it based on the error modeling (for the specific tree) in relation to the ensemble. This weight is typically found by solving a one-dimensional optimization problem, where the goal is to find the weight that minimizes the loss when the tree is added to the current ensemble (Friedman, 2002, Ridgeway, 2020).

The predictions are updated by adding the weighted decision tree's output from the current iteration to the predictions from the previous iteration. This is done in a gradual manner, controlled by a learning rate parameter (usually between 0 and 1). The learning rate

essentially regularizes the contribution of each tree to the ensemble, preventing over-fitting. This process of fitting a new tree and calculating new weights is repeated a specified number of times or until the loss reaches an acceptable level or plateau on the validation set.

The final model is an ensemble of the weighted decision trees. To make predictions, the outputs of all the trees are summed up (with their respective weights). In addition to the learning rate, GB-Trees also have other regularization parameters like tree depth limitation, random subsampling of the training data and features to prevent overfitting.

GB-Trees are powerful and have been shown to be very effective on a wide range of machine learning problems. However, they can be sensitive to overfitting if the data is noisy, and they require careful tuning of the parameters. They are also computationally expensive, especially with large datasets, as the learning is sequential and cannot be parallelized as efficiently as Random Forests (Ridgeway, 2020).

2.4.3 Support Vector Machines

An SVM is a discriminative technique that maps samples in a high-dimensional feature space and finds the best separating hyperplane that divides the input in two classes (Gholami and Fakhari, 2017). Given a set of instance-label pairs $(x_i, y_i), i = 1, ..., l, x_i \in R^n, y_i \in \{-1, +1\}$, the SVM solves the following optimization primal problem:

$$\min \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}(\max(0, 1 - y_i\mathbf{w}^T + \mathbf{x}_i))^2$$

The goal of SVM during training is to find the optimal hyperplane that maximizes the margin between two classes. The margin is defined as the distance between the nearest data

points (of two classes) to the hyperplane. These nearest points are called support vectors. SVM aims to maximize this margin, so in the optimal hyperplane the distance from the margin to the support vectors on each side is maximized (Gholami and Fakhari, 2017). The soft margin allows some samples to be assigned on the wrong side of the hyperplane, but at a cost. The `C` parameter determines the trade-off between maximizing the margin and minimizing the classification error. A small value of `C` creates a wider margin but may tolerate more misclassifications, while a large value of `C` aims for a perfect classification, at the risk of over-fitting.

Kernel methods are a away to transform high-dimensional data into a feature space of pairwise similarity comparisons of the original data points. For non-linearly separable data, this transformation can be useful by making the data linearly separable for the hyperplace. The kernel function therefore reduces the dimensionality by calculating the inner products of the function output values of all data pairs. Common kernel functions include the polynomial kernel, radial basis function (RBF) and the sigmoid kernel (Hofmann et al., 2008).

To predict unseen data, the SVM projects the new data points on the high-dimensional space and calculates the signed distance from the point to the hyperplane. If the distance is positive, the data point belongs to one class, and if it is negative, it belongs to the other class.

Advantages of SVMs include the flexibility in choosing the kernel for various decision functions. This renders it well-suited for problems where samples are not linearly separable (often seen in biological contexts). SVMs are able to handle high-dimensional datasets, however this can sometimes come with a cost in memory or speed (Gholami and Fakhari, 2017).

2.4.4 Multi-Layer Feed-Forward Neural Networks

Multi-layer feed-forward neural networks (MLFNN) are a subset of artificial neural networks (ANNs) that follow the basic principles of ANNs: each neuron within the network is able to receive a signal, process it and send it as output to another neuron. Each neuron should be connected to at least one other neuron and their connection (edge) is defined by a real number. This value assigns an importance of the given edge in relation to the whole network (Svozil et al., 1997).

MLFNNs are designed in layers, where the first layer is the input layer, the in between layers are hidden layers and the last layer becomes the output layer. In a MLFNN, the signal moves in only one direction—forward—from the input nodes, through the hidden nodes, and to the output nodes, which is different from other types of ANNs where feedback loops or cycles are incorporated. Dense layers, also known as fully connected layers, are those in which each neuron receives input from all the neurons of the previous layer, fully connecting the layers (Glorot et al., 2011, Svozil et al., 1997).

The ReLU (Rectified Linear Unit) function is commonly used as an activation function in dense (hidden) layers of the MLFNN. It introduces non-linearity into the model, allowing the network to learn more complex patterns. The ReLU function is defined as $f(x) = \max(0, x)$. Therefore if the input is > 0, it will output the value directly, but if it is < 0, it will output 0. ReLU is favored because of its non-linear capabilities and simplicity which allows for computational efficiency in training (Glorot et al., 2011).

2.4.5 Metrics

A number of metrics were used to evaluate the different models across the chapters. The main metric calculated for evaluating each model's accuracy and effectiveness is the Area Under the Receiver Operating Characteristic (ROC) curve (AUROC) score. The ROC curve is a probability curve that shows the performance of a binary classifier across various thresholds. The curve is created by plotting the true positive rate (sensitivity or recall, described below) against the false positive rate (1 - specificity, described below) at various threshold settings. The AUROC value is the area under the ROC curve, ranging from 0 to 1. AUROC overall represents the degree of separability between the classes (Fawcett, 2006).

Precision is the proportion of true positive predictions among all positive predictions. It represents how many of the predicted positive instances are actually positive (Powers and Ailab, 2020).

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}}$$

Recall (also known as Sensitivity or True Positive Rate) is the proportion of true positive predictions among all the real positive instances. It represents how many of the true positive instances were correctly predicted by the model (Powers and Ailab, 2020).

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}}$$

The False Positive Rate (FPR) is the proportion of false positive predictions among all the known negative instances. It represents how many of the ground truth negatives were

incorrectly predicted as positive by the model (Powers and Ailab, 2020).

$$\text{FPR} = \frac{\text{False Positives (FP)}}{\text{False Positives (FP)} + \text{True Negatives (TN)}}$$

The F1 Score is the harmonic mean of precision and recall. It is often used when the classes are imbalanced as a more accurate evaluation metric for models. If either precision or recall is low, the F1 score will also be low, ensuring that a balance between the two metrics is achieved (Powers and Ailab, 2020).

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

When imbalance of classes poses an issue in evaluating the model another probabilistic curve can be used to assess performance. The Precision-Recall (PR) curve illustrates the trade-off between precision and recall across different threshold points. The PRAUC is the area under the PR curve (Davis and Goadrich, 2006).

The average precision (AP) in the context of the precision-recall curve is a summary measure that quantifies the relationship between precision and recall for a binary classification problem at different thresholds. It provides a single value that summarizes the precision-recall curve. The average precision is computed as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight (Powers and Ailab, 2020):

$$\text{AP} = \sum_n (R_n - R_{n-1}) P_n$$

Where:

- P_n is the precision at the n$^{\text{th}}$ threshold.
- R_n is the recall at the n$^{\text{th}}$ threshold.
- R_{n-1} is the recall at the (n-1)$^{\text{th}}$ threshold.

The AP calculates in a different way the area under the precision-recall curve, and it provides a single value that summarizes the shape of the curve. A perfect classifier will have an AP of 1, while a random classifier will have an AP value close to the proportion of positive samples.

Chapter 3

Introduction to the Dataset and Assessment of -Omics Sub-groups with Unsupervised Learning

3.1 Introduction

A variety of studies perform phenotypic clustering to look at subgroups of DCM or HCM. These use different omics inputs and focus mostly on diseased cohorts (Tayal). In addition, a lot of the methods employed to analyze these cohorts are linear methods (PCA, mixed linear regression), although recently more complex dimensionality reduction methods (DDRTREE) are seen being used (Curran et al., 2023).

Here we want to explore whether the raw genomic inputs generated in an additive model (0,1,2 encodings) can yield subclusters that are informative to explaining the phenotypic variability of HCM and DCM. This is approached by using information from mostly healthy volunteers of the UKBB, and a targeted disease HCM cohort. Both common SNPs and rare variants in genes associated with HCM and DCM are being analyzed, with an aim to evaluate whether there are hidden structures that can be driven by non-linear relationships between these variants. Additionally, being able to link genomics-derived groups with other types of -omics or clinical data available, such as the CMR imaging traits, could unlock informative links.

This chapter will introduce the two main datasets used in the subsequent analyses, as well as the data types utilized. It will also set the context for Chapter 4, and a novel ML method developed to mitigate correlation artifacts when using the Random Forest algorithm.

3.2 Datasets

For the entirety of analyses presented in this thesis two datasets were utilized. The UK Biobank (UKBB), a large voluntary-based cohort of participants, and the Royal Brompton

Hospital (RBH) Cardiovascular Biobank, a specific disease cohort, managed by collaborators.

3.2.1 UK Biobank (UKBB)

The UKBB initiative is a prospective study that gathered detailed genetic and phenotypic information from approximately 500,000 individuals in the UK, at an age between 40 and 69 years old when recruited. For each participant a wide variety of measurements and health-related data was collected, from lifestyle indicators to bodily fluids for biomarker exploration. For the purpose of this thesis, the two types of genomic data information was analyzed (genotyping and whole exome sequencing) as well as the cardiac magnetic resonance imaging, available in a subset of participants.

Genomic data

Two types of genomic data have been utilized in the analyses presented here. Whole-genome genotyping data for 488,377 participants, the majority of which was implemented with the UK Biobank Axiom Array, was released at the first stage of the initiative. Markers were chosen to capture genome-wide genetic variation with SNP and insertions and deletions; some markers were chosen for association with disease and to include coding variants across a range of minor allele frequencies (MAFs), and others to include rare variants. Detailed information on the protocol can be found at Bycroft et al. (2018).

The whole exome sequencing (WES) data was released in 3 stages, with the final release of the 454,787 participants occurring in the end of 2021. The data collected averages over 600 coding variations per gene, encompassing roughly 50 predicted loss-of-function (pLOFs) mutations per gene (Backman et al., 2021). Exomes were captured by a customized version

of the IDT xGen Exome Research Panel v.1.0. The basic design targets 39 megabases of the human genome, covering 19,396 genes across autosomes and sex chromosomes. To enhance the coverage in underperforming loci, extra probes were added. Further information on the Regeneron protocol can be found in Hout et al. (2019).

Imaging data

The cardiac magnetic resonance (CMR) protocol consists of a 20-min scan performed using a 1.5 T scanner (MAGNETOM Aera, Syngo Platform VD13A, Siemens Healthcare, Erlangen, Germany). The protocol includes bright blood anatomic assessment (sagittal, coronal, and axial), left and right ventricular cine images (long and short axes), myocardial tagging (three short-axis slices), native T1 mapping, aortic flow, and imaging of the thoracic aorta (Raisi-Estabragh et al., 2021). Further information on the exact imaging protocol details and ethical considerations are explained in Raisi-Estabragh et al. (2021).

39,559 out of the ∼50,000 Cardiac Magnetic Resonance (CMR) images have been segmented by a method developed by collaborators in the O'Regan lab (Bello et al., 2019). Briefly, the left and right ventricles from the CMR images are segmented using previous knowledge of the cardiac anatomy from a set of manually annotated atlases. Each of these segments are then co-registered to ensure the anatomical consistency between subjects, creating a template space of the heart structure. This template space is referred to as the atlas. The atlas acts as a reference for the subsequent transformation of the anatomical and functional information of each subject image. For example, when an image goes through transformation to the template space, the algorithm ensures that the position and orientation of the points across all subjects are identical but it still keeps the differences in size and shape

(Bai et al., 2015, Marvao et al., 2015). This results in a 3D shape cardiac model with 46,808 points/vertices that can be visualized in a standard coordinate space. Slight differences in the spatial coordinates across all subjects account for the size and shape variations. The QCed outputs include TXT files for each individual's left ventricular dimensions in systole and diastole. For each of the 46,808 vertices, the following measurements were taken:

- location in space: in x, y, z coordinates
- wall thickness in diastole; measured by calculating the distance of vertices between the endocardial and epicardial layers of the ventricle wall (it is only clinically relevant to study wall thickness in diastole since when the cardiac muscle contracts it is hard to distinguish wall thickness from the papillary muscles and trabeculations)
- signed distance measurements; a metric of how each point in the model compares to a mean shape (positive numbers correspond to a larger heart, etc.), this serves as the 3D estimation of the clinical diastolic and systolic volume metrics for each individual
- curvature; a metric for the regional curvedness of the left ventricle surface

In addition to these 3D shapes, phenotypes that relate to the wall thickness, chamber volumes, strains and trabecular traits have also been derived from the non-rigid image registration motion tracking analysis. The final traits are shown in Table 3.1, and more information on the methodology and guidelines used can be found in de Marvao et al. (2021).

3.2.2 Royal Brompton Hospital Cardiovascular Biobank

A total of 710 individuals diagnosed with HCM, who either visited the inherited cardiomyopathy service or were referred for CMR imaging, were consecutively registered in a

Category	Traits
Ascending Aortic (AA)	AAo max area, AAo min area, AAo distensibility
Descending Aortic (DA)	DAo max area, DAo min area, DAo distensibility
Left Atrium (LA)	LAV (Volume) max, LAV min, LASV (Systolic Volume), LAEF (Ejection Fraction)
Right Atrium (RA)	RAV max, RAV min, RASV, RAEF
Left Ventricle (LV)	LVEDV (End Diastolic Volume), LVESV (End Systolic Volume), LVSV, LVEF, LVCO (Cavity Obliteration), LVM (Mass), WT Global, WT Max Global, Err Global, Ell Global, Ecc Global
Right Ventricle (RV)	RVEDV, RVESV, RVSV, RVEF
Mean Wall Thickness (WT)	WT AHA 1, WT AHA 2, ..., WT AHA 16
Max WT	WT Max AHA 1, WT Max AHA 2, ..., WT Max AHA 16
Radial Strain Rates (Err)	Err AHA 1, Err AHA 2, ..., Err AHA 16
Circumferential Strain Rates (Ecc)	Ecc AHA 1, Ecc AHA 2, ..., Ecc AHA 16
Longitudinal Strain Rates (Ell)	Ell 1, Ell 2, ..., Ell 6

Table 3.1: Summary of CMR derived traits from motion tracking analysis. Abbreviations are explained in first occurrence on the table.

prospective registry as part of the NIHR Royal Brompton Hospital Cardiovascular Biobank project, from 2009 to 2015. A specialist in cardiomyopathy independently verified the HCM diagnosis based on recognized clinical and CMR guidelines. All the patients satisfied the diagnostic criteria set by the American Heart Association (see Intro). All the patients were subjected to CMR to evaluate the volumes and functions of their cardiac chambers using a 1.5T Siemens Sonata or Avanto system (Siemens Medical Systems, Erlangen, Germany). Segmentation of the images was performed as described above. For generating genotyping data, two ICC -specific panels were used: the Illumina TruSight Cardio Sequencing Kit1, which covers 174 genes linked to ICCs. This sequencing was performed on the Illumina MiSeq and NextSeq systems. The second panel was a custom-made Agilent SureSelect gene panel related to ICCs, using the Life Technologies SOLiD 5500xl platform. More information on the bioinformatic pipelines and variant calling procedures can be found on Walsh et al. (2017).

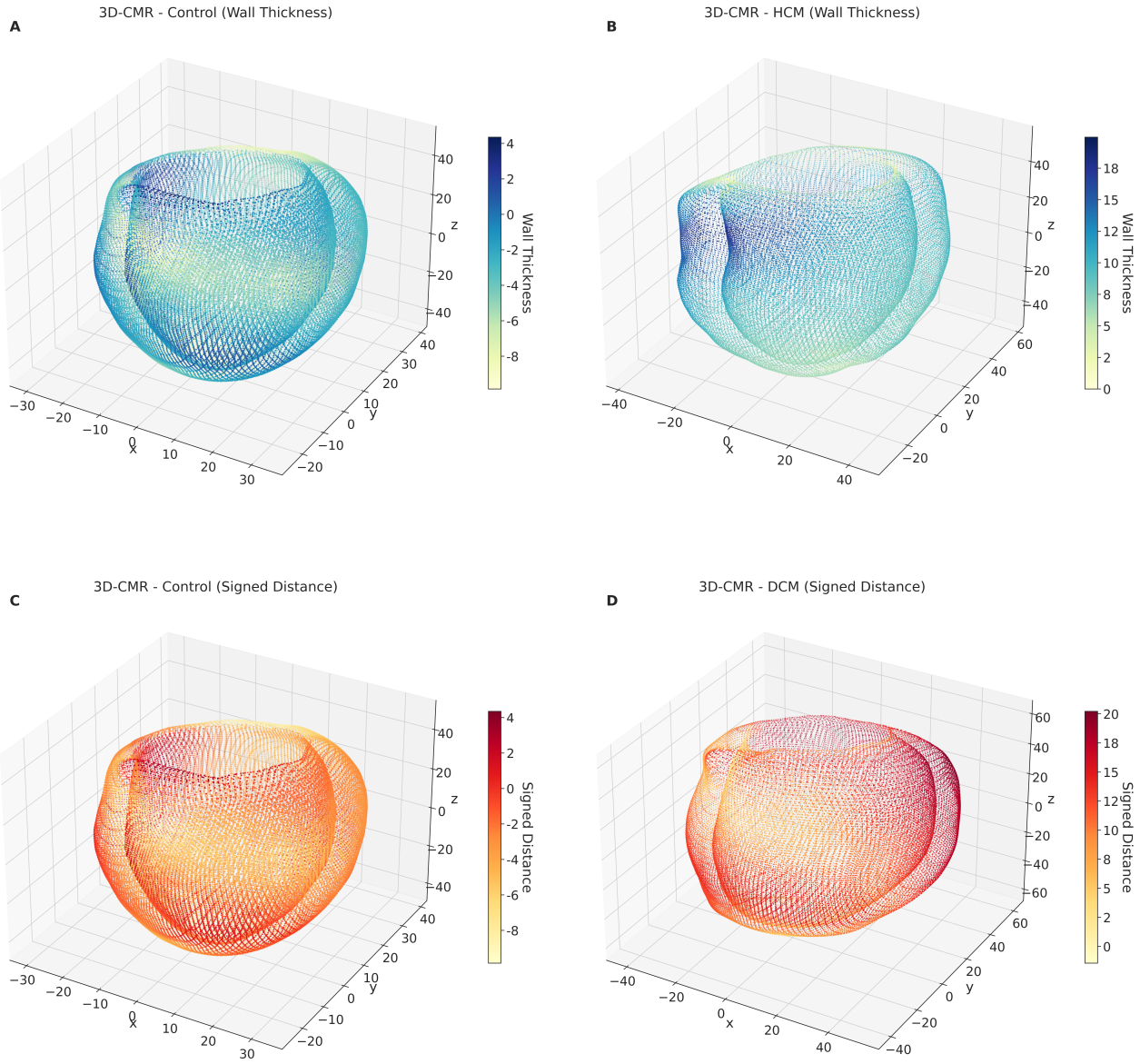


Figure 3.1: 3D plots of the CMR imaging meshes of a normal heart (Control, A, C), HCM (B), and DCM (D). A and B panels are color coded by wall thickness, while C and D panels are coded by signed distance (a more appropriate measurement for dilation in the left ventricle).

3.3 Methods

3.3.1 Common Variants

The first set of inputs generated for the unsupervised pipeline comprised of common SNPs (MAF $>$ 0.001). These were selected by curating GWAS outputs performed on the HCMR cohort (Tadros et al., 2023). SNPs that showed *p*-value significance higher than 0.001 and 0.0001 were considered, so as to filter out unrelated SNPs and reduce the excessive number of commons SNPs generated by the UKBB genotyping array. The final inputs included either 7,491 SNPs or a subset of the more significant 2,170 SNPs. To address the correlation relationships among the common SNPs we used the clumping method to keep SNPs with weaker linkage disequilibrium. We performed clumping on the larger set of SNPs, and after removal the total number of variants was 1,045.

A separate HCM cohort, in addition to the UKBB, was also analyzed to evaluate any significance in HCM stratification driven by common variation. The genotyping array performed on the RBH HCM cohort was different to the UKBB, resulting to a total of 11,316 SNPs or the subset of the more significant (p-value $\leq$ 0.0001) 4,378 SNPs.

For DCM, summary statistics were used from a large meta-analysis effort that combined 16 studies participating in the Heart Failure Molecular Epidemiology for Therapeutic Targets (HERMES) Consortium (Lumbers et al., 2021). *p*-value significance thresholds were kept the same as in HCM with the total number of SNPs at 0.001 adding to 16,175 and at 0.0001 to 3,850 SNPs. After accounting for LD the number of SNPs reduced to 2,876. More details on the summary statistic GWAS studies are given in Chapter 6, section 6.2.1.

3.3.2 Rare Variants

The second set of inputs included rare loci (MAF < 0.001) on specific genes associated with HCM or DCM, as they have been reported by our group or other studies (see Introduction 1.2.1). The encodings for these, similarly to the common variant inputs, followed an additive model with 0,1,2 for no variant, heterozygous and homozygous, respectively.

3.3.3 CMR imaging

For HCM, the wall thickness vector of the 3D meshes was used as an input to the UMAP dimensionality reduction method, so as to look at the separations within the UKBB as well as with the RBH cohort. The number of UKBB HCM cases is 34, for RBH 540, and for the Controls 699. The number of Controls was decided by bootstrap subsampling from the larger set of Controls with available CMR data. This was done for computational efficiency purposes for the novel algorithm developed and presented in Chapter 4 and so as to be able to compare the clustering structures before (as shown in this Chapter here) and after applying the new dimensionality reduction (Chapter 4). UMAP was also applied on the entirety of the available CMR data (n=39,559) without showing any novel phenotypic differences while also identifying segmentation errors that were also manually found and removed after QC (results not shown here).

3.3.4 Unsupervised Pipeline

We performed a 3-step analysis pipeline on the genomic data, that has previously been used and developed by myself in Mujagic and Kasapi *et al.* to derive -omics driven subgroups

within an Irritable Bowel Syndrome (IBS) cohort. The pipeline consists of a dimensionality reduction step, a clustering step and ultimately the statistical testing of the clusters. We improved the first step of the pipeline by incorporating a more stable dimensionality reduction method – UMAP –, in comparison to the kernel-tSNE that was used in the IBS study. The rest of the pipeline was followed as published.

1. **Dimensionality Reduction**

   UMAP (Uniform Manifold Approximation and Projection) was used to reduce the high-dimensional number of features (common SNPs or rare variants). UMAP was chosen over t-SNE for its ability to preserve global structure better (this way we avoided running the algorithm a number of times and measuring the stability of cluster membership across runs) and for its superior run time performance in this high-dimensional dataset. In all input cases UMAP generated coordinates for two components, so as to enable visualization in addition to analysis. Parameter tuning was applied on the minimum distance and number of neighbors to consider. For minimum distance, values were tested in steps of 0.1 (0.1, 0.2, . . . , 0.9) and the number of neighbors were tested in steps of 10 (10, 20, . . . , 100). Number of components was set to 2 for visualization purposes and the distance metric was set to manhattan (given the 0,1,2 encodings). Principal Component Analysis (PCA) was applied as a baseline dimensionality reduction method after mean-centering the inputs.

2. **Clustering**

   The 2-dimensional output coordinates from UMAP were used as input in the clustering step. K-means clustering was used as the method of choice. To find the optimal number

of clusters, the within-cluster-sum of squared errors (WSS) for 2 to 10 number of clusters was calculated. When plotting the WSS against the number of clusters we pick the point where WSS starts to flatten also visible as an elbow on the graph. This point is the value with the maximum second-order difference in WSS.

3. **Statistical Testing**

   At the third step, a number of statistical tests were employed to assess the statistical significance of the resulting groups. The Kruskal-Wallis test by ranks was used to determine whether there are any statistically significant differences among the clusters (when $k>2$) (McIntosh and Lawrie, 2010). The Mann-Whitney U-test then determined the statistical significance between all sets of pairs of clusters (McIntosh and Lawrie, 2010). Both tests are non-parametric rank-sum tests that can be used in groups of the same or different sample size. The tests were applied on a number of demographic and clinical outcomes (race, sex, age, systolic/diastolic blood pressure, etc), phenotypic (HCM, DCM, heart failure) and CMR imaging derived phenotype measurements (see Table 3.1) to determine statistical significance between the clusters. The Hommel correction (Chen et al., 2017, Hommel, 1988) – a measure of family-wise error rate (FWER) multiple testing correction based on Simes' test (Simes, 1986) – was used to correct for multiple testing.

3.4 Results

3.4.1 HCM UKBB

The most evident structures are shown with the common SNP inputs of p-value ≤ 0.0001. This amounts to 2,170 SNPs that are able to separate the samples in three most distinct groups, as shown in Figure 3.2. The UMAP optimal parameters generating these results are 0.2 for minimum distance and 30 for number of neighbors. Part A of Figure 3.2 indicates the optimal number of clusters generated by the within-cluster-sum of squared errors scores. The highest second order difference when looking at all values from 1 to 10 is found in k=3, where the graph shows a sharp elbow-like structure. In Figure 3.2B we find that the blue cluster (referred to as Group 1) is significantly different from the orange cluster (Group 2) when looking at maximum and particular spacial measures of wall thickness, as derived from 2D phenotyping of CMR imaging. Group 2 is also significantly different from the green cluster (Group 3) in these measures.

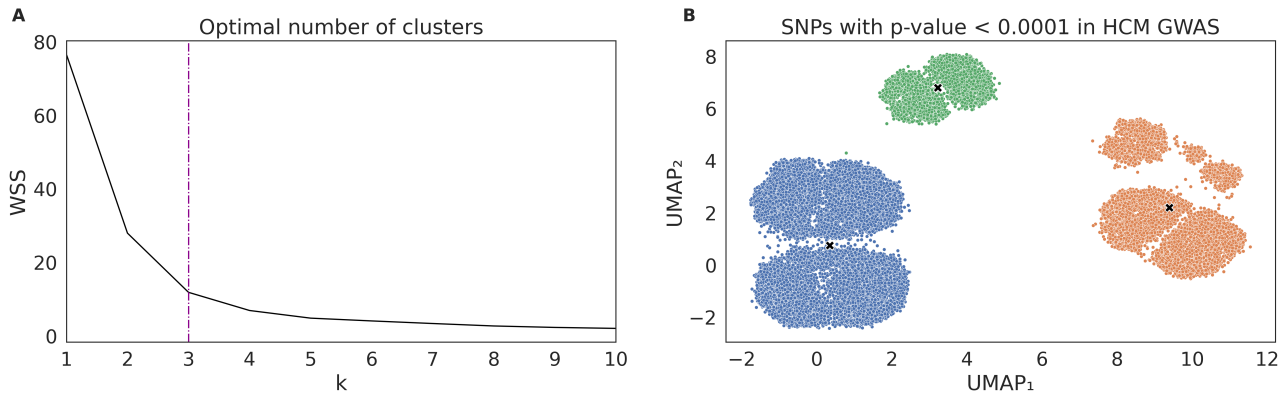


Figure 3.2: UKBB cohort: Unsupervised pipeline results for HCM-associated common variants. Left graph shows the WSS values in the y-axis across the tested number of clusters (k=1-10) in the x-axis. The optimal number of clusters corresponds to k=3, as shown by the purple dashed line. Part B shows the k-means clusters on the UMAP embeddings; x markers show the cluster centroid.

3.4.2 HCM RBH Cohort

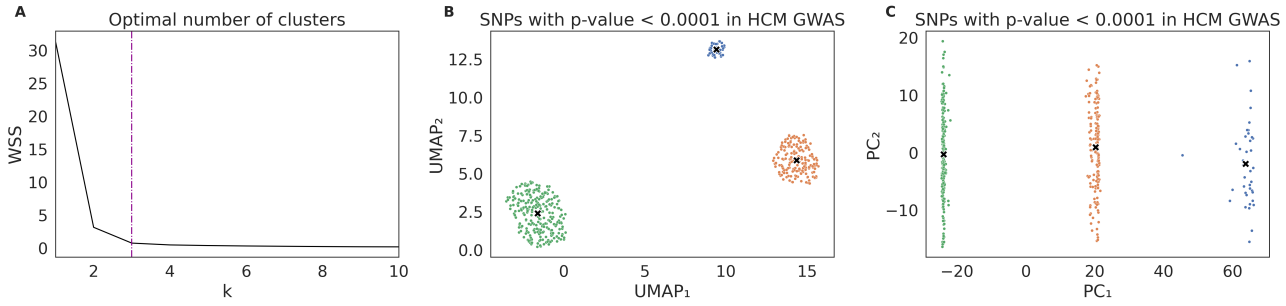


Figure 3.3: RBH HCM cohort: Unsupervised pipeline results for HCM-associated common variants. Part A shows the WSS values in the y-axis across the tested number of clusters (k=1-10) in the x-axis. The optimal number of clusters corresponds to k=3, as shown by the purple dashed line. Part B shows results of k-means clustering on the UMAP embeddings, indicating the 3 separate clusters; x markers present the cluster centroids. Part C shows identical clustering patterns for the 3 groups, when using PCA.

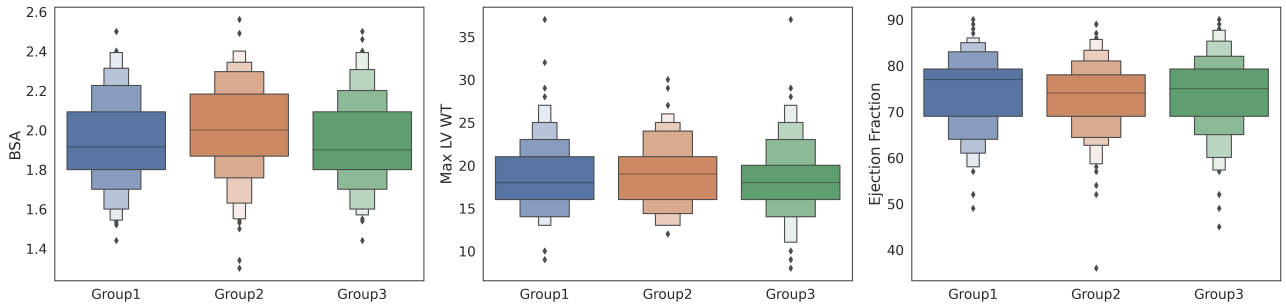


Figure 3.4: RBH HCM cohort cluster differences on CMR derived phenotypes. First graph on the left presents differences in body surface area (BSA). Middle graph presents differences in the left ventricle's maximum wall thickness. Right graph shows the ejection fraction.

3.4.3 DCM UKBB

The most evident structures are shown with the common SNP inputs of p-value ≤ 0.0001. This amounts to 3,850 SNPs that are able to separate the samples in two most distinct groups, as shown in Figure 3.5. The UMAP optimal parameters generating these results are 0.3 for minimum distance and 10 for number of neighbors. Part A of Figure 3.5 indicates the optimal number of clusters generated by the within-cluster-sum of squared errors scores. The highest second order difference when looking at all values from 1 to 10 is found in k=2, where the

graph shows a sharp elbow-like structure. In Figure 3.5B the two clusters are shown, contrary to the HCM samples above, DCM did not show any statistically significant differences in these structures.

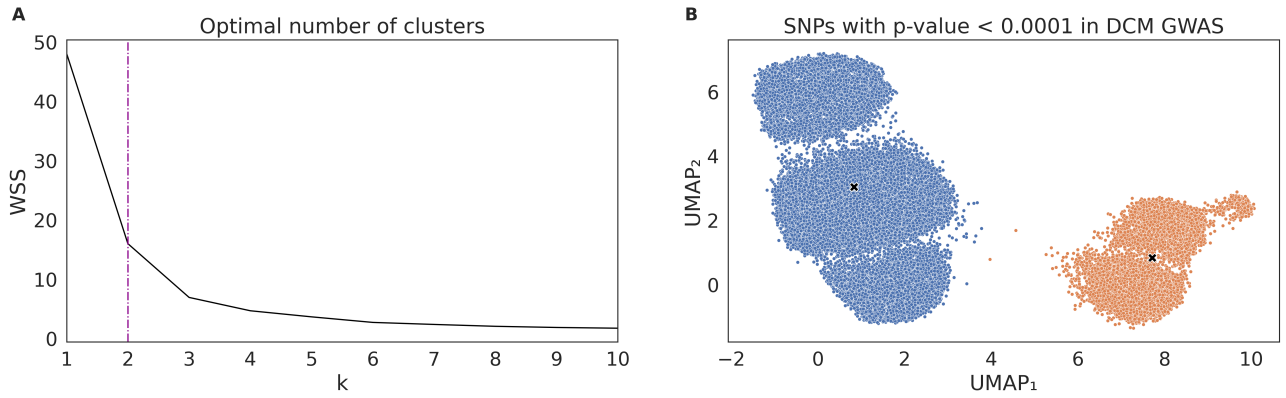


Figure 3.5: UKBB cohort: Unsupervised pipeline results for DCM-associated common variants. Left graph shows the WSS values in the y-axis across the tested number of clusters (k=1-10) in the x-axis. The optimal number of clusters corresponds to k=2, as shown by the purple dashed line. Part B shows results of k-means clustering on the UMAP embeddings, indicating the 2 separate clusters; x markers present the cluster centroids.

3.4.4 Clusters in common SNPs are attributed to correlation

Figures 3.6 and 3.7 show results for the UKBB HCM and DCM common variants as well as the RBH HCM variants after clumping has been applied. These structures show no evident separation between the points, with cloud-like structures dominating the clustering patterns. When applying the k-means algorithm on the RBH data, the method is trying to find the optimal separation, which in this case just separates the cloud of points in a lower, upper left and upper right group. However, these do not seem to have significant differences when testing a variety of clinical or imaging features.



Figure 3.6: UKBB cohort: Unsupervised pipeline results for HCM- and DCM-associated common variants after clumping. Part A shows results from the HCM-associated variants and Part B shows results from DCM-associated variants.

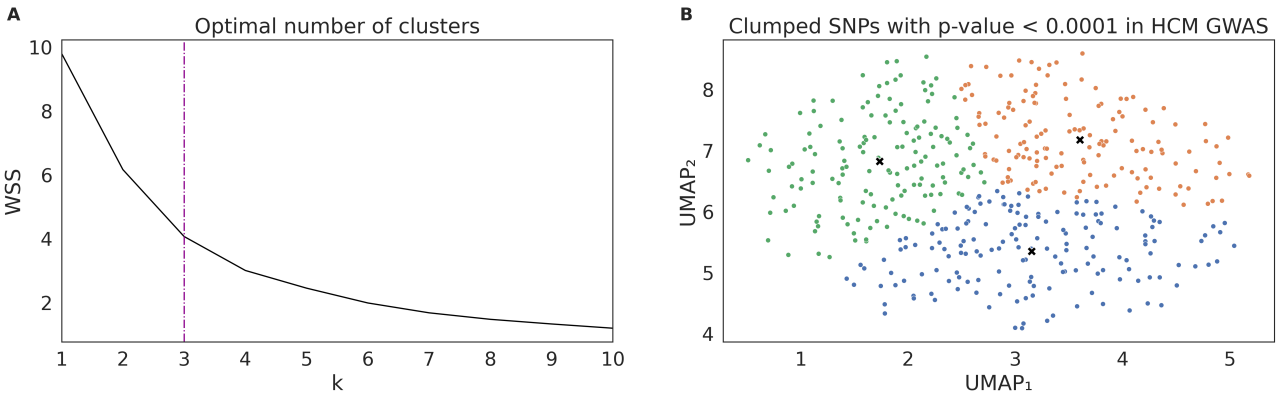


Figure 3.7: RBH HCM cohort: Unsupervised pipeline results for HCM-associated common variants after clumping. Part A shows the WSS values in the y-axis across the tested number of clusters (k=1-10) in the x-axis. The optimal number of clusters corresponds to k=3, as shown by the purple dashed line. Part B shows the clustering results on the UMAP embeddings; x markers indicatethe cluster centroids.

3.4.5 3D-CMR Imaging

Figure 3.8 shows results of the UMAP dimensionality reduction applied on the CMR imaging wall thickness vector. There is an obvious separation between the UKBB Controls (green squares) and the RBH HCMs (blue Xs), while the UKBB HCMs (red dots) are found more in the middle of the cloud of points. The clustering performance across 20 iterations of UMAP and k-means when setting the number of clusters to 2 (HCM and Controls) reaches an average accuracy performance of 0.79.

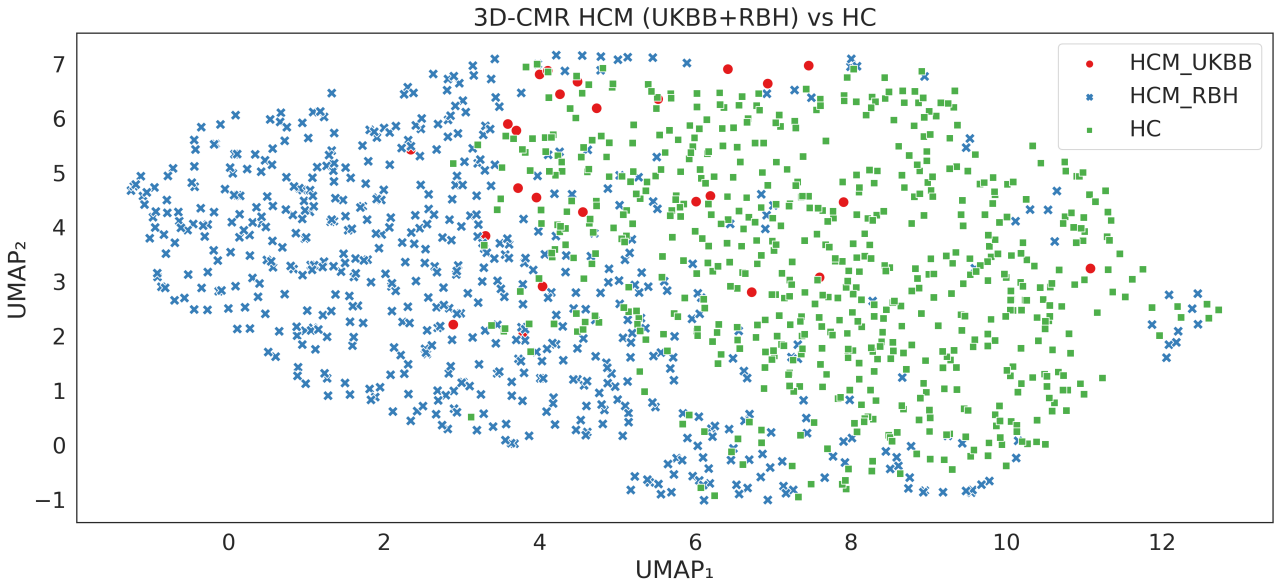


Figure 3.8: 3D-CMR Imaging UMAP embeddings. UKBB Controls are shown in green squares, UKBB HCMs in red dots, and RBH HCMs in blue Xs).

3.5 Discussion and Future Work

Cardiomyopathies have proven to have complex genetic architectures, with a variety of common and rare variants playing key roles in the manifestation of the phenotype. Trying to unlock these architectures and their intricacies by using methods that can handle non-linear relationships is a potential next step in creating omics-derived phenogroups.

The results shown here, highlight the great importance of correlation in genomic data. The stark differences in structures and clusters before and after clumping, emphasize the driving effect of linkage disequilibrium, especially when assessing common SNPs. Leveraging this correlation rather than simply removing it by pruning variants, as is common in current practice (Privé et al.), can open new ways of exploring genomic data with dimensionality reduction methods or other machine learning algorithms. What is also evident is that genomic information solely, whether that is common or rare variation or both, does not drive phenotyp-

ically informative structures similar to what can be seen with the use of other -omics, such as metabolomics in IBS (Mujagic et al., 2022). In the context of cardiomyopathies, profile regression mixture modeling (Liverani et al., 2015), which is a form of Bayesian clustering, has been used with clinical, genetic, CMR, and proteomic data types to inform precision phenotyping groups in DCM (Tayal et al., 2022). In HCM, CMR left ventricular 3D structures can inform genotyping differences via a tree-based dimensionality reduction method (DDRTREE) (Mao et al., 2017) and create phenotypic subtypes based on left ventricular topical differences. The opposite however, i.e. using the genotyping as input for embeddings to inform imaging traits, as was attempted in this chapter, has not been fruitful.

The evidence on non-separation when taking the pruned common SNPs is not surprising, especially if it is compared to the distributions we see in PRS, as it will be discussed in Chapter 6. PRS, which is generated by the effect sizes of such common SNPs tend to aggregate in middle range values and form normal distributions (Zheng et al., 2023a,b). The lack of outliers or different distributional patters can be associated with the lack of spatially separated groups when using non-linear dimensionality reduction methods; due to no strong driving effect from the pruned SNPs.

A number of methods have been developed to mitigate the effects of correlation in genomic loci, i.e. SUPERGNOVA, which estimates local genetic correlations using summary data from GWAS (Zhang et al., 2021). Another method using GWAS summary statistics to decipher genomic correlations is LOGODetect. This method can autonomously locate genetic areas that consistently relate to numerous phenotypes, using a scan statistic approach, and is resilient to overlapping samples from different studies (Guo et al., 2021). Leveraging this new knowledge of interactions based on correlation is useful in the context of employing machine

learning methods to analyze genomic data.

Correlation is also highly evident in the CMR imaging data, where spatial and temporal correlations are a known shortcoming in the acquisition and acceleration of image datasets (Tsao et al., 2003). The 3D mesh points exhibit spatial relationships where points that are closer based on the x,y,z coordinates are expected to be more correlated in terms of wall thickness or other ventricular traits. This is inherent to the process of capturing the images, where there's a notable amount of repeated data over time. Images are captured at one specific moment in the cardiac cycle with the help of ECG-gating and while the patient holds their breath. As a result, the majority of the image remains static. The primary difference between consecutive time frames is the gradual movement of the contrast agent. This repetitive nature of the images allows for data undersampling not just in the spatial domain but also over time (Motwani et al., 2013). Taking these structural biases into account when analyzing these 3D meshes is imperative to avoid erroneous assumptions, especially when assessing the feature importance of these high-dimensional data and want to correctly apply dimensionality reduction on the features.

3.6 Conclusion

The use of genomic datasets as the main driving factor for clustering and deriving phenogroups does not yield novel results. However, the findings presented in this Chapter highlight the great importance of correlation in these genomic datasets. The stark differences in structures and clusters before and after clumping, emphasize the driving effect of linkage disequilibrium, especially when assessing common SNPs. Leveraging this correlation rather

than simply removing it by pruning variants can open new ways of exploring genomic data with dimensionality reduction methods or other machine learning algorithms.

The addition of other -omics datasets or clinical data to the genomics information represents one of the ways for exploring new subgroups within the phenotypes and uncovering differences between cases or individuals at risk. It was confirmed here that imaging data (CMR segmentations) show good cluster separations between HCM cases and Controls with the use of dimensionality reduction methods such as UMAP. The clustering performance across when testing the number of clusters with 2 (HCM and Controls) as ground truth reaches an average accuracy of 0.79, with room for improvement that is leveraged in the next Chapter. Specifically, correlation is not only evident in the genomics dataset, but also in the imaging data; with the latter showing strong spatial correlations that can potentially hinder the efficacy of using the meshes in their entirety. This indicates the need for further dimensionality reduction, while taking into consideration the correlation component.

Chapter 4

LAVASET: A Novel Ensemble Method for Correlated Datasets

4.1 Introduction

Random Forest classifiers (RFs) are frequently used to analyze biological data for prediction and feature selection tasks. RFs can deal with a large number of variables, achieve reasonable prediction scores, and yield highly interpretable feature importance values (Breiman, 2001). As such, they are appropriate models for feature selection and further dimension reduction (DR) for integrated datasets. The premise of the original Breiman RF algorithm is to assemble an ensemble of trees that complement each other and increase the variability of which predictors are selected (Nguyen et al., 2021). However, each node and subsequent split still only considers one predictor variable, which limits both the predictive ability and the correct variable importance assignment in complex biological settings. Nguyen et al., 2021 have recently developed a new information criterion statistic to evaluate a variable's contribution in the predictive ability of the model. It comprises of different levels of probabilities.

In addition, permutation-based variable importance has been extensively studied in RFs. Studies calculating the behavior of variable importance in different scenarios, i.e. collinearity between predictors that are associated with the target outcome, show the level of bias in assignment of variable importance (Nicodemus et al., 2010).

Few techniques have been proposed for enhancing feature importance calculations in datasets with highly correlated variables. The Boruta algorithm (Kursa and Rudnicki, 2010) uses a 'shadow' feature approach where it duplicates the original dataset and shuffles the feature values. Boruta trains a classifier on the enhanced dataset and assigns a value of importance to each of the features. Then the algorithm performs a number of iterations where it compares the importance of the real features to the shadow (shuffled) features, while recording how many times the original features outperform the shadow ones (hits). A

threshold is set and depending on the number of hits the respective feature is assigned an importance or removed from the matrix of further iterations. Boruta in reality encompasses permutation importance, by shuffling the original features, and recursive feature elimination (Guyon et al., 2002), by iteratively considering and removing features that do not reach a threshold.

These methods are efficient in removing noisy features that might not reflect real signals, especially by eliminating through iterations, however they are not sensitive in picking all the correlated variables of an important feature. Addressing relationships between collinear predictors in high dimensional datasets is imperative for reducing the number of variables that are assigned high importance and thereby making the DR more efficient. Here, we propose a novel method (LAtent VAriable Stochastic Ensemble of Trees, LAVASET) that derives latent variables based on distance characteristics of each feature and thereby aim to incorporate the correlation factor in the splitting step (Kasapi et al., 2023). Hence, grouping correlated variables and ensuring similar importance assignment for these.

This chapter will introduce the novel method developed by myself (preprint (Kasapi et al., 2023), under review at journal), which is tested in a variety of -omics datasets, while focusing mostly on the dimensionality reduction and feature importance extraction of the CMR imaging segmented data from the UKBB and RBH cohorts.

4.2 Methods

4.2.1 Datasets

The algorithm (detailed below) is demonstrated on two different imaging datasets with feature importances calculated as a result of a prediction/classification problem between disease and healthy control groups.

Acute myocardial infarction (1D)

Electrocardiogram (ECG) data from the Physikalisch-Technische Bundesanstalt (PTB) dataset (Bousseljot et al., 1995) were downloaded from PhysioNet (Goldberger et al., 2000). Data on acute myocardial infarction (MI) was extracted and compared against no acute MI. All data without a reason for admission or with an unknown diagnosis was excluded. This resulted in 175 control and 346 acute MI ECGs. The individual ECGs were processed to correct signal drifts. An average cardiac cycle was extracted for each individual that was normalized to 750 data points (0.75s). The 3 Frank leads (vector cardiogram, VCG) were used as input to the algorithm and visualize the feature importance in the conventional 12-leads by making use of the (absolute) Kors regression transformation (Kors et al., 1990) for the 8 independent leads.

Hypertrophic Cardiomyopathy (3D)

To test LAVASET's performance on high dimensional, spatial, 3D datasets, we used data meshes derived from Cardiac Magnetic Resonance (CMR) images of the cardiac left ventricle. Segmentation of these images (to produce the meshes) was performed using a deep learning

framework developed by collaborators (Bai et al., 2018). Measurements of myocardial wall thickness were taken along radial segments that connected the inner endocardial and outer epicardial surfaces (Duan et al., 2019). This process produces a 46808 $\times$ 4 matrix where for each of the 46,808 points there is a value for the x, y, z coordinate of the ventricle and the wall thickness. These segmentations were performed on images available by the UK Biobank dataset and an in-house Hypertrophic Cardiomyopathy (HCM) cohort (Royal Brompton Hospital Cardiovascular Biobank). The total number of samples summed to 1,273 (HCM=634, Control=639).

4.2.2 Approach

LAVASET operates given a number of prerequisites and hyperparameters that can be optimized (Figure 4.1). A user-specified distance matrix is calculated to select the k closest feature points to the feature of interest (FOI), which form the FOI neighborhood. This FOI neighborhood submatrix is defined as:

$$FOIn = M_{:,\mathcal{N}(1,k)} \tag{4.1}$$

where M denotes the matrix containing all the input features and the set $\mathcal{N}(1,k)$ represents the indices of the k features in M that are closest to FOI (based on the user-specified distance matrix), including the index for FOI itself. The user specifies the maximum number of features to consider for each split (with default the square root of the total number of features), and these are randomly selected from the entire dataset in each step. For each selected FOI, the first left singular vector (PC1) of the respective FOI neighborhood is calculated. We calculate

this via Singular Value Decomposition (SVD) of the sub-matrix $FOIn$:

$$FOIn = U\Sigma V^T \tag{4.2}$$

Here, $FOIn$ is the scaled FOI neighborhood for the selected VOI, U is the matrix of left singular vectors, Σ is a diagonal matrix containing singular values, and V^T (or V transposed) is the matrix of right singular vectors. The PC1 or first left singular vector is the first vector in matrix U, and this is now the latent variable for the FOIn. Loadings for each latent variable are calculated as:

$$L = U_1 \cdot \ FOIn \tag{4.3}$$

where the first left singular vector U_1 is multiplied by the original $FOIn$ sub-matrix. The input matrix for determining the best split will now consist of the latent variable values instead of the original feature values. The best-split variable and value are evaluated by the traditional Gini index method by deducting the sum of the squared probabilities of each class from one. Once the split occurs, the Gini gain is calculated for the selected latent variable and node by subtracting the sum of the Gini index weights of the two child nodes from the parent node. This is repeated recursively until all samples are split into pure leaf nodes, similar to the classic CART algorithm (Breiman et al., 1984).

4.2.3 Feature importance calculation

Feature importance scores are calculated for each feature by weighing its contribution to the PC score (left singular vector). Specifically, for every selected feature where the PC score is calculated, the loadings vector of the score (with a shape equaling the number of

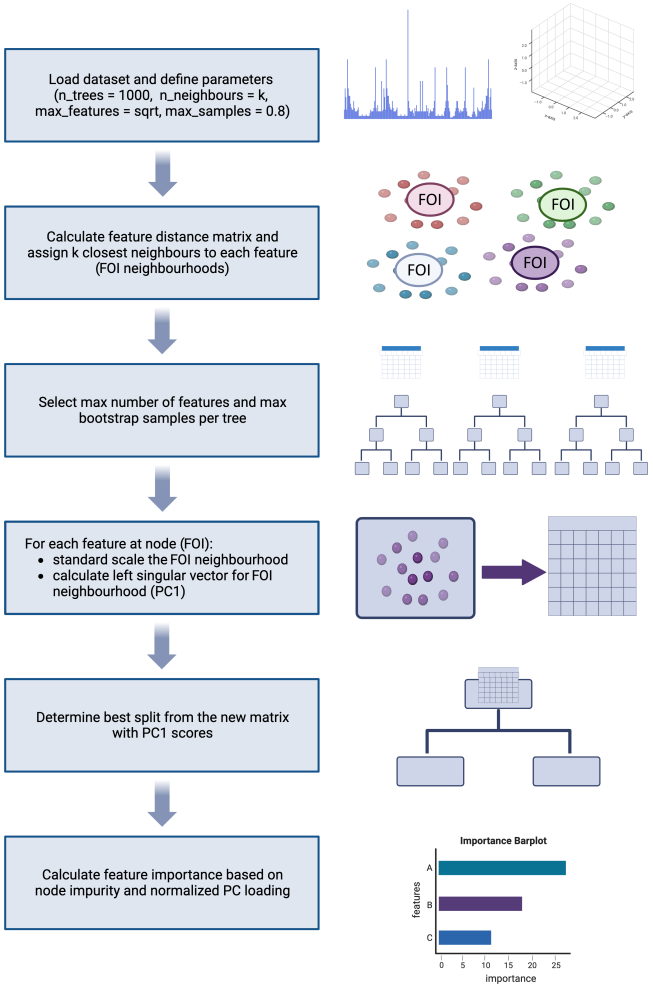


Figure 4.1: LAVASET high-level pipeline indicating the different and novel approach in the node splitting step and feature importance calculation. Model input is customizable and LAVASET can perform on different types of omics data, from 1D to 3D.

neighbors considered) is calculated as shown in Equation 4.3. The loadings vector L is then multiplied by the Gini gain value assigned to the selected latent variable. This results in a feature importance score not only for the FOI but also for its neighbors.

The LAVASET algorithm is designed to allow for multiple variations in the calculation of a feature's importance after model construction. LAVASET outputs for each feature_i a vector of values shown in Equation 4.4, where $\sum_{t=1}^{T} evaluated_t$ represents the summation over all trees from 1 to T of the count of times a feature is evaluated in each tree. Similarly, $\sum_{t=1}^{T} selected_t$ and $\sum_{t=1}^{T} gini_t$ represent the total counts of times a feature is selected for a split and the

accumulated Gini values in each tree, respectively. All features of the original matrix M are equally *evaluated* for a split with a frequency that follows a Gaussian distribution. The subset of features *selected* for splitting a node are those assigned a feature importance.

$$[\sum_{t=1}^{T} evaluated_t \quad \sum_{t=1}^{T} selected_t \quad \sum_{t=1}^{T} gini_t] \tag{4.4}$$

From the values in Equation4.4 we can calculate a normalized Gini feature importance by dividing by the sum of all importances. We also evaluate the ratio of the number of times a feature is selected over the count of times it is considered for a split, and incorporate this in our feature importance final value. Using the same output, another use case would be to compare the ratio of Gini values over the count of times a feature is selected for split, which can give an idea of the value magnitude assigned to a feature at a given split. Results presented in this paper utilize the normalized feature importance multiplied by the ratio of $\sum_{t=1}^{T} selected_t$ over the $\sum_{t=1}^{T} evaluated_t$ for each feature. This is shown in equation 2 where *importance*(i) is the importance of the feature i, T is the total number of trees, *importance*(i, t) represents the Gini importance of the feature *i* in the tree *t*, and *F* is the total number of features. The sum in the numerator of the first fraction goes over all the trees from 1 to *T* for the feature *i*. The sum in the denominator of the first fraction goes over all the trees and all the features (*f* represents each feature in the feature set), which normalizes the Gini of the feature *i*. The ratio of the sums in the second fraction measures the frequency with which feature *i* was selected when it was evaluated.

$$importance(i) = \frac{\sum_{t=1}^{T} gini(i,t)}{\sum_{f=1}^{F}\sum_{t=1}^{T} gini(f,t)} \times \frac{\sum_{t=1}^{T} selected(i,t)}{\sum_{t=1}^{T} evaluated(i,t)} \tag{4.5}$$

4.2.4 Model evaluation

Standard metrics (accuracy, precision, recall, F1-score) are used to compare the classification performance of RFs and LAVASET. For both the simulated and MIBS cohort inputs, LAVASET and RF were run 20 distinct times (20 run pairs with 100 and 1,000 trees, respectively). Identical random state seeds were assigned per run pair for the sample bootstrapping, to account for the randomness between the comparisons. The data were split into training and test sets (80% and 20% respectively), with the sets always being kept identical across the runs. Mean accuracy, precision, recall and F1-scores are reported for the 20 runs.

The neighbors for the VCG data were calculated on the basis of the time of the cardiac cycle. I.e. selecting Frank's lead x for variable (time) i also includes the other 2 leads (y, z) at time i. Including more neighbors takes place in steps of 6 (x, y and z for both i $\pm$ 1). Where the first and last time points are also considered neighbors at i $\pm$ 1.

Finally, we assessed LAVASET's performance on feature extraction for DR, specifically for the 3D CMR dataset. Due to computational constraints, the 3D dataset was tested on 100 and 200 neighbors and 150,000 trees. The neighbors were decided by the x, y, and z coordinates (spatial distance) in an iterative manner by considering the 100 neighbors (for each point) that remain neighbors for over half of the 1,273 samples. The most important features (above the 50% importance value threshold) were selected for LAVASET and RF models (both 150,000 trees), and for each set, we performed further DR and clustering to evaluate how the clusters separated the HCM and Control samples. Uniform manifold approximation and projection (UMAP) and k-means clustering with k=2 (expected number of groups) were used in both cases (McInnes et al., 2018). UMAP components and k-means transformations were evaluated for 20 different random state restarts to ensure robustness. Each restart produced 2 clusters

that were scored by the standard metrics, to determine how closely they matched the true sample classifications.

4.3 Results

4.3.1 CMR Imaging

Both the LAVASET and RF models demonstrate equivalent values across all metrics. LAVASET shows an accuracy of 0.878, precision of 0.837, recall of 0.904 and F1-score of 0.869. RF shows an accuracy of 0.875, precision of 0.916, recall of 0.851, and F1-score of 0.882. Given these high values, we are confident about the use of both models for feature extraction, as means to DR. Figure 4.2 shows the results on 100 neighbors and 150,000 trees. LAVASET is able to assign importance values to a larger surface area of the left ventricle, while also demonstrating the most important areas. RF, on the other hand, picks patches of the left ventricle as most important, disregarding other anatomical parts that could be important. To test how informative these selected features are, we used them as input in further DR and clustering via UMAP and k-means. Table 4.1 shows the clustering performance across 20 iterations of UMAP and k-means, indicating that the features selected by LAVASET perform constantly better than those selected by RF. We also highlight that in this specific dataset, feature extraction improves the models considerably, given the significantly lower performance observed when taking all features as input.

	LAVASET Features	RF Features	All Features
Accuracy	**0.890**±0.01	0.870±0.002	0.797±0.004
Precision	0.905±0.005	**0.913**±0.004	0.817±0.006
Recall	**0.859**±0.009	0.816±0.003	0.763±0.011
F1-score	**0.881**±0.004	0.862±0.002	0.789±0.005

Table 4.1: HCM CMR dataset, feature DR performance. Scores for accuracy, precision, recall, and F1-score across 20 iterations of UMAP and k-means (k=2) on the selected feature sets by importance in LAVASET, RF and without selection. Values shown are the mean ± standard deviation across the 20 iterations.

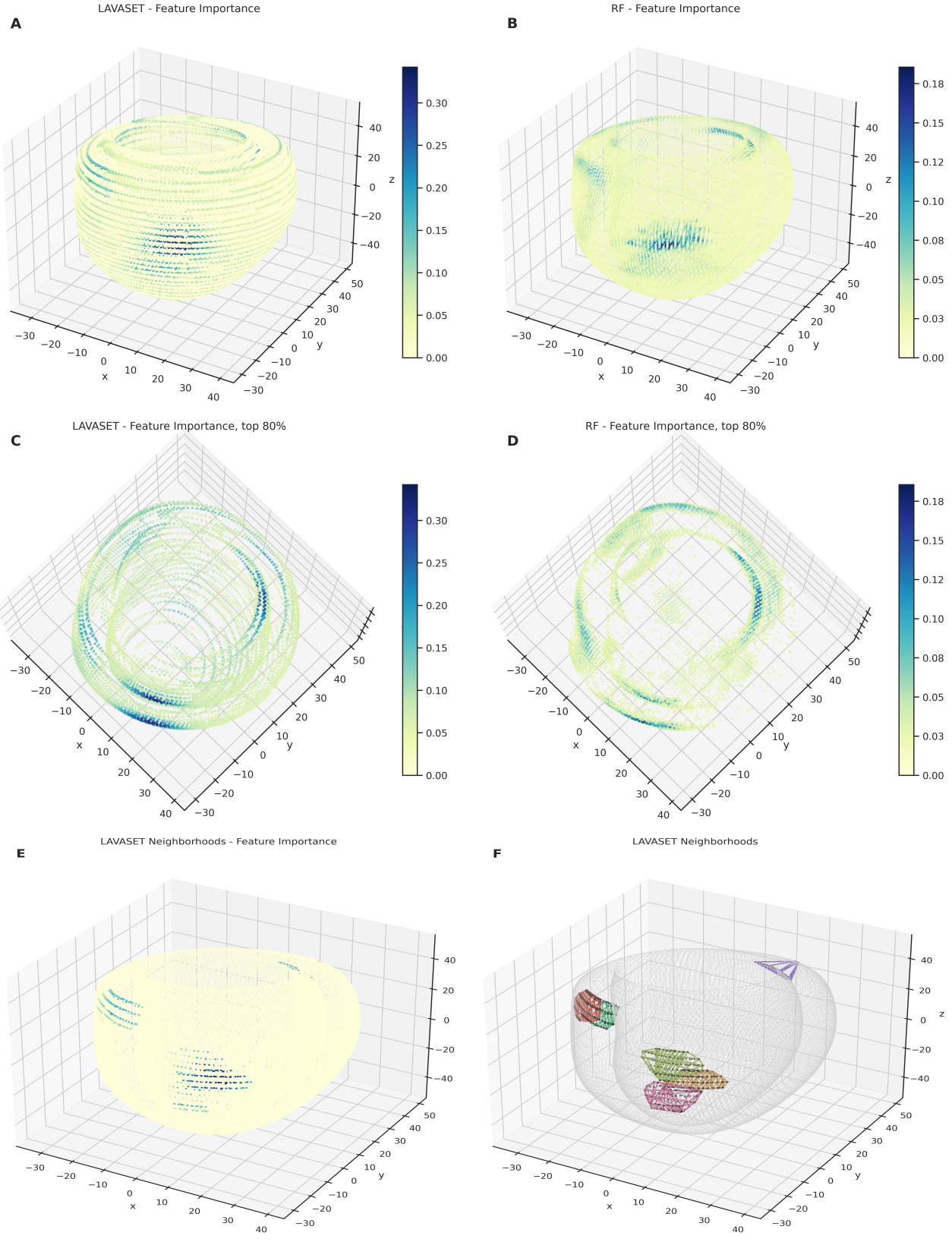

A
LAVASET - Feature Importance
B
RF - Feature Importance
C
LAVASET - Feature Importance, top 80%
D
RF - Feature Importance, top 80%
E
LAVASET Neighborhoods - Feature Importance
F
LAVASET Neighborhoods
x
y
z

Figure 4.2: 3D representations of the left ventricle CMR data points (x, y, z coordinates). Points represent an averaged template of HCM and Control ventricles. The inner and outer structures formed show the endocardium and epicardium, respectively. Colorbar shows the feature importance gradient, indicating that in panel A (LAVASET) the assignment of higher importance is encompassing the entirety of the ventricle structure. In panel B, assignments are given in a patch-like manner for RF. Panels C and D show the points in the 80% quantile from a top view to facilitate the distinction between the inner and outer walls of the ventricle. Panels E and F show 6 distinct neighborhoods of FOIs. In panel D convex hulls are drawn for each neighborhood to represent the pattern of points per neighborhood. Black points indicate the neighbors and the colored connecting lines emphasize the different neighborhoods, and the string-like pattern of neighbor points. Panel C shows the feature importance for the respective 6 neighborhoods.

4.3.2 ECG

We assessed LAVASET's capabilities on more complex data types where it is common practice to use transformations of the raw reading data to infer further information. An example of such data types is transforming Electrocardiogram (ECG) readings to Vectorcardiogram (VCG) inputs. For this task, we used the ECG data from the Physikalisch-Technische Bundesanstalt (PTB) dataset (Bousseljot et al., 1995), as described in Methods. LAVASET-1K performed similarly to the RF-1K in this dataset, with accuracy values showing identical results across 20 runs (0.81 ± 0.01) and other metrics having only small differences. Figure 4.3 shows the results when taking 20 nearest neighbors (across the time axis as described in Methods) and 1,000 trees. LAVASET is able to capture peaks and anomalies across the ECGs that indicate differences between the Control and MI cases. Given the nature of the dataset and the idiosyncrasies of different MI cases based on the location of the infraction, setting a binary classification by incorporating all types of MI as one target will only capture important features related to all these different MI phenotypes.

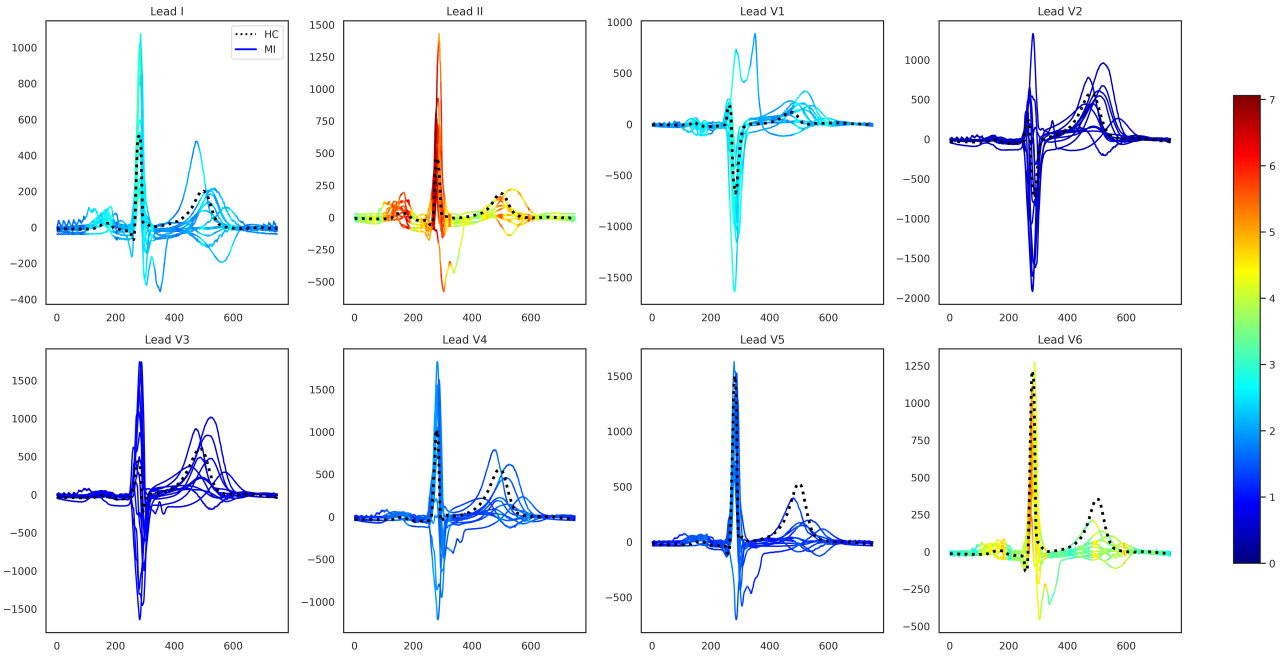


Figure 4.3: ECG feature importance (Kors regression back-transformed from the VCG feature importance) normalized across the 8 independent ECG leads. Subplot titles indicate the respective leads. X-axes show the time in milliseconds (ms) and y-axes the voltage magnitude in millivolts (mV). The dotted black line indicates a healthy sample and the solid lines represent the 10 MI types (acute, anterior, anteriolateral, anteroseptolateral, anteroseptal, inferior, inferolateral, inferoposterolateral, lateral, posterior) in this dataset and are colored by relative feature importance.

4.4 Discussion and Future Work

We have presented a novel ensemble learning method that aims to enhance variable importance and ensure that correlated features are evaluated appropriately and not independently. LAVASET produces non-inferior performance results to traditional RFs in all but one of the examples tested above, and in both simulated and real datasets. While not sacrificing performance in most examples, it is able to assign feature importances in a superior way, by not missing features that are as 'important' by means of correlation, and ensuring that importance values are correctly distributed. In contrast, we have shown that RF allocates feature importance to isolated points, whether that is peak points in a spectrum or points on a 3D mesh, without any clear consistency of the magnitude of importance. Even when

substantially increasing the number of trees in RF, the selection of points and assignment of importance are nearly identical to those selected with the least number of trees, and still do not capture the points selected by LAVASET.

The motivation behind developing LAVASET stems from the idea of enhancing traditional RFs, in a manner similar to how Group Lasso (Yuan and Lin, 2005) enhances the Lasso algorithm (Tibshirani, 1996). In the main premise of Group Lasso, we also assume that there are groups of features that are expected to have similar effects or are related to each other. By penalizing the sum of the absolute values (L1 norm) of the coefficients within each group, Group Lasso encourages the model to select entire groups of features together or exclude them altogether (Yuan and Lin, 2005). Like in Group Lasso, LAVASET is particularly useful when dealing with high-dimensional data where groups of features exhibit similar importance or are structured in some meaningful way. This was evident by the variety of datasets we tested, where the number of features ranged from 750 to 46,000. In this high-dimensional context, LAVASET exhibits stability in its output and remains relatively unaffected by minor variations in the number of neighbors comprising the groups. However, Group Lasso requires non-overlapped groups of features, whereas in LAVASET this can be varied. In fact, in LAVASET, different FOIs can have different numbers of neighbors to increase flexibility. Likewise, LAVASET emulates the kernel filter in convolutional neural networks (CNNs) in that it combines multiple features into a single output (for splitting in LAVASET), however, it does so without condensing the output and attributing the feature importance across the initial features. Other work has investigated the relations between individual features in terms of the similarity of performance at different splits. This methodology is able to discern correlations between individual features, however the mutual forest impact is constrained to evaluating

pairs of features only (Voges et al., 2023). LAVASET can, in theory, be combined with this to perform an a posteriori analysis of evaluating correlations between groups of latent variables.

LAVASET's main limitation derives from the requirement of defining the aforementioned 'groups' (akin to the kernel size in CNNs). This parameter is user-specified and assumes an understanding of the relationship between the variables. In the examples presented in this Chapter, this relationship is defined and assigned by distance, whether that is 1D across time-series data or 3D spatial distance. The influence of the neighborhood definition is evident, especially in the 3D CMR example. Figure 4.2 clearly shows that the string-like pattern of importances calculated by LAVASET is predominantly driven by the assignment of neighbors for each FOI (Figure 4.2F). This inherent limitation, however, is what enables LAVASET's flexibility in creating the groups of neighbors. Distance is only one of the metrics that can be employed. A few other potential examples include genomic distance (combine SNPs via linkage disequilibrium), mass spectrometry isotope patterns (proteomics, metabolomics), or hierarchical relationships (e.g. taxonomy of microbiota). This attribute renders LAVASET more versatile than other similar methods, while giving the user the ability to tailor the algorithm to their specific needs. Hence, it is applicable to a wide variety of datasets and biological questions.

This flexibility is also translated to LAVASET's code implementation. The algorithm also exploits parallelization in order to speed up computations. The body of the code is written in Python 3.10, while utilizing established C++ scripts for efficiency. Given the nature of the code and algorithm, LAVASET can be run in batches, if needed, or can be easily altered to incorporate additional metrics to distance.

To enhance and expand LAVASET's capabilities, we are working on incorporating the

gradient boosting algorithm as one of our built-in additional components. This will extend LAVASET's core methodology to other ensemble methods and benefit from iterative learning and the specific advantages of boosting trees.

Another additional feature to LAVASET's toolkit would be the ability of combining datasets and performing multi-modal classification and feature importance assignment. This has been achieved with the use of Variational Autoencoders (VAEs) where ECG and MRI-based 3D anatomy data were combined in a single model. These two different types of datasets, which in LAVASET have been considered separately, contain structural and myoelectric cardiac information that create a complex interplay that can be useful in improving the diagnostic models (Beetz et al., 2022, Radhakrishnan et al., 2023). VAEs have been successful in capturing the specificities and challenges of each data modality in different branches of the model, while creating a latent space where separation of the ECG and MRI variables is possible. To achieve this cross-modality in LAVASET, one could envision training LAVASET in two steps (similar to the two VAE branches), one for each data modality, so as to create the latent space of neighbors per modality, and then combine the two as the final part of training.

4.5 Conclusion

This chapter presented LAVASET, a novel ensemble method capable to improve feature interpretability by selecting relevant groups of features instead of individual features. Its novel functionality is most useful in datasets with correlations between features. In cases where this is missing from the data, then traditional RFs are more appropriate.

High spatial dependence between features is found in the CMR imaging dataset of the

UKBB. These 3D mesh representations of the left ventricle hold important anatomical and physiological information for the HCM phenotype in a high-dimensional and highly correlated manner. Within this context and due to its novelty, LAVASET is able to assign importance values to a larger surface area of the left ventricle, while also demonstrating the most important areas. RF, on the other hand, picks patches of the left ventricle as most important, disregarding other anatomical parts that could be important.

One of the advantages of RFs in comparison to other classification methods is its ability to assign feature importances to the original features; here, LAVASET retains this advantage by directly assigning importances to the original feature sets, using the loadings derived from singular vector decomposition to guide the relative importances. This renders LAVASET a highly interpretable method, that creates a latest space whose variables directly relate to the original feature space. In comparison to other methods that use graph approaches to incorporate some degree of correlation, LAVASET is uniquely designed to be unaffected by local correlation/neighborhood restrictions and can expand to high-dimensional datasets with spatially coherent features.

LAVASET offers interoperability to the user both by the structure of its code and via the neighbors parameter. It can be applied to almost all -omics data types to identify all relevant known or unknown important features and effectively perform feature extraction for DR.

Chapter 5

Pre-Screening Cardiomyopathies with Machine Learning

5.1 Introduction

For rare cardiovascular conditions such as cardiomyopathies, there is a distinct lack of literature on creating risk models for preliminary screening and prediction [PubMed search with keywords: cardiomyopathy, prediction, screening, machine learning, genomics. 14 related and unrelated results but not encompassing of all terms, as of November 8]. This stands in contrast to more prevalent cardiovascular conditions, which have been extensively studied using either multivariate linear regression models (Siontis et al., 2012) or machine learning approaches (Ahmad et al., 2018, Alaa et al., 2019). Moreover, a number of these studies rely on and incorporate detailed medical data in their training, which is often invasive or not readily accessible in a standard NHS primary record. As we move towards an era where genome sequencing becomes part of routine care, it is imperative to use these personalized data types to streamline patient care and contribute to better diagnosis and prognosis of these diseases. It is time to harness the easily accessible genomic data and clinical biomarkers present in health records to improve the pre-screening of both common diseases as well as rare, life-threatening conditions.

Cardiomyopathies can serve as a prime example of such pre-screening models given their known genetic architecture and its role to disease progression. Our knowledge of the underlying genetic components expands as we move towards evaluating both rare and common genetic variation in these cardiac phenotypes. Together with a population-based cohort such as the UKBB, this can serve as a foundational framework for evaluating the feasibility and efficiency of such pre-screening models. It paves the way for developing automated diagnostic models with high clinical utility.

This chapter will showcase traditional and fully interpretable machine learning models

that can differentiate between individuals with cardiomyopathy from those without, using easily accessible clinical data combined with genomic information.

5.2 Methods

5.2.1 Dataset and basic model architecture

Individuals from the UKBB dataset were chosen for this analysis. A combination of individuals with available WES and CMR data were used in the models, while for case numbers participants with just WES were also included, to preserve a good sample size. To define the HCM and DCM cases in the UKBB, diagnostic codes from ICD (International Classification of Diseases) and linked primary care records were used. The sample size summary is presented in Table 5.1.

	UKBB WES	UKBB CMR	UKBB WES + CMR	RBH
HCM	**581**	37	34	**518**
DCM	**1,262**	46	43	-
Controls	451,413	39,417	**36,393**	-

Table 5.1: Sample breakdown by cardiomyopathy case, Control, and dataset. Numbers highlighted in bold represent the final sample size used in the analyses, after QC, removing missing variables and ensuring availability of all data types.

To ensure robustness of results and optimize the model performances, each classifier was trained 100 distinct times, where each time random training and test splits (80:20 ratio) were done *ab initio*. Splits of sets were tracked to ensure that the same train-test split was assigned across different models. To minimize the issues arising for class imbalance, in each run, the majority class was subsampled (to include a number of samples equal to $\pm$ 100 to the minority class). The resampled data was then split into training and testing sets. Four

different classifiers were trained with this architecture, all described in detail below.

5.2.2 Support Vector Machines

Linear Support Vector Classifier

An SVM is a supervised method that maps samples in a high-dimensional feature space and finds the best separating hyperplane that divides the input in two classes. This was described in detail in 2.1 in Chapter 2. To evaluate the most appropriate model for this context both a linear and a non-linear SVM classifier were trained. For the linear model the kernel is defined by:

$$K(x, x') = x \cdot x' \tag{5.1}$$

Where x and x' are two data points, and $\cdot$ denotes the dot product. The features were auto-scaled using Z-score normalization to ensure that they have a mean of 0 and a standard deviation of 1. The final model architecture is determined through a hyperparameter tuning process, specific to the linear kernel. Specifically, the regularization strength parameter$C$ is optimized over a range of values 1×10^{-11} to 1×10^{-2} on a logarithmic scale. The optimal C value is selected based on maximizing the area under the receiver operating characteristic curve (AUROC) using 5-fold cross-validation. Once the best C value is identified, the linear model is updated with this parameter and trained on the dataset. The model's performance is subsequently evaluated using metrics such as AUROC, precision, recall, F1 score, and precision-recall AUC (see Chapter 2.4.5). The function returns a comprehensive set of results, including these metrics, the best parameters, and feature weights for each run. Feature weights are used in order to calculate the feature importances. In the case of the linear SVM these

weights are the coefficients assigned to each feature when solving the optimization problem as seen in Equation 2.1. The weight vector $\mathbf{w}$ has as many coefficients as there are features in $\mathbf{x}$ where each component corresponds to a feature in $\mathbf{x}$. The magnitude (or absolute value) of each component of the weight vector indicates the importance of the corresponding feature. A larger magnitude indicates a higher importance, and vice versa. The sign of the weight indicates the direction of the relationship between the feature and the class label. A positive weight suggests that as the feature value increases, the prediction moves in the direction of one class, while a negative weight suggests the opposite. In the context of SVM, feature importance does not necessarily indicate causality but rather how influential a feature is in the decision-making process of the classifier. Features with larger weights play a more significant role in determining the position and orientation of the optimal hyperplane.

Non-Linear Support Vector Classifier

In the non-linear SVM model the Radial Basis Function (RBF) kernel was chosen. The RBF computes the similarity between two points based on the Euclidean distance. In this case, the decision boundary can be a curve or a more complex shape, allowing the SVM to capture intricate patterns in biological data contexts. The RBF kernel introduces one additional parameter, the γ parameter, which determines the shape of the decision boundary. A smaller γ value produces a more flexible boundary, while a larger γ results in a more rigid, smooth boundary. The RBF kernel is defined as:

$$K(x, x') = \exp(-\gamma||x - x'||^2) \tag{5.2}$$

Where:

- γ is a positive constant.
- $||x - x'||$ is the Euclidean distance between the data points x and x'.

The features were auto-scaled using Z-score normalization to ensure that they have a mean of 0 and a standard deviation of 1. For the model's hyperparameter tuning, a grid search was conducted over a specified range of values for the hyperparameters C and γ. The range for C was defined using a logarithmic space from 10^{-2} to 10^{10} with 13 values. For γ, the range was defined from 10^{-5} to 10^{3} with 9 values. This resulted in a grid of 117 possible combinations of C and γ. The `GridSearchCV` method from `scikit-learn` was employed to search over the defined hyperparameter grid, similar to the linear model above. The search was conducted using 5-fold cross-validation and aimed to maximize the AUROC. Performance was evaluated by the same metrics as in the linear example.

5.2.3 Random Forest

For the RF model a different set of hyperparameters was optimized, to fit with the specific model requirements. Values ranging from 1,000 to 5,000, with 10 linearly spaced numbers were evaluated for the number of trees. The maximum number of features for split in each node was always selected to be the default RF setting (squared root of the total number of features). Maximum depth values were assessed from 50 to 500, with 5 linearly spaced numbers, as well as removing the depth parameter. Minimum samples split and minimum samples per leaf were assessed for every integer value between 1 and 10. Bootstrap of samples was always activated to allow for calculating the out-of-bag score. The optimal

hyperparameters were identified by looking at which combination maximized the AUROC, through 5-fold cross-validation. Performance metrics such as AUC, precision, recall, F1 score, and out-of-bag score are computed. In the RF algorithm implemented in `scikit-learn`, feature importances are calculated based on the average decrease in the Gini impurity caused by splits over a particular feature, averaged over all trees in the forest. Specifically, when a feature is used to split the data into child nodes, the impurity (or error) of the parent node minus the weighted impurity of the child nodes gives the decrease in impurity due to that feature. This value is accumulated for each feature and then averaged over all trees to give the feature importance. Features that frequently result in purer nodes will have a higher importance.

5.2.4 Gradient-Boosted Trees (GB-Trees)

To optimize the performance of the GB-Trees model, a grid search approach was employed again. The parameters under consideration for optimization were the number of estimators, with values ranging linearly from 100 to 5,000. The learning rate parameter shrinks the contribution of each tree. It was explored with values ranging from 0.01 to 0.2. The chosen values aim to assess the model's sensitivity to the rate of learning, from a slow rate (0.01) to a relatively faster rate (0.2). To measure the quality of a split the mean squared error with improvement score by Friedman or the mean absolute error were assessed. Subsample values were chosen between 0.5 and 0.9 (0.1 step-size), allowing the model to explore the effect of using a subset of the data versus the entire dataset for training each tree. The grid search was performed using 5-fold cross-validation, ensuring a robust evaluation of the model's performance across different subsets of the data.

5.2.5 Genomic Variables - Rare Variants

Individuals were grouped into 4 genetic strata, based on the type of rare variant they carried on HCM or DCM-associated genes. The two pipelines shown below describe how individuals (and variants) are grouped based on criteria specific to HCM and separately for DCM. For each Control and HCM case 4 groups were created based on the HCM pipeline (PHENOCOPY, SARCo, SARCVUS, SARCPLP). And again, separately, for each Control and DCM case another 4 groups were defined (PHENOCOPY, DCMO, DCMVUS, DCMPLP). The pipelines used have been developed and described in de Marvao et al 2021, and were replicated here. Variant annotation in these pipelines was conducted using various tools, including the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) with plugins for gnomAD (Collins et al., 2020), LOFTEE (Karczewski et al., 2020), and SpliceAI (Jaganathan et al., 2019).

HCM pipeline

Variants 100 bp +/- the region of 32 HCM or LVH-associated genes as defined in CardioClassifier (Whiffin et al., 2018) were extracted from the WES data of 454,756 UKB participants. Protein-altering variants variants (PAVs) that had a MAF $<$0.1% in gnomAD and UKBB were identified. Intron variants that were pathogenic in ClinVar (Landrum) were manually curated for functional evidence of splicing. Splice region variants and other splice variants were included if they were predicted to cause splicing through SpliceAI (threshold $>$0.8). Individuals carrying a variant from this list are added to the PHENOCOPY group. The variant list was then shortened to only include the 8 definitive HCM genes (*MYBPC3*, *MYH7*, *TNNT2*, *TNNI3*, *TPM1*, *ACTC1*, *MYL3*, *MYL2*). LOFTEE was used to identify low confidence pLoF variants and identify PAVs that are mislabelled. The variants were then filtered for

disease-causing mechanisms; all *MYBPC3* PAVs were kept, for the other 7 genes, only variants influencing gene product structure (indels, missense, start and stop lost) or gene product level but nonsense-mediated decay escaping, were kept. Individuals carrying a variant from this list are added to the SARCo group.

Variants were further filtered to only include those that met a filtering allele frequency (faf) of $<$0.00004 in gnomAD (more stringent to the MAF $<$0.1% used above). Carriers of the 2,423 variants remaining include a list of individuals that carry a HCM VUS variant that is rare enough to have a substantial effect on disease. After exclusion for DCM variants, 10,209 carriers were identified for the 2,405 variants. These individuals carrying a variant from this more stringent list are added to the SARCVUS group.

Finally, the most filtered list includes variants that would be called pathogenic/likely pathogenic (P/LP) if identified in a patient with HCM; using Cardioclassifier (Whiffin et al., 2018) and ClinVar (Landrum et al., 2014, 2018), the variants were curated for evidence of pathogenicity. Carriers of the 168 variants remaining include a list of individuals that carry an HCM P/LP variant. 929 carriers were identified for the 168 variants. These individuals carrying a variant from this most stringent list are added to the SARCPLP group.

DCM pipeline

Variants located 100 bp +/- the region of 19 genes linked to DCM with varying levels of evidence (definitive, strong, or moderate) were extracted from the WES data of 454,756 UKBB participants. PAVs that had a MAF of less than 0.1% in both gnomAD and UKBB were selected. Intron variants labeled as pathogenic in ClinVar underwent curation for splicing evidence. Only specific splice variants were considered, based on SpliceAI predictions. *Titin*

truncating variants in $PSI > 90\%$ exons or variants with a splice prediction of pathogenic (with assessment) were included. From the first step, 105,507 individuals carried one of the 12,201 remaining variants, and are placed in the PHENOCOPY group.

The list of variants was then refined to include only 12 genes with definitive or strong evidence of causing DCM (*BAG3*, *DES*, *DSP*, *FLNC*, *LMNA*, *MYH7*, *PLN*, *RBM20*, *SCN5A*, *TNNC1*, *TNNT2*, *TTN*). LOFTEE was used to identify low confidence pLoF variants and identify PAVs that are mislabelled. The variants were then filtered for disease-causing mechanisms; all PAVs in the genes *BAG3*, *LMNA*, *PLN*, *SCN5A*, *RBM20*, and *DSP* were kept; only variants influencing gene product structure (indels, missense, start and stop lost) or gene product level but nonsense-mediated decay escaping, were kept for *MYH7*, *DES*, *TNNC1*, and *TNNT2*; only variants predicted to influence gene product level (e.g. frameshift, splice, stop gained) were kept for *TTN* and *FLNC*. Carriers of the 7,549 variants remaining include a list of individuals that carry a DCM VUS variant. After exclusion for HCM variants, 62,292 carriers were identified for the 7,467 variants and were added to the DCMo group.

Further refinement was done to the variant list based on specific allele frequency criteria. The variant list was shortened to only include variants that met a faf of < 0.00004 in gnomAD. Compound heterozygous carriers of common TTNtvs (same two *TTN* variants identified in > 10 individuals) were removed from the analysis as these were likely rescued. After this step, 26,181 individuals carried one of the 6,950 remaining variants, and consisted the DCMVUS group.

The variant list was then filtered further to only include variants that would be assigned a P/LP interpretation if identified in a patient with DCM. The variants were curated for any evidence of pathogenicity in CardioClassifier and ClinVar. 1,351 variants were assessed,

of which 1,018 had strong evidence of pathogenicity in DCM and 82 were identified with evidence for HCM. These 82 were excluded. Finally, 2,195 carriers were identified for the 1,018 P/LP variants, creating the final group of DCMPLP.

Figure 5.1 gives an overview of the chapter's methodology and outputs.

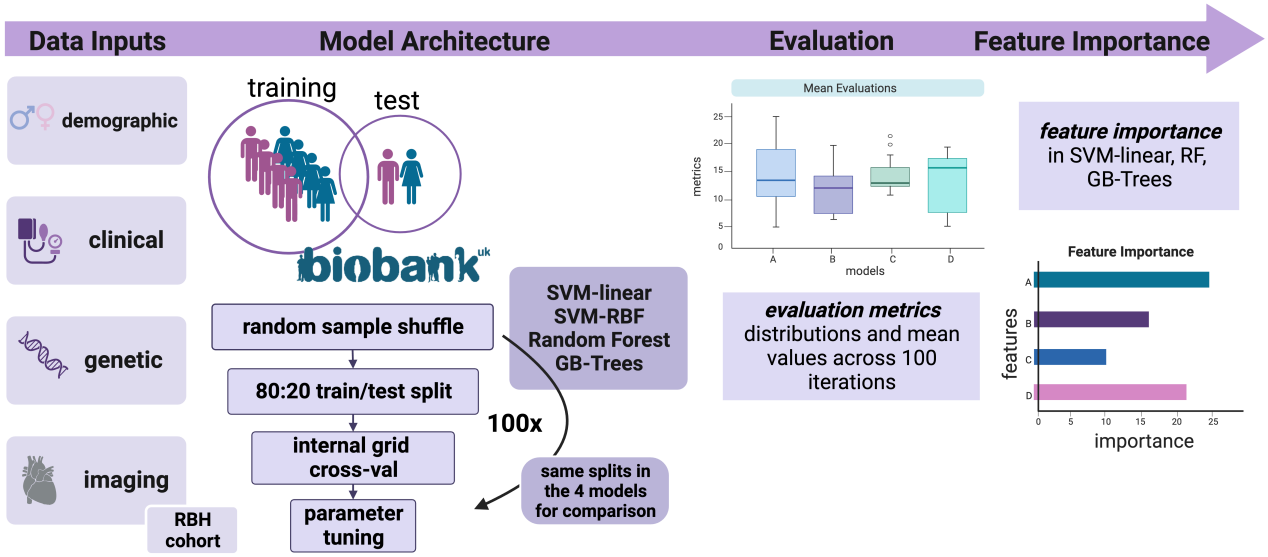


Figure 5.1: Overview of chapter analyses, summarizing the data inputs, model architecture employed, and evaluation outputs.

5.2.6 HCM Models and LAVASET-derived Imaging Features

To assess whether a multi-omics approach of combining genomic, clinical, and imaging features would enhance the HCM model predictions, another set of inputs was tested. LAVASET-derived most important imaging features were incorporated with the rest of the

variables. These most important features included all wall thickness mesh points above the 50% importance value threshold, as derived by the CMR imaging models presented in Chapter 4.5.

Due to number constraints of HCM participants in the UKBB that also have CMR imaging data, the HCM cohort from the RBH was incorporated in these models. The harmonization of the two cohorts results to a decreased number of clinical variables (due to missing values for these in the RBH cohort), with the finalized list of features being: British Caucasian, Sex, Age, Systolic BP, BMI, the genetic variables (PHENOCOPY. SARCo, SARCVUS, SARCPLP), and the 726 LAVASET-derived imaging features (to compare with the complete number of clinical features used in the non-CMR models, see Table 5.2).

To avoid score inflation due to the use of selected features by the same cohort, the test set in these models was kept identical to the test set in the LAVASET model. Controls not incorporated in the LAVASET model for computational cost reasons, were included in the training set. The total number of HCM cases used in training was 506 and in testing 112. The total number of Controls in the training set was 24,954 and 5,040 for the test set (80:20 split). As a result of the constraint to the specific one test set, the 4 models were run once, following the hyperparameter tuning and cross-validation steps described before. The significant class imbalance is taken under consideration during model evaluation, and metrics such as the Precision-Recall curve and average precision (AP 2.4.5) are mostly considered in the Results.

5.3 Results

5.3.1 UKBB Feature Evaluation

Tables 5.2 and 5.3 summarize the list of features used as input in the prediction models for HCM and DCM respectively. For most variables their appearance is common in both sets of models, with the exception of atrial fibrillation which was removed from the DCM models during QC (a skewed majority of DCM cases showed either fibrillation or flutter, leading to erroneous predictions). Demographic features such as sex, age, and race were included, with race in the UKBB being defined as British Caucasian or other. Lifestyle indicators such as exercise, smoking, alcohol intake as well as common cardiovascular or metabolic phenotypes were also included (hypertension, coronary disease, diabetes). Participants in the Controls were chosen to represent a typical population sample, factoring in common predictors like hypertension, without excluding common metabolic or cardiovascular disorders. Specifically, 19% of the Controls participants have hypercholesterol, while for HCM the proportion goes up to 40% and DCM 41%. For hypertension, 29% of Controls exhibit the phenotype, 63% of HCMs, and 67% of DCMs. Systolic blood pressure is consistently higher in both cardiomyopathy groups in comparison to the Controls group (HCM Mann-Whitney p-value$=1.43 \times 10-1$, DCM p-value$=2.26 \times 10-3$) and so is mean diastolic blood pressure, however, with a lower difference than systolic and no statistical significance.

	Controls (n=36,393)	**HCM (n=581)**
Age ($years$)	54.91 ± 7.48	58.50 ± 7.56
Alcohol Intake	16.99 ± 16.55	16.63 ± 17.26
Atrial Fibrillation	1661	241
BMI (Kg/m^2)	26.60 ± 4.22	28.77 ± 4.77
British Caucasian	31754	438
Cardiac Arrest	52	27
Coronary Disease	1608	146
Diabetes	1870	49
Diastolic BP ($mmHg$)	81.47 ± 10.06	82.04 ± 10.87
Handgrip Left (kg)	31.62 ± 11.09	31.36 ± 12.62
Handgrip Right (kg)	33.85 ± 11.15	33.19 ± 12.47
Hba1c ($mmol/mol$)	35.09 ± 5.07	37.29 ± 6.97
Hypercholesterol	6924	236
Hypertension	10409	364
Moderate Activity ($days/week$)	3.36 ± 2.34	3.18 ± 2.53
PHENOCOPY	100	4
Respiratory Failure	146	26
SARCPLP	70	67
SARCVUS	761	107
SARCo	2174	168
Scoliosis	147	6
Sex	17429	382
Smoking	0.53 ± 0.19	0.55 ± 0.19
Stroke	508	36
Systolic BP ($mmHg$)	137.13 ± 18.11	142.87 ± 20.81
Valve Disease	108	44
Vigorous Activity ($days/week$)	1.82 ± 1.87	1.52 ± 2.10

Table 5.2: HCM vs Controls: clinical and genomic features from the UKBB. Integers indicate binary variables and shows the count of presence of respective variable. Continuous variables are presented with meant and standard deviation for each group. Sex variable is count of males. Activity, smoking and alcohol intake are measured based on the UKBB scale: data-fields 884, 914, 1239, 1558.

	Controls (n=36,393)	**DCM (n=1,262)**
Age ($years$)	54.91 ± 7.48	59.88 ± 6.91
Alcohol Intake	16.99 ± 16.55	20.92 ± 23.35
BMI	26.60 ± 4.22	29.37 ± 5.62
British Caucasian	31754	1070
Cardiac Arrest	52	69
Coronary Disease	1608	382
DCMPLP	155	118
DCMVUS	2093	217
DCMo	5144	345
Diabetes	1870	179
Diastolic BP ($mmHg$)	81.47 ± 10.06	81.77 ± 11.73
Handgrip Left (kg)	31.62 ± 11.09	30.92 ± 11.22
Handgrip Right (kg)	33.84 ± 11.15	32.98 ± 11.16
Hba1c ($mmol/mol$)	35.09 ± 5.07	39.16 ± 9.76
Hypercholesterol	6924	517
Hypertension	10409	849
Moderate Activity ($days/week$)	3.36 ± 2.34	3.10 ± 2.61
PHENOCOPY	9356	527
Respiratory Failure	146	81
Scoliosis	147	11
Sex	17416	885
Smoking	0.53 ± 0.19	0.57 ± 0.26
Stroke	52	69
Systolic BP ($mmHg$)	137.12 ± 18.10	139.18 ± 20.78
Valve Disease	6918	517
Vigorous Activity ($days/week$)	1.82 ± 1.87	1.27 ± 2.07

Table 5.3: DCM vs Controls: clinical and genomic features from the UKBB. Integers indicate binary variables and shows the count of presence of respective variable. Continuous variables are presented with meant and standard deviation for each group. Sex variable is count of males. Activity, smoking and alcohol intake are measured based on the UKBB scale: data-fields 884, 914, 1239, 1558.

5.3.2 SVM-linear models can distinguish best between HCM and Controls

Figure 5.2 shows the results of the 4 different ML methods, across the 100 distinct runs, when looking at HCM vs Controls prediction models. The test set samples were changing for each run, but being kept the same across the 4 models to allow for comparison, as described in Methods. Figure 5.2 presents the distributions of scores across the 100 runs while Table 5.4 shows the mean and standard deviation for the 100 runs. Across 3 evaluation metrics (AUROC, f1-score, and recall) the SVM with the linear kernel seems to perform marginally better than the other 3 methods, whereas in precision the RBF kernel seems to be performing by 0.004 better, however not significant as indicated by the *p*-value of 6.51×10^{-1}. The SVM-linear AUC scores were statistically different to the ensemble methods (RF, GB-trees), as well as significantly higher than all others for f1-scores and recall.

Model	**AUROC**	**F1-Score**	**Precision**	**Recall**
SVM-linear	0.867 ± 0.022	0.774 ± 0.032	0.814 ± 0.042	0.740 ± 0.036
SVM-RBF	0.866 ± 0.023	0.762 ± 0.036	0.818 ± 0.043	0.714 ± 0.046
RF	0.858 ± 0.023	0.757 ± 0.03	0.802 ± 0.046	0.718 ± 0.038
GB-Trees	0.857 ± 0.023	0.758 ± 0.032	0.813 ± 0.046	0.712 ± 0.04

Table 5.4: Performance metrics for HCM vs Controls models. First value indicates the mean of the 100 distinct runs and second value the standard deviation.

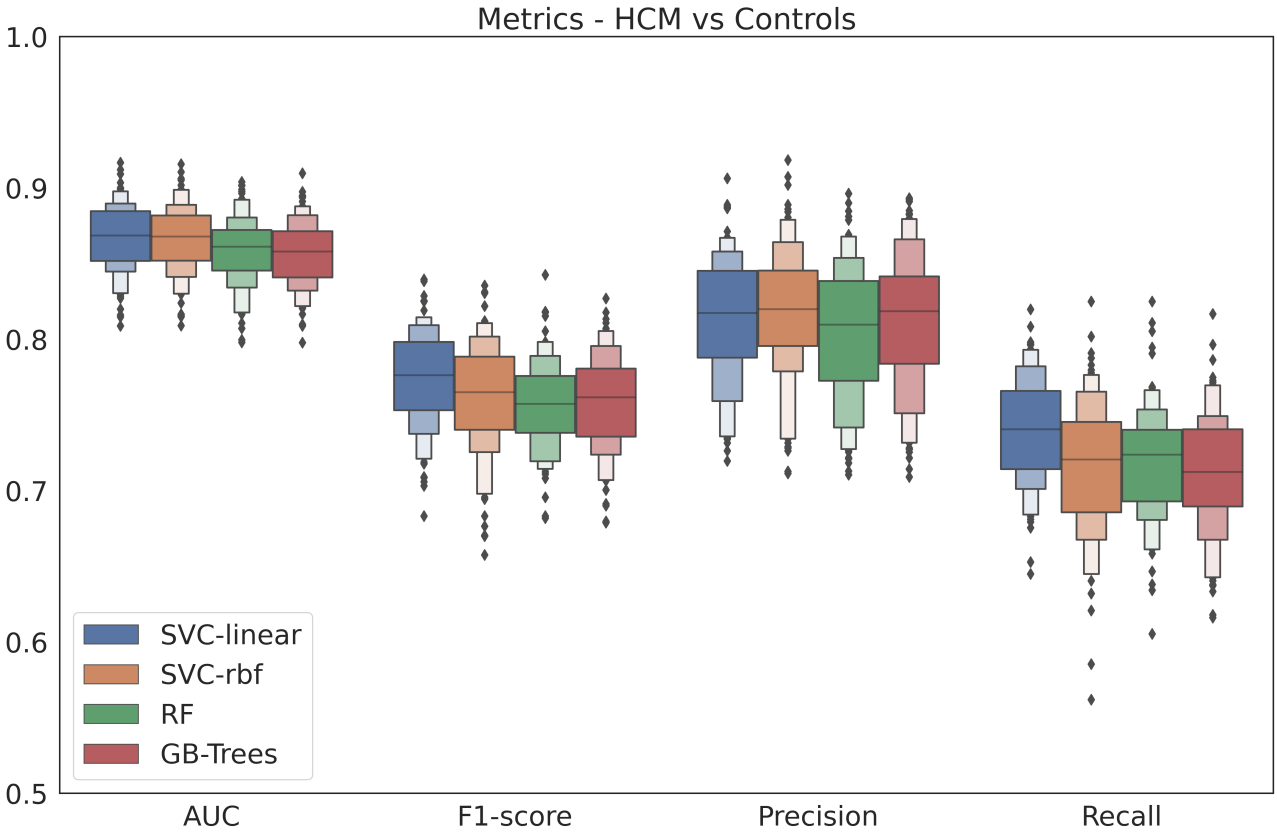


Figure 5.2: Performance evaluation metrics for the 4 different ML methods, in the HCM vs Controls prediction models. Boxplots show the distribution of values across the 100 distinct iterations. Blue indicates the Support Vector Machine (SVM) with linear kernel (SVM-linear), orange the SVM with the radial basis function kernel (SVM-RBF), green the Random Forest (RF), and red the Gradient Boosting Trees (GB-Trees) results. Median line is shown in black within the box. *P*-values are as follows: **AUC**: *SVM-linear vs SVM-RBF* $= 6.51 \times 10^{-1}$, *SVM-linear vs RF* $= 7.60 \times 10^{-3}$, *SVM-linear vs GB-Trees* $= 1.03 \times 10^{-3}$, *SVM-RBF vs RF* $= 2.85 \times 10^{-2}$, *SVM-RBF vs GB-Trees* $= 5.30 \times 10^{-3}$, *RF vs GB-Trees* $= 5.65 \times 10^{-1}$. **F1-score**: *SVM-linear vs SVM-RBF* $= 1.09 \times 10^{-2}$, *SVM-linear vs RF* $= 9.86 \times 10^{-5}$, *SVM-linear vs GB-Trees* $= 5.57 \times 10^{-4}$, *SVM-RBF vs RF* $= 2.95 \times 10^{-1}$, *SVM-RBF vs GB-Trees* $= 4.72 \times 10^{-1}$, *RF vs GB-Trees* $= 7.45 \times 10^{-1}$. **Precision**: *SVM-linear vs SVM-RBF* 4.65×10^{-1}, *SVM-linear vs RF* $= 7.58 \times 10^{-2}$, *SVM-linear vs GB-Trees* $= 9.03 \times 10^{-1}$, *SVM-RBF vs RF* $= 1.40 \times 10^{-2}$, *SVM-RBF vs GB-Trees* $= 4.12 \times 10^{-1}$, *RF vs GB-Trees* $= 1.12 \times 10^{-1}$. **Recall**: *SVM-linear vs SVM-RBF* $= 2.51 \times 10^{-5}$, *SVM-linear vs RF* $= 5.91 \times 10^{-5}$, *SVM-linear vs GB-Trees* $= 9.17 \times 10^{-7}$, *SVM-RBF vs RF* $= 5.35 \times 10^{-1}$, *SVM-RBF vs GB-Trees* $= 7.39 \times 10^{-1}$, *RF vs GB-Trees* $= 2.99 \times 10^{-1}$.

5.3.3 Important features in HCM models include genetic and clinical characteristics

Figure 5.3 presents the feature importance values generated by the SVM-linear (top, blue) and RF (bottom, green) HCM vs Controls models. The bars show the distribution of feature importance scores as they were generated for each of the 100 distinct iterations. Among the most important features selected by the SVM-linear models are presence of atrial fibrillation or flutter and having a SARCPLP variant. Sex, presence of coronary or valve disease, hypertension, episode of cardiac arrest, having SARVUS and SARCo variants, and BMI also show a relatively higher importance towards predicting for HCM cases. A higher systolic blood pressure and blood glucose level is also associated with the HCM cases prediction. Lighter blue boxes indicate the negative feature importances, usually associated with the 0, in this case Controls, class label.

Being British Caucasian shows up as most important for predicting the Controls participants, potentially driven by the higher percentage of British Caucasians in the Controls group (87%) in comparison to the HCM group (75%). Right hand grip is ranked as the next most important, and it does indeed show a higher mean in the Controls group than the HCM group as per Table 5.2. The difference however is marginal which is also supported by the absolute value of the importance which is less than 0.1. The rest of the features shown in the graph are all within the limits of $[-0.1, 0.1]$ indicating that those features are relatively less important in the predictions.

RF feature importances do not show a direction of importance towards the labels, and present importance values based on the Gini index calculation of each split, as described in

Chapter 4.5. A key difference between the SVM-linear and RF results is the high importance of age shown in the latter. Moreover, the SARCPLP variants are ranked as less important by the RF model in comparison to the SVM-linear model. This is an interesting discrepancy, given that even in terms of absolute sizes (without looking at the sign of the score), the SARCPLP feature is assigned a very high importance in the SVM model, while in the RF it is assigned an importance almost a third of the higher importance value. Systolic and Diastolic BP measures show very similar importances, in line with what it was shown in the SVM-linear results when taking the absolute values of coefficients.

Similar results on feature importances are also shown in the GB-Trees, with AF, SARCPLP and hypertension being the dominant features. Barplots are shown in Appendix Figure 7.1.

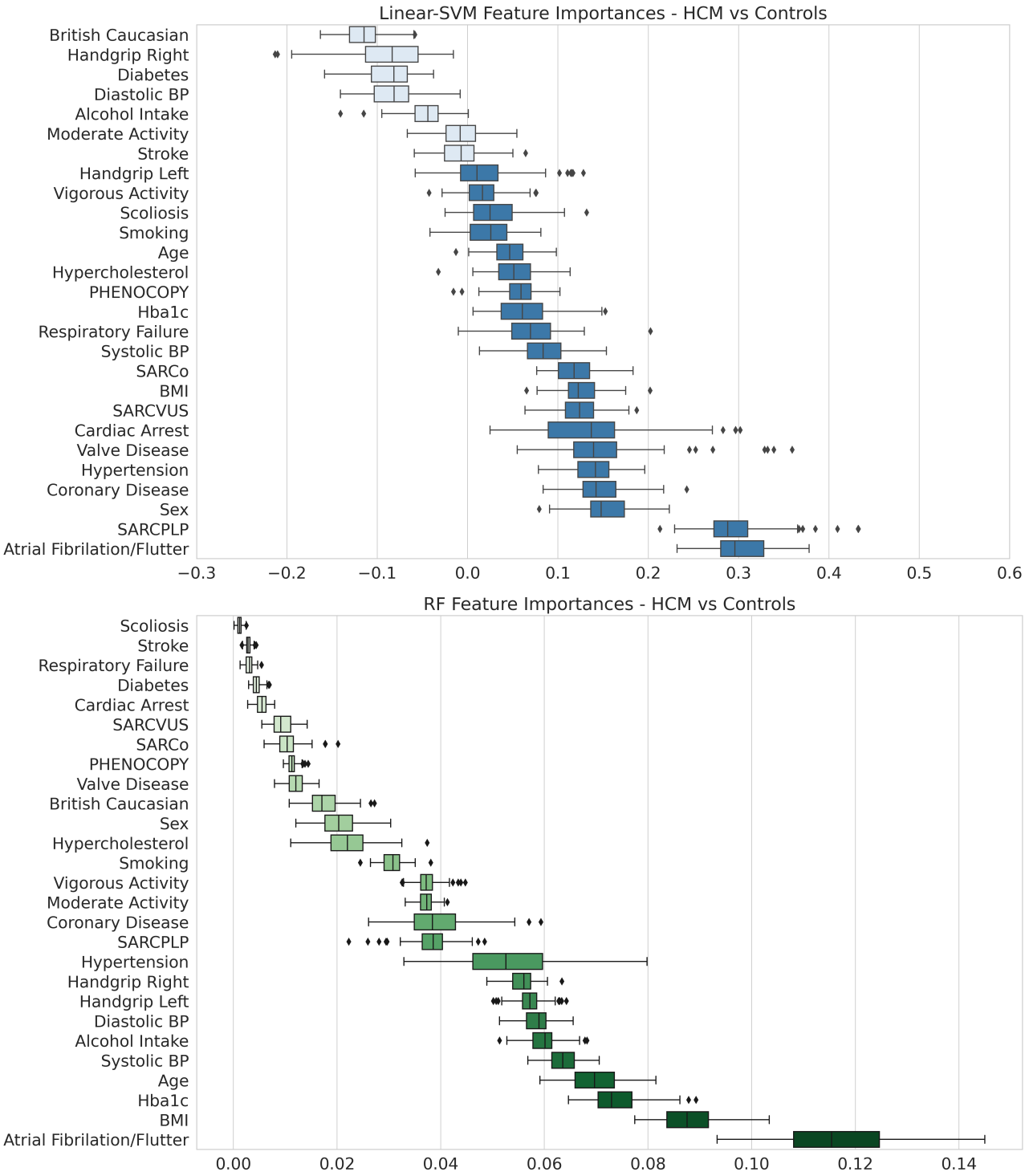


Figure 5.3: Feature Importances for the HCM vs Controls prediction models. Bars show the distribution of feature importance scores as they were generated for each of the 100 distinct iterations. Top graph shows the results from the SVM-linear model. Lighter blue boxes indicate the features with a negative median value across the 100 iterations, therefore contributing more to the Controls label prediction. Darker blue boxes indicate the features with a positive mean value, which contribute more to the HCM label prediction. Bottom graph in shades of green shows the results from the RF models, based on the Gini index value feature importance. The Sex variable was coded based on the UKBB data encoding, with 1 depicting males and 0 depicting females.

5.3.4 SVM models can distinguish best between DCM and Controls

Figure 5.4 shows the results of the 4 different ML methods, across the 100 distinct runs, when looking at DCM vs Controls prediction models. The test set samples were changing for each run, but being kept the same across the 4 models to allow for comparison, as described in Methods. Figure 5.4 presents the distributions of scores across the 100 runs while Table 5.5 shows the mean and standard deviation for the 100 runs. In contrast to the HCM vs Controls models, here, both SVM kernels (linear and RBF) have the highest scores in AUROC and non-significantly different values in f1-scores. Precision scores have slight differences across the 4 different methods but without statistical significance being shown (Figure 5.4).

Model	AUROC	F1 Score	Precision	Recall
SVM-linear	0.851 ± 0.028	0.737 ± 0.038	0.756 ± 0.05	0.723 ± 0.05
SVM-RBF	0.851 ± 0.026	0.734 ± 0.037	0.766 ± 0.051	0.709 ± 0.056
RF	0.841 ± 0.027	0.724 ± 0.035	0.766 ± 0.05	0.69 ± 0.052
GB-Trees	0.832 ± 0.029	0.717 ± 0.037	0.759 ± 0.055	0.684 ± 0.054

Table 5.5: Performance metrics for DCM vs Controls models. First value indicates the mean of the 100 distinct runs and second value the standard deviation.

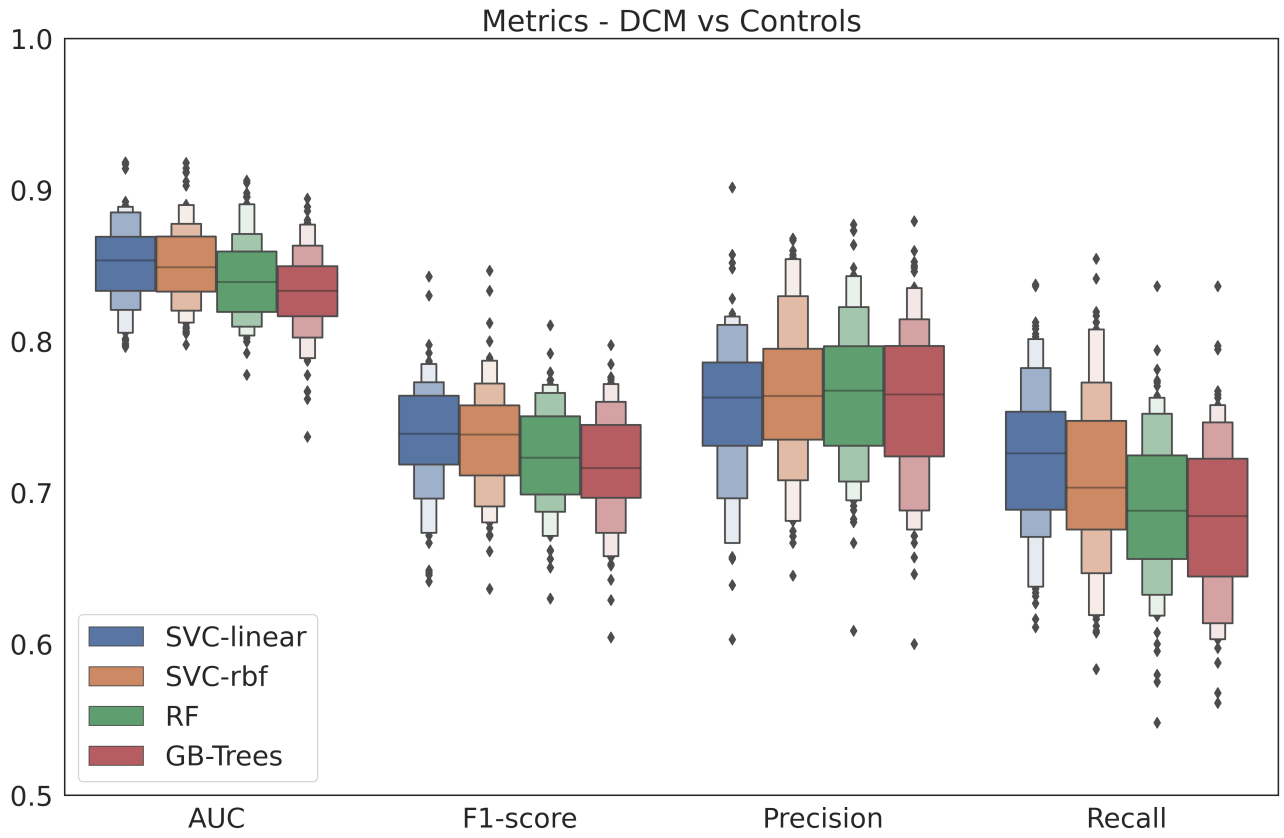


Figure 5.4: Performance evaluation metrics for the 4 different ML methods, in the DCM vs Controls prediction models. Boxplots show the distribution of values across the 100 distinct iterations. Blue indicates the Support Vector Machine (SVM) with linear kernel (SVM-linear), orange the SVM with the radial basis function kernel (SVM-RBF), green the Random Forest (RF), and red the Gradient Boosted Trees (GB-Trees) results. Median line is shown in black within the box. *P*-values are as follows: **AUC**: *SVC-linear vs SVC-RBF* $= 9.88 \times 10^{-1}$, *SVC-linear vs RF* $= 8.71 \times 10^{-3}$, *SVC-linear vs GB-Trees* $= 2.27 \times 10^{-6}$, *SVC-RBF vs RF* $= 6.73 \times 10^{-3}$, *SVC-RBF vs GB-Trees* $= 1.18 \times 10^{-6}$, *RF vs GB-Trees* $= 2.21 \times 10^{-2}$. **F1-score**:*SVC-linear vs SVC-RBF* $= 5.81 \times 10^{-1}$, *SVC-linear vs RF* $= 9.01 \times 10^{-3}$, *SVC-linear vs GB-Trees* $= 1.99 \times 10^{-4}$, *SVC-RBF vs RF* $= 3.82 \times 10^{-2}$, *SVC-RBF vs GB-Trees* $= 1.29 \times 10^{-3}$, *RF vs GB-Trees* $= 2.00 \times 10^{-1}$. **Precision**: *SVC-linear vs SVC-RBF* $= 1.53 \times 10^{-1}$, *SVC-linear vs RF* $= 1.65 \times 10^{-1}$, *SVC-linear vs GB-Trees* $= 6.80 \times 10^{-1}$, *SVC-RBF vs RF* $= 9.58 \times 10^{-1}$, *SVC-RBF vs GB-Trees* $= 3.38 \times 10^{-1}$, *RF vs GB-Trees* $= 3.61 \times 10^{-1}$. **Recall**: *SVC-linear vs SVC-RBF* $= 7.53 \times 10^{-2}$, *SVC-linear vs RF* $= 1.11 \times 10^{-5}$, *SVC-linear vs GB-Trees* $= 4.22 \times 10^{-7}$, *SVC-RBF vs RF* $= 1.35 \times 10^{-2}$, *SVC-RBF vs GB-Trees* $= 1.55 \times 10^{-3}$, *RF vs GB-Trees* $= 4.35 \times 10^{-1}$.

5.3.5 Important features in DCM models include genetic and clinical characteristics

Figure 5.5 presents the feature importance values generated by the SVM-linear (top, blue) and RF (bottom, green) DCM vs Controls models. The bars show the distribution of

feature importance scores as they were generated for each of the 100 distinct iterations. The first set of important features contributing to DCM with median values between 0.1 and 0.2 across the 100 runs, are sex, presence of coronary or valve disease, incident of cardiac arrest or respiratory failure, hypertension, age, BMI, blood glucose level and having DCMPLP variants. Higher alcohol intake and presence of DCMVUS variants have median values close to 0.05. DCMo variants, diastolic BP, having hyper-cholesterol and higher moderate activity show some small contribution to the DCM class, while also contributing to the Controls class for a majority of iterations as indicated by the error bars.

The lighter-colored bars show importances for features that are more Controls-contributing. The most pronounced feature in this category is the right hand grip measurement, which when looking at he magnitude of importance on the absolute value, it is still not equivalent to the top features of the positive group. Diabetes and vigorous activity are two additional features that show higher Controls-contribution to almost the entirety of runs. Although the left hand grip strength and systolic blood pressure typically show a negative median, the error bars suggest that they were assigned a positive importance value in several iterations (DCM contribution).

The bottom graph of Figure 5.5 presents the importances assigned by the RF model. Age, BMI, and blood glucose levels rank as the most significant features. Following them are hypertension, systolic and diastolic blood pressure, strength of both left and right hand grips, and alcohol consumption. The genomic features (DCMPLP, DCMVUS, DCMo, PHENOCOPY) are ranked in lower places and have been assigned values less than the 50% threshold (age median $= 0.9$) of the total importances.

Similar results on feature importances to the SVM-linear are also shown in the GB-Trees,

with hypertension, coronary artery disease, age, and BMI being dominant features, while sex is found further down in the ranking. Barplots are shown in Appendix Figure 7.1, bottom figure.

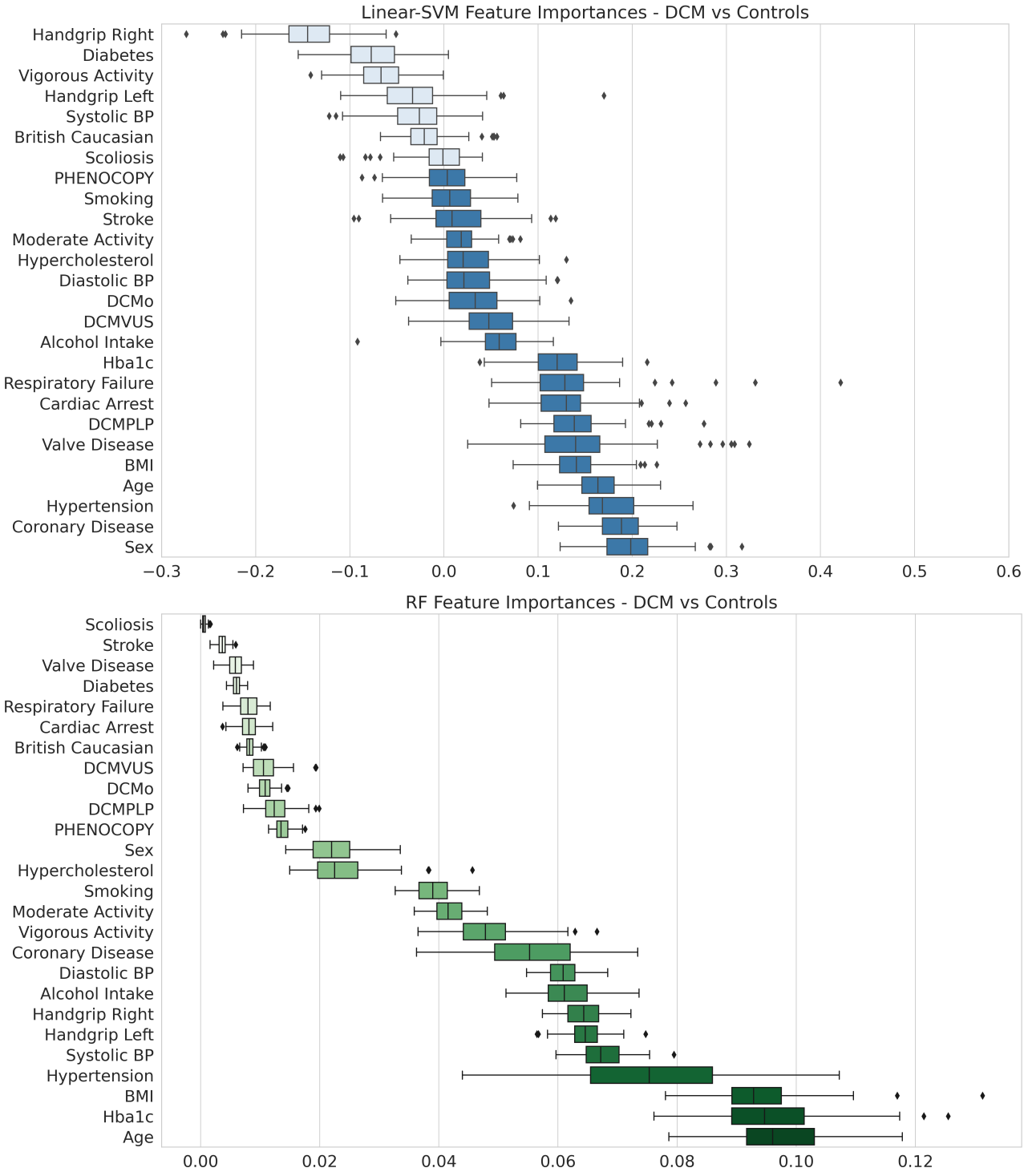


Figure 5.5: Feature Importances for the DCM vs Controls prediction models. Bars show the distribution of feature importance scores as they were generated for each of the 100 distinct iterations. Top graph shows the results from the SVM-linear model. Lighter blue boxes indicate the features with a negative median value across the 100 iterations, therefore contributing more to the Controls label prediction. Darker blue boxes indicate the features with a positive mean value, which contribute more to the DCM label prediction. Bottom graph in shades of green shows the results from the RF models, based on the Gini index value feature importance. The Sex variable was coded based on the UKBB data encoding, with 1 depicting males and 0 depicting females.

5.3.6 HCM and DCM P/LP variants and common clinical factors show importances in both phenotypes

Figure 5.6 shows the importance values intersection between the HCM and DCM models. The top figure presents results from the two SVM-linear models, indicating that HCM and DCM associated P/LP variants are the most important features driving predictions for the two diseases, along with sex, BMI, and presence of cardiovascular related factors. The figure below shows the mean of importances of the two ensemble methods (RF and GB-Trees). Ensemble methods indicate that hypertension is the most importance feature separating the cardiomyopathies from the Controls, while P/LP variants show a stronger importance in the HCM ensemble models than the DCM. Finally right hand grip strength presents as a strong predictor for Controls in both HCM and DCM SVM-linear models.

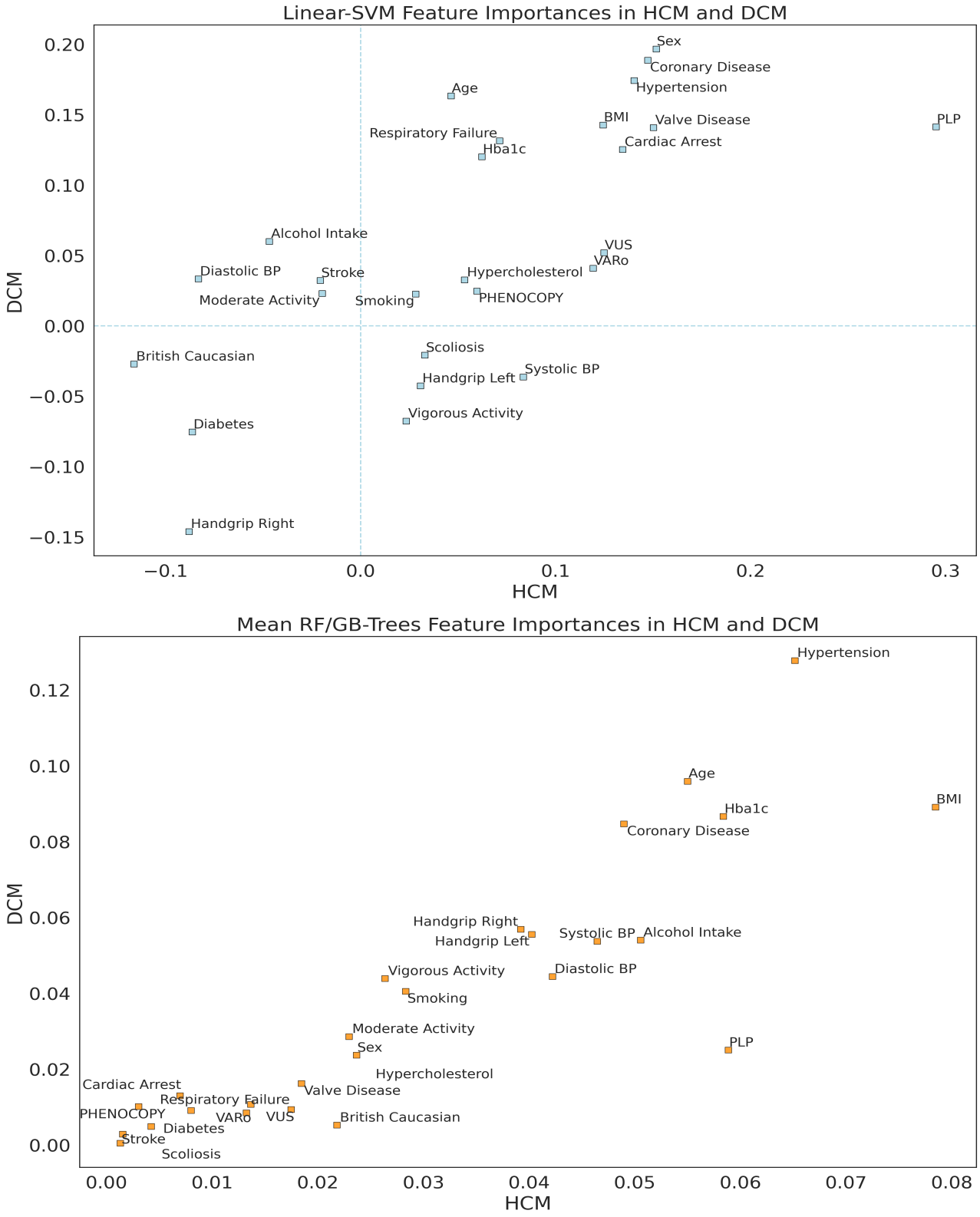


Figure 5.6: Feature Importance values intersection between the HCM and DCM prediction models. Top graph shows the results from the SVM-linear model. Bottom graph shows the results averaged from the RF and GB-Trees models, based on the Gini index value feature importance.

5.3.7 HCM models including LAVASET-derived CMR imaging features perform better

Table 5.6 shows the metrics from the HCM vs Controls model when including the important features derived from LAVASET. The first 4 metrics show results from the test set, whereas the average precision (AP) quantifies the relationship between the precision and recall at different thresholds while training. In all metrics the SVM-RBF outperforms the other methods. The ensemble methods have lower values on recall, and perfect precision scores, indicating that the models with CMR features mislabel HCM samples as Controls, increasing the number of false negatives. Direct comparisons cannot be made with the HCM vs Controls models presented before given the different training/test split architecture and the additional cohort; however, in the best performing model (SVM-RBF) we do see that all metrics are relatively higher when CMR features are included. Figure 5.7 presents the Precision-Recall curves for each model to facilitate the comparison in different thresholds while training. The AP is shown in the legend and indicates that the model performing best in the test set (SVM-RBF), is also the one with the biggest area under the curve in the Precision-Recall curve.

Model	AUROC	F1 Score	Precision	Recall	AP
SVM-linear	0.937	0.825	0.879	0.777	0.908
SVM-RBF	0.978	0.889	0.923	0.857	0.956
RF	0.966	0.466	1.00	0.304	0.939
GB-Trees	0.943	0.756	1.00	0.607	0.940

Table 5.6: Performance metrics for HCM vs Controls models, when incorporating the LAVASET-derived CMR imaging features.

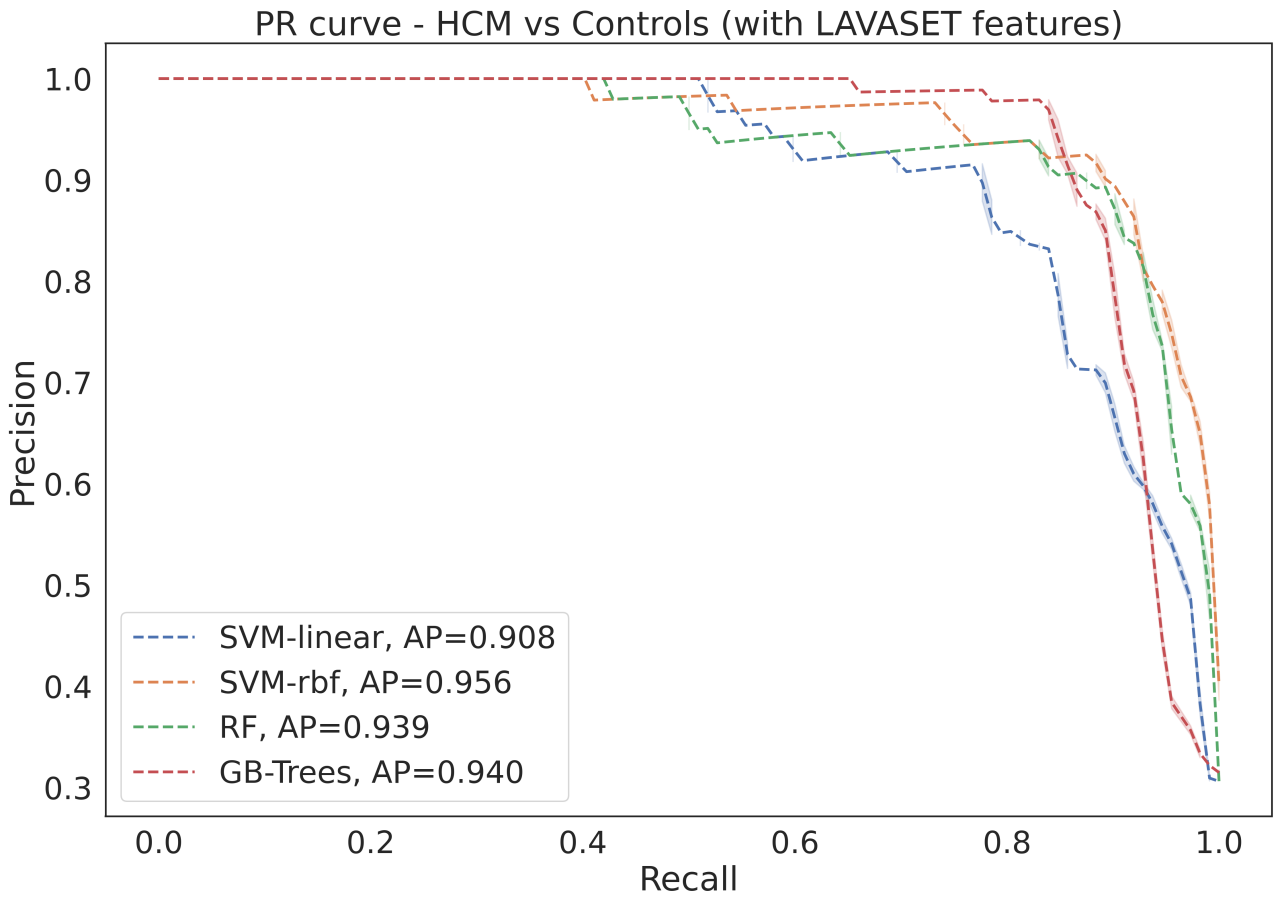


Figure 5.7: Precision-Recall curve for the 4 models. Average Precision (AP) is calculated and shown in legend for each model.

5.4 Discussion and Future Work

Cardiomyopathies are known to show a significant genetic component. This can be attributed to either a single or multiple rare variants influencing the phenotype (Marian and Braunwald, 2017, Mazzarotto et al., 2020) or, based on recent discoveries, to the effect of common variants (Zheng et al., 2023a,b). As genome sequencing becomes a standard part of routine care (NHS, 2023), it's crucial to leverage this personalized data to enhance patient care, leading to improved diagnosis and prognosis of these conditions.

Using existing data from patient records to pre-screen for rare, burdensome diseases, without the need for expensive tests at first level, can refine subsequent testing and boost the efficiency of referrals. Therefore, computational models that can triage, at an initial stage, potential patients and give a possible assessment and risk for cardiomyopathy, are a future possibility.

5.4.1 ML models can differentiate between CMs and Controls

This chapter demonstrated that traditional machine learning methods can differentiate between individuals with cardiomyopathy and those without, utilizing readily available clinical data and genomic rare variant information. Specifically, in the HCM vs Controls models, the best performing method, SVM-linear, is able to identify 3 out of 4 HCM cases correctly, missing however about 25% of the cases. The precision of the model at identifying the HCM cases is higher to its specificity, where about 4 out of 5 HCM-predicted cases, are true HCM patients. Having a higher precision, and subsequently a lower false positive rate (FPR=14.6%) is beneficial in a pre-screening tool for various reasons. A high FPR means that many individuals

without HCM will be incorrectly identified as having it. This can cause unnecessary stress, anxiety, and subsequent medical follow-ups. With an FPR of 14.6% less than 2 out 10 non-risk individuals will be mislabeled as at risk, making this SVM-linear model a potential candidate for a future pre-screening diagnostic tool.

For the DCM models, the results were not as clear as to which model outperformed all others, given AUCs were tied for linear and RBF SVMs. Recall was significantly higher in the SVM-linear model, while precision was not statistically significant between the two SVM methods, as indicated by the *p*-values in Figure 5.4 caption. In comparison to the HCM models, there is a relative decrease in performance of DCM vs Controls which is evident in all the metrics presented. Specifically, the SVM-RBF shows a recall (or sensitivity) of 0.72, meaning that the model identified correctly 72% of the participants with DCM. A precision of 0.75 means that in 75% of the instances where the model classified a participant as having DCM, that classification was true. The DCM SVM-linear model shows an FPR of 18.7% indicating that out of the total of Controls, the model incorrectly predicted approximately 18.7% of them as DCMs. With a specificity lower than in the HCM models (85.4% for HCM, 81.3% for DCM), in this case there is a higher chance of screening Controls participants incorrectly and labeling them as at risk for DCM.

For rare cardiovascular diseases like cardiomyopathies, there is a noticeable gap in the literature when it comes to establishing baseline models for pre-screening and prediction. This is in contrast to the more common cardiovascular diseases, where there is a plethora of research employing machine learning techniques. A study by Hossain et al. (2023) published a heart disease risk prediction model that exhibits an accuracy of 94.9% using the RF method. Weng et al. (2017) compared different machine learning models for predicting cardiovascular

event risk with their best performing model, a neural network, achieving an AUC of 76.4% with sensitivity of 67.5% and specificity of 70.7%. Beunza et al. (2019) also presented a performance comparison of multiple machine learning algorithms, drawing data from the Framingham heart database, to predict coronary heart disease. They achieved the highest AUC value of 0.75 with the support vector machine approach. Shah et al. (2020) reported a peak accuracy of 90.78% in their heart disease predictions using the K-NN algorithm. Lastly, Jindal et al. (2021) trained an RF model for heart disease risk prediction which demonstrated an accuracy of 86.9%, a sensitivity of 90.6%, and a specificity of 82.7%. The metrics showcased here for the HCM and DCM models align with the findings from these studies, without however being able to make direct comparisons as to whether these models are better.

Another novel and unique characteristic of this methodology relies on the multiple iterations for creating robust predictions. This is necessary given the nature of features, especially due to the rare variant information incorporated. Rare variants, as the name indicates, are not found in a large number of participants, in some cases even having variants present only in one or two individuals. Therefore, to prevent the issue of a biased test and training set where a number of these rare variants are potentially completely excluded, the train-test split shuffling across 100 iterations was introduced. In this way, the model is able to train in an out-of-bag manner, create a robust learning, and mitigate issues such as the rare variable instance. Rare variants, however, are not the sole genetic component for cardiomyopathies. They account for less than 50% of cases (Verdonschot et al., 2020) and recent GWAS efforts have shown that SNP heritability has a strong polygenic influence on the phenotype which can account for the rest of the cases. To address this, a future consideration would be to include PRS as an additional feature to the models. This would be possible in

cases where whole genome sequencing or genotyping is offered to participants or outpatients, and not just specific genetic panels that are disease focused. In the given study, PRS was not initially included because of sample size constraints. Since scores were available for only around 70% of the participants, incorporating them would have substantially reduced the number of cases.

5.4.2 Models are interpretable and important features can be evaluated

SVMs and ensemble methods, such as RFs and GB-Trees, offer certain advantages especially when it comes to interpretability. The decision boundary in SVMs is determined by support vectors, which are data points that lie closest to this boundary. This makes it relatively straightforward to understand which data points influence the decision-making process, and completely interpret the model's predictions. Moreover, for linear SVMs the coefficients of the weight vector, in the decision boundary equation ($w \times x + b = 0$), correspond to the features in the input data. The magnitude of each coefficient indicates the importance of that feature in determining the decision boundary. Decision trees, which are the building blocks of RFs and GB-Trees, are inherently interpretable. They split on features at various thresholds, allowing for a clear, hierarchical understanding of decision-making. Through this understanding and by calculating the impurity of each split based on a given feature, a final feature importance can be assessed based on the aggregated quality of splits that each feature outputs. The complete explainability makes SVMs and ensemble methods ideal candidates for clinical tools, especially at early stages of researching their utility in such uses.

An additional advantage of the interpretability lies on feature ranking and deriving feature importances directly from the model's training. In the HCM models presence of atrial

fibrillation or flutter (AF) and the SARCPLP variants were highlighted as key features. HCM often is associated with with AF, with studies indicating that approximately 20% to 25% of HCM patients also have AF (Falasconi et al., 2020). Presence of AF with HCM has been shown to have a higher incidence of stroke or heart failure (Maron et al., 1995, Patten et al., 2018), however this is not the case in this UKBB subset, as derived by the feature importances and the prevalence of stroke events in the HCM group. SARCPLP variants given their pathogenic or likely pathogenic significance to HCM are expected to rank high in feature importance, and were selected and included for this purpose.

Other notable features with high importance include sex, the presence of coronary or valve diseases, hypertension, instances of cardiac arrest, possession of SARVUS and SARCo variants, and BMI. HCM has been observed to be more prevalent in men, with males making up about 60% of the HCM cases (Butters et al., 2021, Siontis et al., 2019). This male dominance is puzzling and not yet well-understood, given that HCM is inherited in an autosomal dominant manner, which should lead to an equal distribution between males and females. One possible reason for this disparity might be a diagnostic bias, leading to HCM being less frequently recognized in women (Butters et al., 2021). In the entirety of the UKBB dataset, the prevalence of women is slightly higher (54.4%), so is in the Controls group used in the models (52%). In the HCM group the number of males was predominant with 66% of cases, which follows the pattern of male predominance of the disease, as described before. Based on UKBB's encodings of this binary feature (0: female, 1:male) a higher positive value in the SVM importance results of the sex variable indicates an association of being male with the HCM prediction label; which is in concordance to other studies.

The definition of HCM requires the absence of hypertension or coronary/valve disease in

order to explain the left ventricular wall enlargement. However, patients with HCM can also develop hypertension or valvular aortic stenosis, making it challenging to distinguishing the simultaneous presence of these conditions (Marian and Braunwald, 2017). Hypertension is also a known risk factor and has been proven to be causally related to HCM through Mendelian randomization studies (Harper et al., 2021). Therefore the presence of high importance in conditions like this is expected. So is an elevated systolic blood pressure, which can define the hypertension phenotype (Wang, 2019). Glucose metabolism has been studied extensively in relation to hypertrophy and heart failure, with animal models and clinical trials showing a relationship between an altered glucose metabolism and hypertrophy. Specifically, in the process of cardiac hypertrophy and abnormal restructuring, there's a metabolic shift from fatty acid oxidation to glucose use. This shift is linked to an excess of glycolysis in enlarged hearts, which can lead to hyperglycemia (Kolwicz and Tian, 2011, Tran and Wang, 2019). Murakami et al. (2004) findings indicate that HCM patients without clear signs of diabetes or hypertension may still have insulin resistance. This insulin resistance could be associated with HCM manifestations. Hence, seeing an elevated Hba1c feature showing a higher relative importance and contributing to predicting the HCM group aligns with what is previously known.

For DCM, the ESC guidelines recommend that all patients exhibiting heart failure symptoms undergo standard blood tests to check for associated conditions. These tests should include a complete blood count, assessments of kidney and liver functions, electrolyte levels, thyroid function, fasting glucose, and Hemoglobin A1C (Hba1c). In addition, there have been studies linking insulin resistance with DCM and as a result of ineffective glucose metabolism in the heart. It is well-known that diabetes can cause an impairment in the cardiac muscle,

leading to diabetic cardiomyopathy; however, insulin resistance is also present in DCM patients with no diabetes, which suggests a possible bi-directional underlying mechanism (Mak et al., 2021). As a result, including blood glucose as a predictor in the DCM models is beneficial and its high importance is justified.

The features that consistently influence DCM (based on the SVM-linear) include sex, the presence of coronary or valve diseases, occurrences of cardiac arrest or respiratory failure, hypertension, age, BMI, blood glucose levels, and the presence of DCMPLP variants. The sex variations observed in DCM follow the patterns previously mentioned for HCM. A number of epidemiological studies have shown that DCM is less prevalent in women compared to men (Fairweather et al., 2023, Halliday et al., 2018). Additionally, men exhibit lower LVEF, face poorer outcomes, and have a higher death rate compared to women (Argirò et al., 2022).

The mean age in the entire UKBB dataset is 57.9 ± 8.6 years, with the Controls group having a mean age of 54.9 ± 7.5 and the DCM group 59.9 ± 6.9. Halliday et al. (2018) investigated the different outcomes and severity of disease based on age, showing that there is an increased mortality in cases of more than 60 years of age (and the sex differences in these). However, there are no clinical indications or pathobiological pathways that could explain the difference in mean between these groups. A possible explanation would be participation bias stemming from the volunteer-based approach of the UKBB (Schoeler et al., 2023). The Controls group comprised of participants who all undertook further testing for MRI and CMR imaging. The almost 40,000 participants for which CMR imaging was analyzed have a mean age lower than the total UKBB mean and almost identical to the Controls group. Whereas the diseased groups include mostly individuals that did not undertake CMR imaging, and belong to the remaining 450,000 UKBB individuals. To ensure that age did not solely drive the

predictions, models removing the age variable were also assessed and did not show significant differences to the models presented here. Another way to mitigate this would be to correct for the age differences between the Controls and DCM/HCM groups and therefore remove the imaging participation bias apparent.

Regarding BMI, both the HCM and DCM groups have a mean index within the overweight range. The upper cutoff for a healthy range is 24.9 kg/m^2, making the Controls group with a BMI mean at 26.6 kg/m^2 also higher than the normal healthy range. There have been studies investigating the associations of obesity with cardiomyopathies, given the known link of obesity to worse cardiovascular outcomes (Zaromytidou and Savvatis, 2023); Harper et al. (2021) explored this with two-sample Mendelian randomization and identified a causal relationship of obesity with HCM, but not T2DM. Specifically, they showed a stronger association of BMI with sarcomere negative than sarcomere positive HCMs. Higher BMI in young women was linked to a greater likelihood of being diagnosed with cardiomyopathy later on, particularly DCM. This increased risk began even with a small increase in body weight while those with severe obesity faced a risk that was nearly five times greater (Robertson et al., 2020).

Similar to HCM, coronary artery disease, valve disease, cardiac arrest, respiratory failure, and hypertension are all comorbidities to DCM (esc). Additional risk factors, such as smoking and alcohol intake, when combined with a genetic architecture can increase susceptibility to developing a dilated phenotype. Thus, the association of DCM with the presence of P/LP or VUS variants and increased alcohol consumption, as indicated by feature importance, is justified.

Finally, increased left and right hand grip strength (HGS) measurements show a high association with the Controls group as well as more vigorous activity. It is evident that

patients with phenotypes that can lead to heart failure, as is DCM, strength of hand grip is impaired. There is a negative correlation between the odds for developing heart failure and HGS, indicating that lower values of HGS can be seen in patients with DCM (Li et al., 2023). Enhanced HGS correlated with CMR-derived indicators of cardiac structure and function that indicated reduced cardiac hypertrophy and remodeling (Beyer et al., 2018).

5.4.3 HCM models including LAVASET-derived CMR features perform better

Incorporating CMR features in the HCM vs Controls models increased the performance of the SVM models across all metrics, and resulted in a perfect precision for the ensemble methods. This increase is expected given the strong hypertrophic phenotype of HCM hearts in comparison to the Controls. This is also evident by current medical practice where a CMR image of the left ventricle and the measuring of its hypertrophy is the main diagnostic criterion for HCM (esc). The lower number in recall in association with the perfect prediction scores that the ensemble methods showed, indicate a high number of false negatives and seemingly no false positives. A potential explanation to this would be the large imbalance between the HCM and Controls classes, whereby the methods potentially make predictions by chance and the majority class is favored in these predictions.

This model set up had significant limitations in comparison to the pre-screening models. The necessity of including the RBH HCM cohort due to the low UKBB HCM number with CMR features led to an increased number of missing values in the clinical features. Therefore, most of the features used in the pre-screening models had to be excluded after harmonization of the two datasets; the main reason why the RBH cohort was also not used as a testing set for

the pre-screening models. To prevent inflating the testing scores due to the incorporation of selected features from the same testing cohort, the test set in these models was maintained identical to the test set in the LAVASET model. However, this was a necessity due to the restricted number of cases while the ideal set up would be to derive the important features via LAVASET, train a new model in a new dataset and test on separate data. As a future improvement, and as more CMR imaging data becomes available in the UKBB, and with the influx of more CMR imaging data in the UKBB, the aim would be to set up models using training data that has not been utilized previously and make predictions on the same test set used here.

5.5 Conclusion

This chapter demonstrated that traditional machine learning methods can differentiate between individuals with cardiomyopathy and those without, utilizing readily available clinical data and genomic rare variant information. In both binary classification models evaluated (one for each type of CM: HCM and DCM), the best-performing method was the SVM. Specifically, in the HCM vs Controls models, SVM-linear, is able to identify 3 out of 4 HCM cases correctly, while about 4 out of 5 HCM-predicted cases, are true HCM patients, indicating a relatively high precision.

For the DCM vs Controls models, AUROCs were tied for linear and RBF SVMs, suggesting that both types of kernels can be used for accurate classifications. In comparison to the HCM models, there is a relative decrease in performance of DCM vs Controls which is evident across all metrics evaluated in this Chapter. However, the models still identified correctly 72% of the

participants with DCM, while 75% of the instances where the model classified a participant as having DCM, that classification was true. Findings from both sets of models are results from robust predictions of 100 iterations, to accommodate for the nature of features such as the rare variant information incorporated. This methodology introduces an appropriate and novel approach to utilizing genomic rare variant information in an unbiased manner and creating robust learning.

Moreover, employing classical machine learning models such as SVMs allows for comprehensive interpretability and explainability of the models, resulting to important features being evaluated and ranked. Known CM comorbidities and/or risk factors such as hypertension and alcohol intake were ranked with relative high importance in the models, along with pathogenic or likely pathogenic variants for CM. When incorporating features from LAVASET-derived important CMR features, these ranked the highest in importance for distinguishing between HCM vs Controls models, and increased the overall model performance to almost perfect classification (AUROC=97.8%); which outperforms the best-performing models by over 20% on the AUROC metric.

The models described here can lay the groundwork for assessing the utility of ML in pre-screening and/or stratifying cardiomyopathy patients. There are still several shortcomings to address, most of which relate to the rarity and low prevalence of these phenotypes. Nevertheless, as more datatypes such as CMR imaging and WES become available for larger numbers of affected individuals, the predictive positive value of such models will be able to increase and potentially become useful in the clinic.

Chapter 6

Deep Learning Models for Evaluating PRS in Cardiomyopathies

6.1 Introduction

HCM and DCM have been traditionally considered monogenic diseases, with a Mendelian pattern of inheritance. About a third of HCM cases can be attributed to a causal rare variant and recent studies have shown that common variants can contribute significantly to risk of disease (SNP heritability (h2) 0.29) (Tadros et al., 2021). In DCM, a GWAS analyzing left ventricular phenotypes from cardiac MRI data unveiled 45 new loci of common variants, which cause cardiac structure alterations seen in DCM. A polygenic risk score (PRS) based on these MRI-derived measurements demonstrated a strong correlation with the occurrence of DCM in the general population (Nair and Hinton, Pirruccello et al., 2020, Tadros et al., 2023, Zheng et al., 2023a).

A PRS, as described in detail in Introduction section 1.2.2, summarizes the cumulative risk that can be attributed to common variants. In the initial stages of developing PRS, scores were computed using variants that were statistically significant according to GWAS. Despite this, the risk of over-fitting remains, especially when dealing with small sample sizes and/or a large number of variants. To address this, some methods have incorporated regularization techniques such as the Best Linear Unbiased Predictors (BLUP) and genomic BLUP (gBLUP) (Clark and Werf, 2013). These pipelines have their own limitations, including the potential to inflate the effect sizes of correlated genetic variants.

To mitigate the issue of overestimation, techniques that take into account the linkage disequilibrium (LD) have been recently introduced. One approach selects representative SNPs within LD blocks, using pruning or clumping (known as P+T), while another relies on LD panels, serving as the foundation for algorithms like LDpred (Vilhjálmsson et al., 2015), Lassosum (Mak et al., 2017), and PRS-CS (Ge et al., 2019). Incorporating LD information can

be computationally expensive and still fall short in capturing the complexities of of non-linear effects and interactions between the genetic variants.

Using machine learning models to predict PRS can capture non-linear SNP interactions with fewer genetic architecture assumptions. Employing deep learning methods for calculating and evaluating PRS has been shown in more prevalent diseases such as Alzheimer's disease (Zhou et al., 2023), breast (Badré et al., 2020) and ovarian cancer (Dareng et al., 2022).

Here, we present a different method for calculating PRS by training a feed-forward neural network and generating the PRS from its prediction probabilities on new, unseen data. This ML-based approach shows that PRS in HCM and DCM can be clinically relevant, in concordance to results presented recently by (Zheng et al., 2023b), and is completely data-driven without the need of prior knowledge like LD panels or priors.

6.2 Methods

PRS-CS is one of the most recent methods developed for generating PRS (Ge et al., 2019). Its novelty comes from incorporating continuous shrinkage priors when calculating the SNP effect sizes, which enables multivariate modeling of local LD patterns. These priors are incorporated via a Bayesian regression method. Here, PRS-CS is used as a traditional PRS calculation method for comparison to the DL method, given its proven utility in cardiomyopathies, as presented recently (Zheng et al., 2023a,b). PRS-CS scores used here are generated by Zheng et al. (2023a).

6.2.1 SNP Filtering

SNP filtering was executed based on the *p*-value thresholds generated by previously performed Genome-Wide Association Studies (GWAS) and meta-analyses, including Multitrait Analysis of GWAS (MTAG). The HCM GWAS was performed on HCM cases (n=2,431) and controls (n=40,283) from the Hypertrophic Cardiomyopathy Registry (HCMR). The study included unrelated patients who were diagnosed with HCM based on the presence of unexplained thickening of the left ventricular (LV) wall, defined as a LV wall thickness (LVWT) greater than 15mm, or greater than 13mm with either a family history of HCM or a known genetic variant that is associated with the disease. The HCM cases underwent gene panel sequencing based on clinical indications (Tadros et al., 2023).

For DCM, summary statistics were used from a large meta-analysis effort that combined 16 studies participating in the Heart Failure Molecular Epidemiology for Therapeutic Targets (HERMES) Consortium. Total number of DCM cases was 4,255 with 1,199,156 controls (Lumbers et al., 2021). Six of the 16 studies identified DCM cases using string criteria, involving cardiac imaging, specific diagnostic codes from ICD (International Classification of Diseases), or DCM-specific clinical diagnosis. Under these criteria there were 6,001 cases and 449,382 controls. The remaining 10 studies identified cases based on the presence of left of left ventricular systolic dysfunction (LVSD) without any underlying secondary causes. This less stringent category consisted of 9,298 cases and 1,157,145 controls (Lumbers et al., 2021).

Based on the summary statistics from both the HCM GWAS and the DCM meta-analysis, different sets of SNPs were evaluated as inputs to the neural network models. These sets were created by taking SNPs in different levels of *p*-value significance. Seven different *p*-value thresholds were assessed from 1×10^{-2} to 1×10^{-8} with each subsequent step representing a

difference of one additional order of magnitude.

6.2.2 Neural Network Architecture

For both the HCM and DCM classification tasks, two separate feed-forward neural networks were trained using `Keras` and a number of parameters were tested on the validation set to find the optimal architecture. The models always begin with an input layer consisting of as many neurons as there are features in the training data. This varied based on the 7 different *p*-value thresholds. This is followed by one or more hidden layers with 128 neurons in the first hidden layer. Subsequent hidden layers, if present (steps of 1 to 5 hidden layers were evaluated), have a diminishing number of neurons (neuron count in subsequent layers was determined by dividing 128 by a range of 1 to 4). Each hidden layer used the ReLU (Rectified Linear Unit) activation function. Presence of batch normalization and dropout layers was also evaluated to assess whether they prevent over-fitting. The Batch Normalization was set at momentum of 0.9, and the Dropout rate was evaluated at 0.1 to 0.9. The network concludes with a single-neuron output layer with a logistic activation function, making it suitable for binary classification tasks. The models were compiled using the Adam optimizer with an initial learning rate of 1×10^{-3} and used binary cross-entropy as the loss function. Performance is evaluated using multiple metrics, including accuracy, Area Under the Receiver Operating Characteristic curve (AUROC), Area Under the Precision-Recall curve (AUC-PR), Precision, Recall, and a F1 Score. Training is conducted with early stopping based on validation F1-score, with a patience parameter set to 10 epochs. The training is performed in batch sizes of 500 and is run for a maximum of 100 epochs.

The architecture was constructed with the `Sequential()` function of the `Keras` sub-

package in `TensorFlow`. Model specifications and ranges for parameter hypertuning were informed by other DL-PRS methods in literature (Badré et al., 2020, Dareng et al., 2022, Zhou et al., 2023).

To mitigate the risk of over-fitting arising from linkage disequilibrium (LD), the method of weight correlation descent (WCD) introduced by Kim et al. (2023) was used. WCD serves as a regularization technique by introducing an additional term to the model's loss function, as indicated in Equation 6.1.

$$\tilde{J}(\theta; X, y) = J(\theta; X, y) + \alpha g(w) \tag{6.1}$$

Here, $\tilde{J}(\theta; X, y)$ represents the regularized loss function, $J(\theta; X, y)$ is the original loss function prior to applying WCD, $g(w)$ measures the correlation between the weights of each node in the NN, and α is a coefficient to fine-tune the impact of WCD. A higher α means more regularization. The weight correlation for each layer of the NN is computed using Equation 6.2:

$$\rho(w_l) = \frac{1}{N_l(N_l - 1)} \sum_{\substack{i,j=1 \\ i \neq j}}^{N_l} \frac{|w_{li}^T w_{lj}|}{\|w_{li}\|_2 \|w_{lj}\|_2} \tag{6.2}$$

Here, $\rho(w_l)$ measures the correlation between the weights of nodes within the same layer, N_l is the number of nodes in layer l, and w_{li} and w_{lj} are the weights of the i-th and j-th nodes. The denominator normalizes the term, ensuring that the correlation is not affected by the magnitude of the weights. $g(w_l)$, which is the weight correlation loss for each layer, is

determined as shown in Equation 6.3:

$$g(w_l) = -(N_l - 1)N_{l-1}\ln(1 - \rho(w_l)) - N_{l-1}\ln\left(1 + (N_l - 1)\rho(w_l)\right) \tag{6.3}$$

This term is used to penalize high correlations within the weights of the same layer. Finally, $g(w)$ is calculated by summing the weight correlation loss across all layers, i.e., $g(w) = \sum_l g(w_l)$.

6.2.3 Dataset and Training/Validation/Test Splits

Genotyping data of UK Biobank (UKBB) participants were used for both the HCM and DCM neural network training. Unrelated (excluding 3rd degree or closer), White British participants were only included. Cases of HCM were picked through both self-reported medical histories, hospital admissions and death records, and through cardiac magnetic resonance (CMR) imaging showing a left ventricular wall thickness exceeding 15mm. Unrelated White British HCM cases from the Royal Brompton & Harefield (RBH) Hospitals Cardiovascular Research Biobank were also included to increase the number of cases for training and testing. To define the DCM cases in the UKBB diagnostic codes from ICD (International Classification of Diseases) and linked primary care records were used.

Both cohorts were filtered by standard quality control (QC) criteria and the finalized dataset was refined including only variants that passed genotyping QC (MAF 1%, genotyping rate 0.99, HWE 1×10^{-6}).

The total number of individuals included for the DCM models is 4,290 out of which 1,052 DCM cases and 3,238 Controls. For the HCM models the total number of individuals adding both the UKBB and RBH cohorts sums up to 10,624. HCM cases are 790 and Controls

9,834. Controls in both sets of models were selected via clinical criteria and included only individuals with CMR imaging data (for the downstream PRS evaluation analysis). Clinical criteria included no prior cardiovascular or known common metabolic diseases (i.e. diabetes), as well as no hypertension phenotypes. The sample sizes presented here are used for training the NN models and learning the PRS weightings and are independent of the GWAS and GWAMA discovery cohorts that were described above in section SNP filtering 6.2.1.

To separate the target cohort for PRS tuning and PRS evaluation, each dataset was split in training and validation (PRS tuning), and test sets (PRS evaluation). Training (80%) is the subset used for the training of the various *p*-value threshold inputs and to decide the optimal parameters for each model. The validation set (10%) was used to cross-validate the parameter tuning and pick the best-performing input. Finally, the test set (10%) was used to calculate the PRS for the remaining and unseen individuals. Figure 6.1 shows the exact numbers on each set.

p-value	HCMR GWAS	UKBB-HCM	RBH-HCM	UKBB\RBH Common	**HERMES**	UKBB-DCM
1×10^{-2}	73,294	51,786	66,247	47,465	140,364	92,900
1×10^{-3}	12,447	7,491	11,316	6,945	24,411	16,175
1×10^{-4}	4,885	2,170	4,378	2,014	6,218	3850
1×10^{-5}	3,502	1,198	3,110	1,121	2,508	1,484
1×10^{-6}	2,970	803	2,638	752	1,254	842
1×10^{-7}	2,293	558	2,140	531	918	670
1×10^{-8}	448	403	423	388	615	460

Table 6.1: SNP variants as input for the HCM and DCM neural network models. HCMR GWAS column shows the number of variants per significance threshold as taken from the summary statistics of the base data (GWAS). UKBB and RBH columns shows the number of variants present in the UKBB and RBH respectively. UKBB+RBH Common are the final set of variants used in the dataset for the neural network pipeline. HERMES column shows the number of variants per significance threshold as taken from the summary statistics of the base data (HERMES meta-analysis). UKBB-DCM is the final set of variants used in the dataset for the neural network pipeline.

6.2.4 Statistical Analysis

The non-parametric Mann-Whitney U-test was employed for all statistical tests performed in this chapter. A non-parametric test was chosen given the various group sizes compared when looking at distributions of PRS within quantiles or across higher and lower value ranges.

In the DCM section we calculated the odds ratio (OR) for DCM. OR is a measure of association between an event (in this case the PRS) and an outcome (usually a disease or condition, here DCM). It is a statistic that quantifies how the odds of the outcome changes with the exposure compared to without the exposure.

The odds ratio is calculated as follows:

$$\text{Odds Ratio (OR)} = \frac{\text{Odds of outcome with event}}{\text{Odds of outcome without event}} \tag{6.4}$$

In a 2×2 contingency table format, where:

- a is the number of cases with high PRS and DCM,
- b is the number of cases with high PRS without DCM,
- c is the number of cases with low PRS but DCM,
- d is the number of cases with low PRS and without DCM,

The odds ratio can be calculated as:

$$\text{Odds Ratio (OR)} = \frac{a \times d}{b \times c} \tag{6.5}$$

- OR $= 1$: No association between exposure and outcome.

- OR > 1: The exposure is associated with higher odds of the outcome.
- OR < 1: The exposure is associated with lower odds of the outcome.

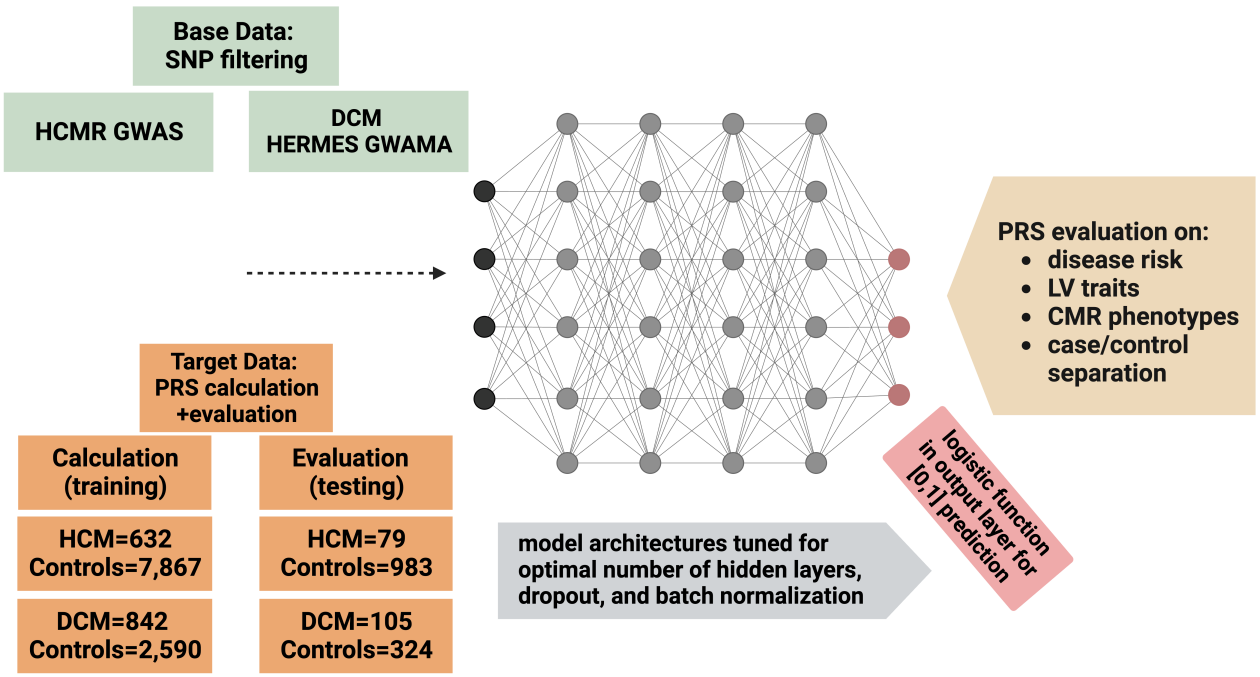


Figure 6.1: High-level overview of the chapter, including the dataset splits, model architecture, and PRS evaluation methods. Neural network depicted is not indicative of the exact architecture.

6.3 Results

Table 6.2 and Table 6.3 present the results from the models with different sets of SNPs as input, based on the *p*-values of different levels of magnitude (1×10^{-2} to 1×10^{-8}). Results are shown for the validation set (10% of total samples), from which the model with the highest F1-score is selected. For HCM this is the model with *p*-value input of 1×10^{-5} and for DCM

HCM Validation Set Metrics						
Model	Accuracy	Precision	Recall	ROC-AUC	AUC-PR	F1-Score
1×10^{-3}	0.91	0.40	0.19	0.62	0.13	0.26
1×10^{-4}	0.94	0.79	0.37	0.75	0.41	0.50
1×10^{-5}	0.95	0.79	0.51	0.82	0.59	0.63
1×10^{-6}	0.94	0.72	0.43	0.80	0.44	0.55
1×10^{-7}	0.95	0.93	0.42	0.80	0.44	0.57
1×10^{-8}	0.95	0.90	0.42	0.79	0.48	0.57

Table 6.2: HCM validation metrics across the different levels of SNP filtering.

DCM Validation Set Metrics						
Model	Accuracy	Precision	Recall	ROC-AUC	AUC-PR	F1-Score
1×10^{-3}	0.51	0.31	0.59	0.55	0.32	0.40
1×10^{-4}	0.54	0.32	0.59	0.53	0.31	0.42
1×10^{-5}	0.52	0.31	0.48	0.54	0.29	0.39
1×10^{-6}	0.49	0.27	0.47	0.51	0.29	0.35
1×10^{-7}	0.53	0.32	0.58	0.56	0.32	0.41
1×10^{-8}	0.48	0.28	0.54	0.54	0.31	0.37

Table 6.3: DCM validation metrics across the different levels of SNP filtering.

the model with 1×10^{-4}. Prediction scores are then calculated based on these inputs for the remaining 10% of the test set in both cases.

6.3.1 PRS can separate between HCM cases and Controls

Figure 6.2 A and B show the distribution of PRS when looking at the HCM and HC groups separately. Figure 6.2A shows the raw scores as direct output from the neural network, scores that as expected are in the range between 0 and 1. There is an evident bimodal distribution of scores when looking at the HCM samples (n=69) which falls in line with the recall values of the test set (0.48). The test set precision climbs to 0.69 which also explains the pattern of distribution shown for the HC samples (n=994). To transform these scores to a format and distribution closer to the more classical PRS method outputs, Figure 6.2B presents the log-transformed scores. In both graphs it is apparent that the median PRS values of HCM

(shown by the dashed blue lines) are further right to the HC median, indicating that HCM PRS is higher than for HC (Mann-Whitney p-value $= 1.32 \times 10^{-19}$. These results are in line with other recent findings using GWAS meta-analysis data (Zheng et al., 2023b). Finally, Figure 6.2C shows the distribution of only HCM samples based on their genotype. A sarcomere positive (SARC+) genotype defines carriers with a rare variant in any if the 8 sarcomeric genes, proven to have high association with the disease. Sarcomere negative (SARC-) are non-carriers of such variant. From the graph it is evident the SARC+ carriers have a lower PRS to the non-carriers, however in the test set there were only 9 carriers, therefore there is not enough statistical power to discern whether this signal is true.

6.3.2 HCM PRS groups show differences in left ventricular phenotypes

HCM is most commonly diagnosed by looking at the maximum wall thickness of the left ventricle. This predictor along with the left ventricular mass and end-systolic volume are strong indicators of dysfunction related to HCM. Figure 6.3 and Figure 6.4 show the distributions of values for these 3 predictors when looking at samples with higher or lower PRS. Specifically, Figure 6.3 shows a strong pattern arising when taking the 27 individuals with PRS equal to or higher than 80% and comparing them to the 953 individuals with PRS equal to or lower than 20%. A similar, but less pronounced, pattern arises when looking at the top and bottom 20% quantiles shown in Figure 6.4. The pattern in this second figure is less pronounced but still follows the biological and anatomical phenotype we would expect. That is, individuals with higher PRS exhibit am increased maximum wall thickness and mass in the left ventricle. This while their end-systolic volume is lower and compromised as expected to individuals with a lower PRS, who we would expect to have a closer to normal (i.e. higher

end-systolic volume) left ventricular function.

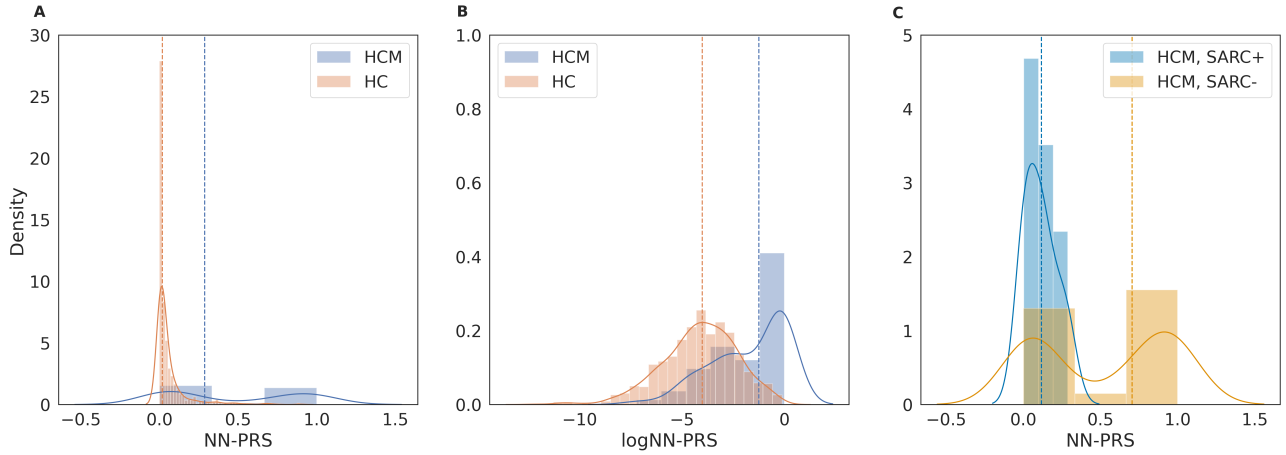


Figure 6.2: Distributions of PRS scores generated by the neural network model for HCM. X-axis values in A are the prediction probabilities generated by the output layer (logistic function) of the NN. X-axis values in B are the log-transformed values from B. Dotted vertical lines represent the median of each distribution with respective colors (HCM: n=69 - blue, HC: n=994 - orange). Differences between the two distributions in both A and B are found significant (Mann-Whitney *p*-value $= 1.32 \times 10^{-19}$). Panel C shows the PRS distribution for just the HCM subset. Sarcomere positive (SARC+, n=9, blue) cases are of those who are carriers of at least one rare variant in any of the 8 sarcomeric genes associated with HCM. Sarcomere negative (SARC-, n=60, yellow) are non-carriers, who still exhibit the phenotype (Mann-Whitney *p*-value $= 1.00 \times 10^{-2}$). NEED TO ADD THE MEDIAN VALUES

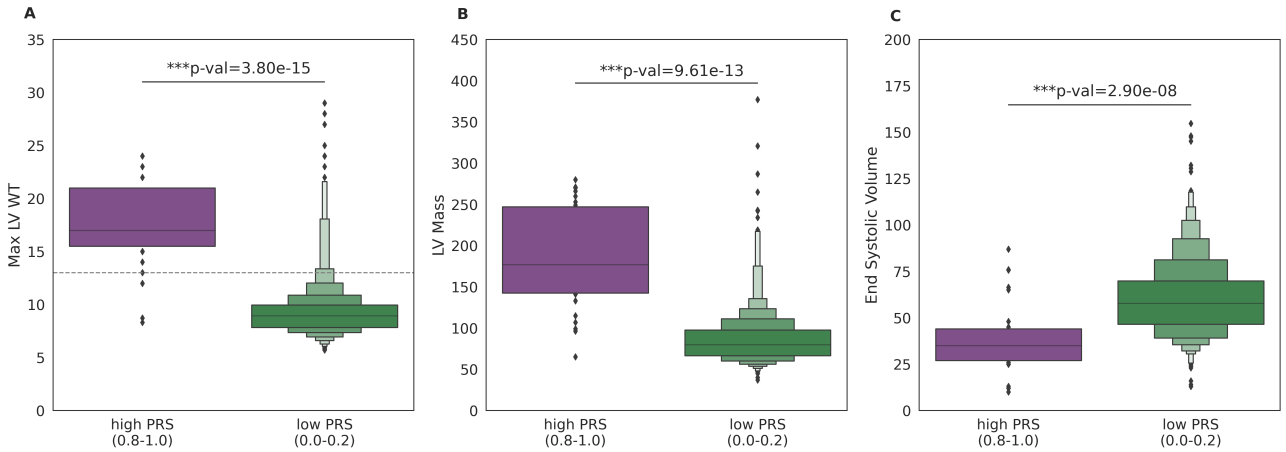


Figure 6.3: CMR imaging outcomes stratified by PRS groups. Three HCM-related outcomes are presented: maximum left ventricular wall thickness (A), left ventricular mass (B), and end systolic volume (C). The dashed gray line in panel A indicates the clinical definition of hypertrophy in millimeters (>13mm). Purple boxes include the number of samples (n=27) with PRS equal to or higher than 80%, green boxes include samples (n=953) with a PRS that is equal to or lower than 20%. All 3 CMR outcomes show significance ((Mann-Whitney *p*-value $= 3.80 \times 10^{-15}$, 3.61×10^{-13}, 2.90×10^{-8} for A, B, C respectively.)

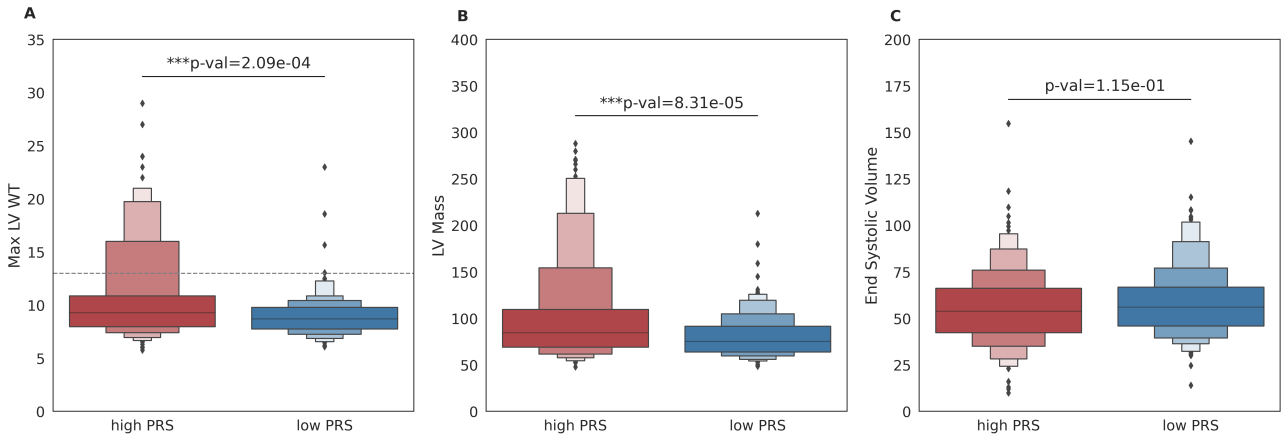


Figure 6.4: CMR imaging outcomes stratified by log-transformed PRS groups. Three HCM-related outcomes are presented: maximum left ventricular wall thickness (A), left ventricular mass (B), and end systolic volume (C). The high and low PRS groups are defined by the highest (>0.8) and lowest (<0.2) quantiles respectively. The dashed gray line in part A indicates the clinical definition of hypertrophy in millimeters (>13mm).

6.3.3 PRS can separate between DCM cases and Controls

Figure 6.5 show distributions of PRS generated by the Bayesian approach PRS-CS and the feedforward neural network. Specifically, Figure 6.5A presents the DCM and HC PRS distributions on the exact same set of samples as the test set. While Figure 6.5B shows the raw scores as direct output from the neural network (NN-PRS), again separated by disease status (HCM n=106, HC n=1038). Figure 6.5C shows the log-transformed prediction probability values (logNN-PRS), consistent to how it was done with the HCM models. In all three graphs it is evident that the mean value of the HCM distribution is higher than the HC mean, however we do see that the neural network PRS separates the two groups more than the PRS-CS scores (Mann-Whitney Part A p-value $= 4.07 \times 10^{-2}$, Part B p-value $= 2.22 \times 10^{-4}$). In Figure 6.5A the means for HCM and HC are -4.26×10^{-7} and -4.34×10^{-7} respectively, wheres in Figure 6.5B the HCM mean is 0.58 and for HC 0.49. For the log-transformed values we still see a similar order of difference between the two distributions with means at -0.67 for HCM and -0.94 for HC. Better separation of the DCM diseased group and the HC individuals using the

NN-PRS is also supported by looking at the R^2 score. Results from a linear regression model using the two sets of scores (PRS-CS and NN-PRS) and setting as y target the disease status indicate that the PRS-CS R^2=4.14×10^{-4} whereas NN-PRS R^2=4.22×10^{-3}.

When looking at the PRS distributions between carriers of a pathogenic/likely pathogenic DCM variant (n=234) and non-carriers (negative genotype, n=830), the median PRS score in the carriers is lower (0.47 vs 0.50 in negative), however this difference is not supported by statistical significance (Mann-Whitney U-test, p-value=0.23).

6.3.4 DCM PRS groups show differences in imaging and clinical features

Data from CMR imaging derived phenotypes for 84 traits were evaluated between the groups of high and low PRS. Figure 6.6 presents the results of statistical testing with Mann-Whitney U-test across the 84 traits, with 6 traits showing some significance between the two groups. Specifically, 4 traits relate to circumferential strain (Ecc) in 4 different segments of the left ventricle ($AHA1, 4, 9, 10$), another one relates to the radial strain (Err_AHA_1) and the final one to longitudinal strain (Ell_4). From the violin plots we can discern that there is a slight decrease in circumferential, radial, and longitudinal strain in the group with high PRS. Circumferential and longitudinal strains (Ecc_AHA_n, Ell_4) are given by a negative value, therefore the high PRS group (light purple) shows a reduction in absolute Ecc and Ell value, as shown by its distribution placed higher on the y-axis (lower negative value) than the low PRS group (light green). For the radial strain trait (Err_AHA_1), which is evaluated in positive digits, the decreased pattern of distribution for the high PRS group is more apparent. The rest of the 84 traits assessed did not show any statistical significance after multiple testing correction and were not explored further.

In Figure 6.7, the top graph shows the odds ratio for DCM across different PRS quantiles, when compared to the median quantile (OR=1, 40%-60%). Error bars were generated by bootstrapping samples between different quantile ranges, so as to calculate the confidence intervals for the specific range. The number of samples to bootstrap was set to 30 and the error bars in the top graph represent 95% confidence intervals. There is an evident upwards trend to the odds ratio as the PRS qunatile range increases, with the highest values found in the 90-95% quantile range (OR=2.50, 95% CI: 0.36-0.65). The difference between this range to the median is also found to be statistically significant (Mann-Whitney *p*-value = 3.14×10^{-2}).

The two bottom graphs of Figure 6.7 show the cumulative count of DCM diagnoses for each year across 3 different PRS groups, high PRS, median PRS, and low PRS, which can also translate to high, medial, and low risk for disease. Cumulative count is defined by the total number of individuals given a diagnosis of DCM at a specific year. In the first graph presented (left, test set) we see results for the test set DCM cases for which we have date of diagnosis data (n=23). Given the low number of samples (low risk PRS group only contains 3 cases), we also present results by including PRS generated for the validation set as well (n=125 total). In both graphs there is a discernible difference at onset age of disease between the high (n=52, including validation samples) and low (n=35, including validation samples) risk PRS groups, however given the sample sizes we can not but just make assumptions of the trends.

Figure 6.8 shows a comparison of the NN-PRS and PRC-CS scores when looking at DCM odds ratio (top graph) and cumulative count for diagnosis in the highest risk groups (80% quantile, bottom graph) for both scoring methods. The highest OR for PRS-CS is found in the 99% quantile while for other high risk quantiles (80-90, 90-95, 95-99) the PRS-CS OR is lower than the NN-PRS OR and in 2 instances on the median line of OR=1, indicating no association

with higher disease risk. The onset age of disease curves between the PRS-CS and NN-PRS show some differences, however these were not statistically significant (Mann-Whitney p-value = 0.37).

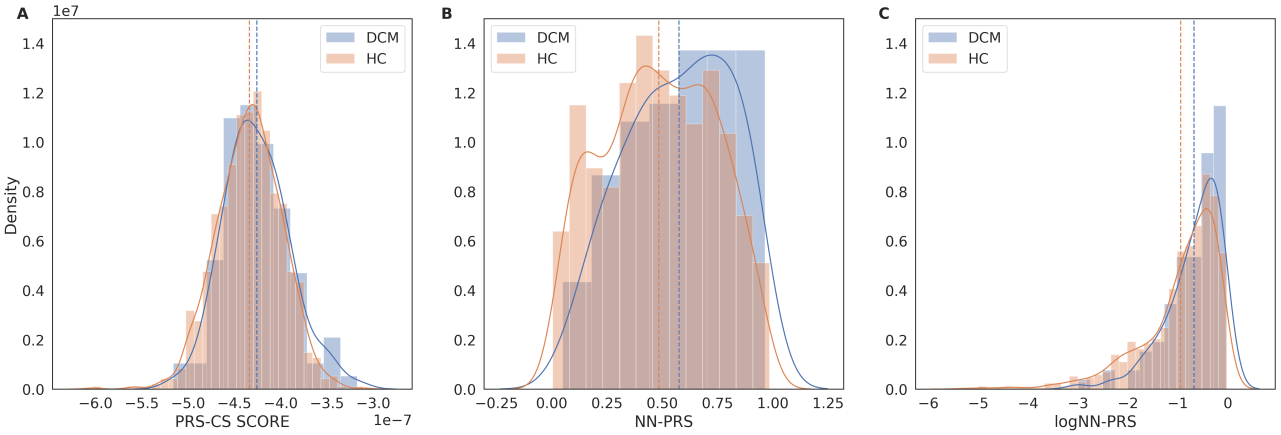


Figure 6.5: Distributions of PRS generated by PRS-CS (A), Neural-Network (B) and Neural Network log-transformed (C). X-axis values in A are generated by the PRS-CS method by calculating the SNP effect sizes and then converted in `plink` for individual sample scores. X-axis values in B are the prediction probabilities generated by the output layer (logistic function) of the NN. X-axis values in C are the log-transformed values from B. Dotted vertical lines represent the mean of each distribution with respective colors (DCM: n=106 - blue, HC: n=1038 - orange).

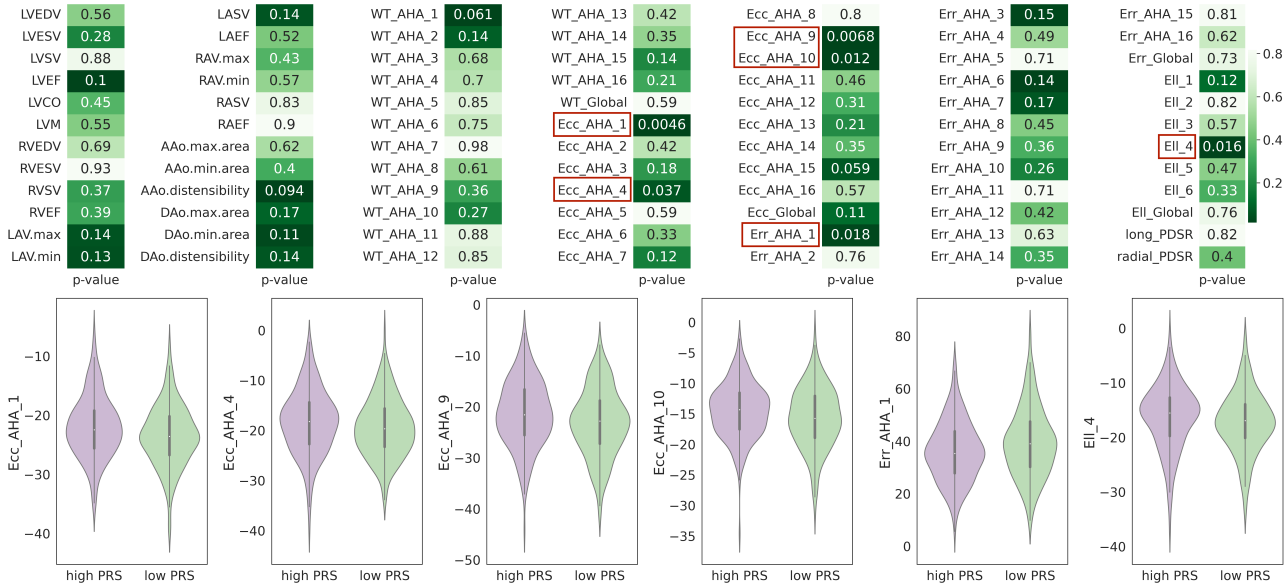


Figure 6.6: Evaluating differences between high and low PRS groups across 84 CMR imaging phenotypes. Top graph shows a heatmap of the Mann-Whitney *p*-values when comparing the logNN-PRS of the high PRS group (quantile > 0.8) to the low PRS group (quantile <0.2), adjusted for multiple correction with Hommel. Colormap shows lighter green values as higher (less significant) and darker green values as lower (more significant). Red boxes show the 6 phenotypes with statistical significance, p-value < 0.05. Violin plots below present the distributions of PRS for the 6 statistically significant CMR phenotypes, with the high PRS group shown in light purple and low PRS group in light green. Median values for each distribution are represented by the white dots in the middle of the plots. Abbreviations and explanations of the 84 CMR-derived imaging phenotypes are shown in Table 3.1.

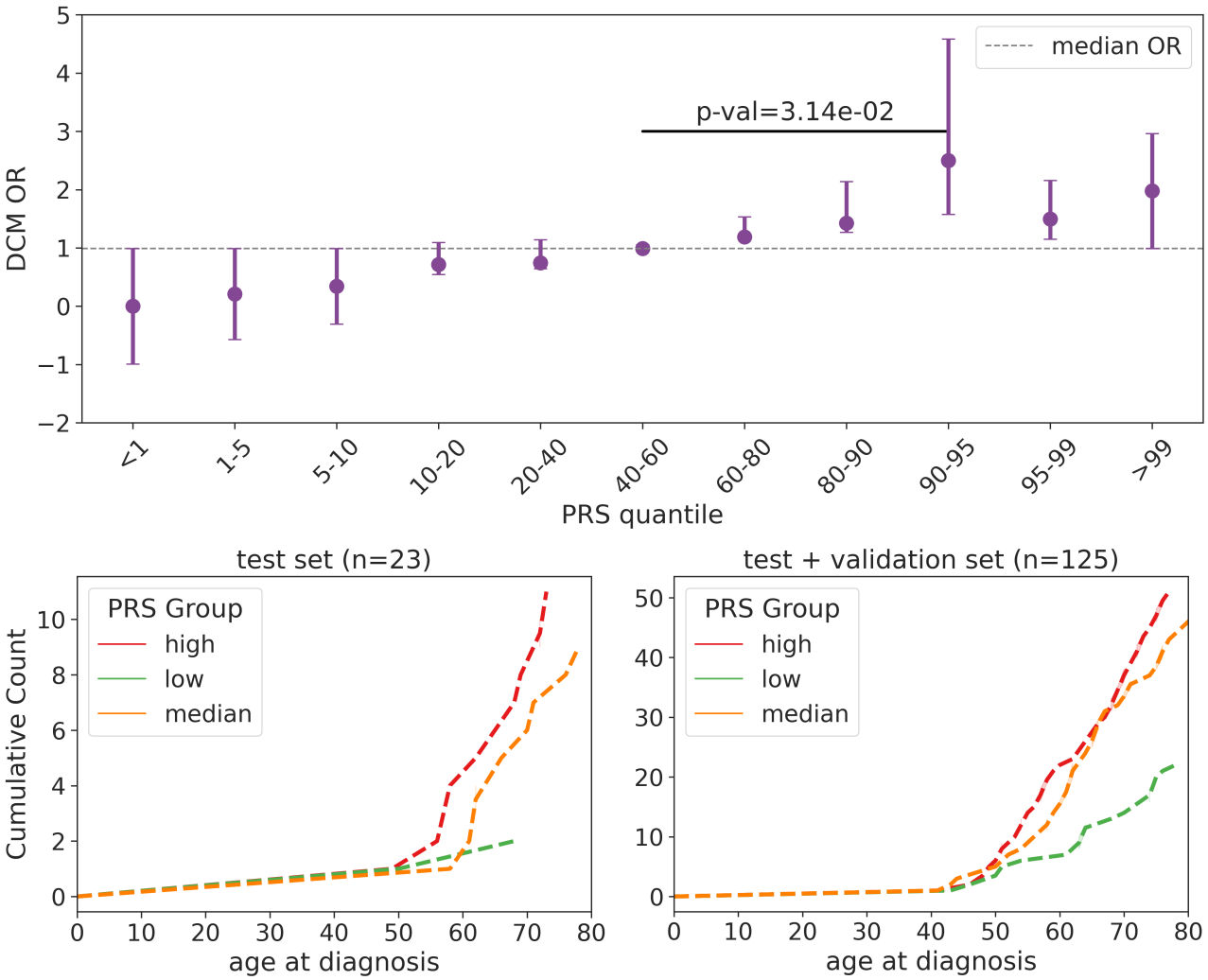


Figure 6.7: Top graph: Odds ratio for DCM across different quantile ranges, compared to the median range (40-60%. Bottom graphs: Cumulative count of DCM diagnoses for each year across 3 PRS groups (high (more than 80%), median (40-60%, and low (less than 20%). Left graph shows results from the test set (n=23) and right graph shows results from both the test and validation set (n=125).

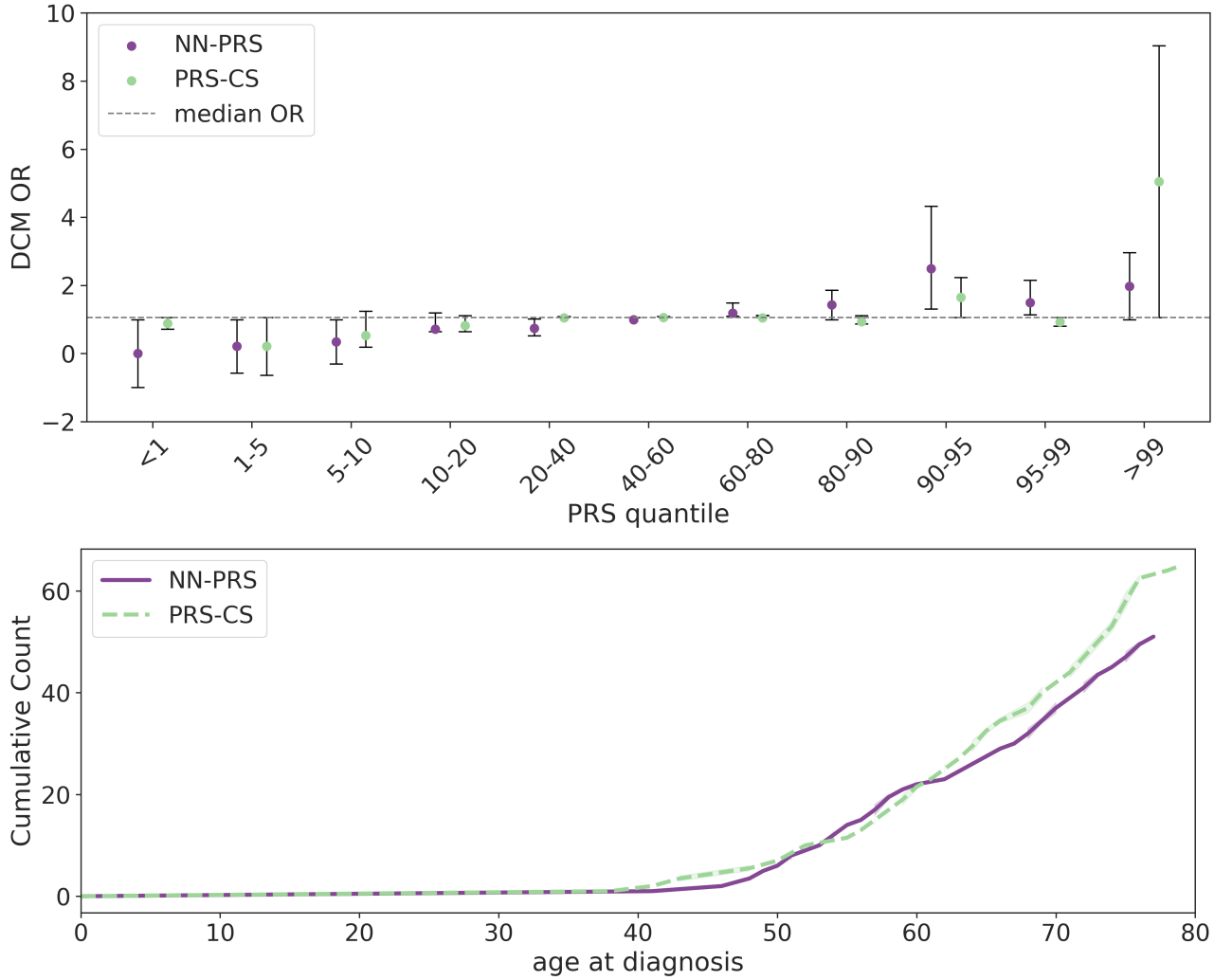


Figure 6.8: Comparison of the NN-PRS (purple) and PRS-CS (light green). Top graph shows the odds ratio for DCM across different quantile ranges for both PRS methods. Bottom graphs shows the cumulative count of DCM diagnoses for each year of the high PRS groups (more than 80%) for both methods.

6.4 Discussion and Future Work

Polygenic Risk Scores have recently become relevant in HCM and DCM, after studies have shown that part of the missing heritability in these –so far described– Mendelian diseases, is attributed to common variants (Pirruccello et al., 2020, Tadros et al., 2021). As the number of GWAS or cohort-based studies increases we will be able to conduct new and more effective genome-wide meta-analyses to identify more loci associated to cardiomyopathies.

A number of studies have shown that use of deep learning models in non-rare diseases such as cancer or Alzheimer's, evaluate the polygenic architecture more efficiently than classical statistical methods (PRS-CS) (Badré et al., 2020, Dareng et al., 2022). In this chapter, this use of deep learning, is explored as a new alternative method for calculating PRS in rare diseases, like HCM and DCM, while also evaluating the potential clinical utility of the PRS, given that this is still in early stages (Zheng et al., 2023b).

The deep learning model architecture described above, has the ability to generate PRS without the use of Linkage Disequilibrium (LD) reference panels. These panels are computationally expensive to generate and implement, and by themselves do not always capture the non-linear interactions. PRS-CS was developed to mitigate this loss of non-linear effects in other previous models (LDpred etc), however it still requires an LD reference panel, and the shrinkage priors that allow for multivariate local LD assessment, are still very dependent on these external LD panels. In contrast, the feed-forward neural network is able to generate equivalent, if not better results, without the need for external cohorts during training, and the additional step in QC that involved pruning or clumping SNPs found in LD. This pure data-in-hand-driven approach has the disadvantage, however, that while training we use a big part of our samples, for which we cannot generate PRS. A way to improve this in the future is

by creating a robust pipeline where samples across the three sets (training, validation, set) are bootstrapped. This will allow for samples to be switched around and used in every part of the pipeline, thus generating PRS for every individual, while also accounting for robustness with high iterations of sampling and model training (Xu and Goodacre, 2018).

In the HCM models, PRS is able to separate between HCM cases and Healthy Controls (Controls) with an ROCAUC of 0.82 in both the validation and test set for the best performing input (SNP filtering *p*-value = 1×10^{-5}, n SNPs). It was shown that the median PRS value for HCM individuals is higher than the median PRS value in Controls, and the distributions of raw prediction probability scores are in line with the precision and recall metrics. The feed-forward neural network model generated a bimodal distribution of scores for the HCM individuals, with peaks around 0 and 1 (based on whether the cases were classified correctly or not). To ensure that this was not a result of cohort artifact, given that about half of HCM cases were included from a different cohort (RBH) to increase sample size, we looked at the distribution of PRS in RBH and UKBB cases and did not find any significant differences.

To evaluate the clinical utility of PRS we look at known predictors of HCM, derived from CMR imaging data. Specifically, HCM is most commonly diagnosed by looking at the maximum wall thickness of the left ventricle. This trait along with the left ventricular mass and end-systolic volume are strong indicators of dysfunction related to HCM (esc). The results described here are in line with the biological and anatomical interpretation expected, where the group with higher risk of developing HCM based on PRS has a higher value of maximum wall thickness and mass of the left ventricle (Rowin et al., 2020). While the end-systolic volume, although not significant shows relatively lower values in the higher risk group, in line with the biological explanation of an increased wall thickness leading to lower volumes in end

of systole (Haland et al., 2017).

A major improvement on the HCM models would be to include summary statistics from a recent meta-analysis study that incorporated the HCMR study which was used here, along with another 6 studies to increase the sample power. Specifically, a total of 5,900 unrelated HCM cases and 68,359 controls of European ancestry were included in this GWAS (Zheng et al., 2023b), increasing the number of cases more than 2x from the HCMR study and therefore potentially including more significant and relevant loci in the summary statistics. This alteration would change the SNP variables used as input to the neural network models and potentially increase the PRS performance, given the more complete genetic information added from the additional significant SNPs.

To evaluate whether the NN-PRS performed better than the PRS generated by PRS-CS, we compared the distributions of scores between DCM and HC groups. In both methods, it is evident that the mean value of the DCM distribution is higher than the HC mean as we would expect, however we do see that the neural network PRS separates the two groups more than the PRS-CS scores. This is assessed by the difference of the means between the two distributions which in the case of NN-PRS is larger than in PRS-CS. When looking at the DCM odds ratios for the different PRS quantiles of NN-PRS and PRS-CS, the differences are not as pronounced. NN-PRS odds ratios show an increasing pattern from the lowest to the 90-95% quantile range, where the maximum value is found, and a decrease in the next two highest ranges. On this 90-95% quantile range it is also where it outperforms the PRS-CS, which however shows a highest value on the 99% quantile. Most importantly, PRS-CS OR fails to show an increasing pattern of DCM risk based on the PRS quantile, with high risk ranges such as 80-90 and 95-99 showing an OR=median=1, indicating no association with higher disease

risk.

An additional observation in the DCM context which differs from what we see in HCM, relates to the pattern of PRS distribution of the NN-PRS, i.e. the raw output of prediction probability values from the final layer (with a logistic activation) of the neural network. In DCM we see that the raw predictions follow a normal distribution (similar to the PRS-CS output) and they do not need to be log-transformed to resemble the more 'classical' PRS distributions. This can be explained when looking at the test set metrics of both models, and identifying that the precision scores are much lower in the DCM models than the HCM, while recall scores show some difference. This translates to the model assigning prediction probabilities that are closer to the mean of a normal distribution rather than the tails.

Given that PRS is able to separate DCM cases from Controls, the next step would be to assess its potential clinical utility on distinguishing and showing expected biological patterns on known CMR phenotypes. DCM's diagnosis is defined by the presence of left ventricular (LV) under certain conditions and by excluding other cardiac abnormalities (Pinto et al., 2016). Differences in distribution between high and low PRS groups were shown for 6 derived CMR phenotypes from 2D imaging data in the UKBB. These six phenotypes relate to the circumferential, radial, and longitudinal strain. Strain evaluates how the heart muscle changes shape in three different ways. Longitudinal strain assesses the contraction of the heart muscle from its base to its apex. Circumferential strain examines the narrowing of the heart's short axis during the contraction phase. Radial strain evaluates the thickening of the myocardium from the inner layer (endocardium) to the outer layer (epicardium) (Zhang et al., 2019). Results shown above indicate a slight decrease in circumferential, radial, and longitudinal strain in the group with high PRS. Given the low sample size of individuals with CMR imaging

data, genotyping data, and a DCM diagnosis the statistical power in discerning a strong association is affected. Nevertheless these results are in concordance to trends found in these strain traits in relation to DCM and heart failure (Vietheer et al., 2021, Yu et al., 2017).

To further assess the utility of PRS in a rare disease like DCM, odds ratio values were examined for different PRS quantiles starting from lower risk ranges and going up to the 99^{th} percentile. The upward trend shown indicates a higher odds ratio with an increase in the PRS, which is in line with a higher association of developing DCM given a higher PRS. In addition, there is some evidence of earlier onset of disease in the group with high PRS, especially when incorporating the validation set in order to increase numbers. However, again the low sample size does not allow for statistical power in the association.

Overall, the low case sample sizes in either diseased cohorts (RBH) or general population biobanks (UKBB) for rare diseases are one of the shortcomings of using a machine learning method to generate PRS, in contrast to the studies mentioned before on non-rare outcomes. To mitigate this the robust pipeline approach of bootstrapping samples across the different sets was described above. Other machine learning methods can also be employed, that have a lower chance of overfitting and can deal with lower samples sizes (Support Vector Machines, Random Forests), however these often do not perform as well as the neural networks. In particular, traditional Random Forests are not able to handle correlated datasets in an efficient way, therefore in order to mitigate the LD associations pruning or clumping of the SNPs would have been required, similar to what is needed in the classic statistical PRS methods (ie PRS-CS) (Choi et al., 2020).

LAVASET, presented previously in Chapter 4.5, was developed to address this issue of Random Forests when dealing with correlated datasets. Therefore, an additional consideration

would be to use LAVASET as another method to train and calculate PRS, as part of incorporating machine learning in these pipelines. LAVASET does however require defining the correlational relationships of the SNPs, which can mostly be done using these LD reference panels. This would defeat the purpose of using machine learning to decipher the SNP correlations while training, which is how the NN architecture described above implements it. Regardless, a next step would be to compare the performance of LAVASET in relation to the classical statistical methods using LD reference panels.

With the increased incorporation of generative models in biomedical contexts, another consideration for the low sample size would be training Generative Adversarial Networks (GANs). GANs are able to sustain good predictions by performing synthetic data augmentation, based on real-data observations, and therefore increase the sample size while training (Hsu and Lin, 2020). A study compared the performance of Deep Boltzmann Machines (DBMs), Variational Autoencoders (VAEs), and GANs in learning joint distributions of real and simulated SNP data under sample size constraints. DBMs are the most efficient in learning the combined distributions of real and synthetic SNP data; while VAEs can understand the underlying distributions of the real dataset, however are not able to effectively generate synthetic data that closely resemble the original. GANs needed larger sample sizes in order to produce equivalent results to the DBMs and VAEs (Nußberger et al., 2021). A combination of these models, or DBMs alone, could potentially be used as alternatives or complimentary to the NN-PRS models presented here, so as to increase the test size numbers for evaluating the PRS, by reducing the samples used in training.

6.5 Conclusion

The complex genetic architecture of cardiomyopathies goes beyond the Mendelian and familial inheritance patterns, as reaffirmed here and in recent GWAS. For this reason, PRS can provide useful insights of the missing genetic information even in such rare diseases. Here, a deep learning model approach (NN-PRS) was introduced for generating and evaluating PRS in HCM and DCM cases separately.

Specifically, the median PRS value of HCM cases is significantly higher to the Controls median, indicating a higher probability of disease for HCM, as expected. A higher median PRS is also evident when looking at PRS distributions between DCM cases and Controls. In these DCM models, it was also shown that NN-PRS achieved a better separation of the DCM diseased group and Controls than the classical PRS-CS method (NN-PRS:R^2=4.22×10^{-3} vs PRS-CS R^2=4.14×10^{-4}).

In addition to raw PRS scores trending higher for CM status, cardiac anatomical phenotypes and related clinical outcomes to HCM and DCM also showed significant association with higher PRS. For HCM, maximum wall thickness of the left ventricle is the strongest clinical indicator of disease, along with the left ventricular mass and end-systolic volume. Here, it was shown that the group with higher risk of developing HCM (based on NN-PRS scores) has higher values of maximum wall thickness and mass of the left ventricle, while lower values in end-systolic volume. These findings are in line with the biological explanation of an increased wall thickness leading to lower volumes in end of systole, and the anatomical expectations of enlarged thickness and mass in HCM phenotypes.

For DCM, differences in cardiac strain distributions, a sign of myocardium deformation, are shown between high and low PRS groups, as derived from 2D CMR imaging data. Specif-

ically, the higher PRS group shows differences in circumferential, radial, and longitudinal strains that follow known patterns of reduced and impaired strain, while the lower PRS groups presented strain measurements closer to unaffected phenotypes. NN-PRS association with the DCM phenotype and clinical outcome was further corroborated by the upward trend of odds ratios and PRS scores, indicating a higher chance of disease for individuals in higher NN-PRS quantiles. Finally, when comparing the cumulative count of DCM diagnoses between the higher and lower PRS groups, there is a discernible difference at onset age of disease between the high and low risk PRS groups, with the high risk group having an earlier and steeper onset of disease.

Overall, NN-PRS showed associations of higher PRS groups with higher risk in cardiomyopathies or its related outcomes and phenotypes (i.e. left ventricle traits). NN-PRS was also able to outperform the traditional PRS-CS method by one order of magnitude in the context of DCM, when stratifying for DCM and Control groups based on the output scores. This paves the way for 1) incorporating PRS and polygenic information in evaluating, assessing and potentially diagnosing cardiomyopathies, while 2) eventually also introducing machine-learning-based PRS methods to improve the calculations.

Chapter 7

Concluding Remarks

7.1 Concluding Remarks

The use of machine learning is expanding constantly in translational settings and new information is discovered on the complex genetic and phenotypic architectures of cardiomyopathies. This can be leveraged for developing new methodologies specific to the the nuanced and complex nature of biological datasets or creating AI tools that can automate or simplify diagnostic procedures securely and with interpretable means.

In this thesis, I have introduced a novel methodology in Chapter 4 to address the prevalent challenge of correlation in biological datasets, facilitating dimensionality reduction and the extraction of key features from CMR imaging datasets, which are integral for the diagnosis of cardiomyopathies. I demonstrated that LAVASET enables the selection of significant features in a manner that is both biologically-informed and inclusive of correlation, thus selecting known cardiomyopathy phenotypes with higher accuracy. Additionally, the utility of LAVASET was presented using an example ECG dataset, laying the groundwork for subsequent research on ECG data within the UK Biobank concerning HCM and DCM. This opens up future aspects for the project, particularly the ability to derive meaningful features from various imaging datasets (CMR and ECG) with the goal of discovering linkages between them and other -omics. An example of that would be establishing novel phenotypes for the identification of new genetic correlations through GWAS, similarly to what has been implemented in previous studies involving CMR imaging of the aortic valve (Córdova-Palomera et al., 2020). One can envision that a GWAS might associate ECG and CMR imaging characteristics identified by LAVASET to uncover novel associations.

I also highlight another application of machine learning, to predict cardiomyopathies using an interpretable approach with available genomic and clinical data from patient records.

Such models can serve two primary purposes. One function could be as an initial screening tool where an AI algorithm reviews general practitioner records to identify patients with potential, undiagnosed cases of cardiomyopathy. The second use case could be to aid in stratifying patients for whom the cardiomyopathy question has been raised by specialists. These individuals would have had additional diagnostic tests by default (such as CMR imaging). An AI algorithm would prove beneficial if it could accurately pinpoint those at low risk who may not need further costly evaluations. The specificity and sensitivity required for these two applications differ significantly. Based on the findings in Chapter 5, the latter application may be more suitable, especially considering the significant improvement in model performance when LAVASET-derived CMR features are included. This is pertinent for individuals already under scrutiny for cardiomyopathy. Additionally, the low prevalence of cardiomyopathies contributes to lower precision in real-world figures, rendering the pre-screening task more challenging. An element not yet included in these models is family history, which if integrated, could enhance the models' effectiveness with or without CMR features, thus potentially making the models viable for both pre-screening in the general population and stratifying those suspected of having cardiomyopathy.

In scenarios where interpretability is less critical, advanced deep learning methods can discern genomic links to define the polygenic architecture of cardiomyopathies. Chapter 6 demonstrated a deep learning model for computing PRS that depends solely on genotype information from the target dataset, bypassing the need for linkage disequilibrium calculation using computationally-expensive LD reference panels like UK Biobank or the 1,000 Genomes Project. This approach customizes the analysis to the data at hand. Deep learning models also streamline the process of generating PRS by defining genomic associations and correlations

during training, allowing for flexible data manipulation such as bootstrapping and subsampling, which reduces computational time. Most importantly however, it is corroborated that PRS' derived from deep learning are associated with the risk of cardiomyopathy and its related traits (Zheng et al., 2023a,b). For HCM, individuals with higher PRS quantiles displayed more pronounced left ventricular traits consistent with HCM-related phenotypes. For DCM, PRS revealed an increasing pattern of DCM odds ratios from lower to higher scores, with a more distinct onset of disease in the highest PRS group (80%). These findings underscore the value of PRS as a significant metric in evaluating cardiomyopathies, evolving the traditional view of these diseases as strictly monogenic or Mendelian and driven only by rare variants. Finally, incorporating complete genomic profiles, by adding PRS to the interpretable models discussed above and in Chapter 5, will increase performance and render them more comprehensive.

In conclusion, throughout this thesis I have shown how machine learning methods can be used to further advance our knowledge of polygenic architecture in cardiomyopathies, assess their current utility state for stratifying or pre-screening, and explore new methodologies to deal with such multidimensional complex biological datasets.

Appendix

7.2 Appendix

SPRINGER NATURE LICENSE TERMS AND CONDITIONS

Nov 10, 2023

This Agreement between Ms. Melpomeni Kasapi (”You”) and Springer Nature (”Springer Nature”) consists of your license details and the terms and conditions provided by Springer Nature and Copyright Clearance Center.

License Number: 5658980947445

License date: Oct 30, 2023

Licensed Content Publisher: Springer Nature

Licensed Content Publication: Nature Reviews Cardiology

Licensed Content Title: Dilated cardiomyopathy: the complexity of a diverse genetic architecture

Licensed Content Author: Ray E. Hershberger et al

Licensed Content Date: Jul 30, 2013

Type of Use: Thesis/Dissertation

Requestor type: academic/university or research institute

Format: electronic

Portion: figures/tables/illustrations

Title of new work: Interpretable machine learning methods for investigating and pre-screening cardiomyopathies with genomic and imaging data

Institution name: Imperial College London

Expected presentation date: Nov 2023

Portions: Figure 4: Morphology of the cardiomyopathies.

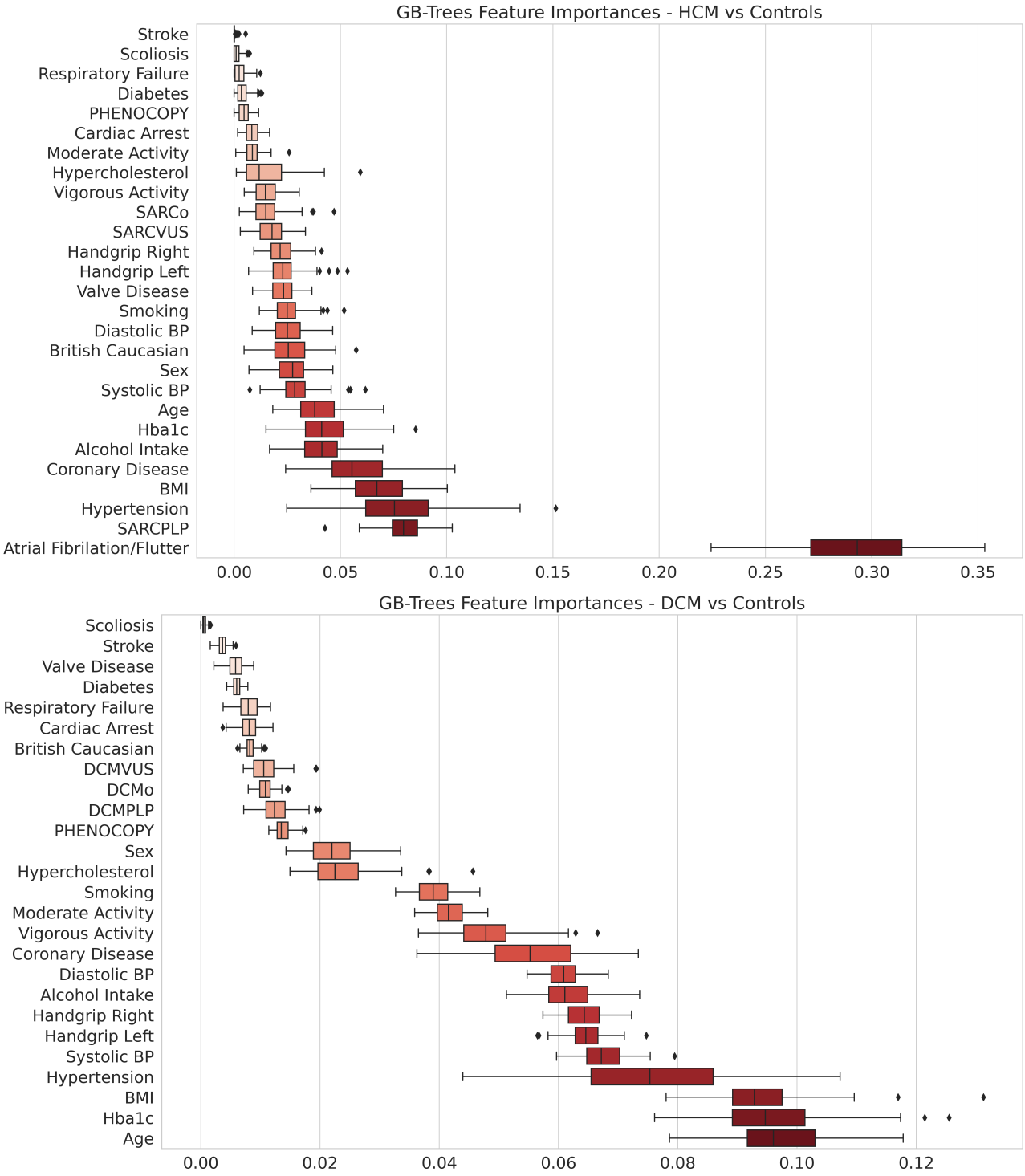


Figure 7.1: Feature Importances for the HCM/DCM vs Control prediction models in GB-Trees. Bars show the distribution of feature importance scores as they were generated for each of the 100 distinct iterations. The Sex variable was coded based on the UKBB data encoding, with 1 depicting males and 0 depicting females.

Bibliography

2023 esc guidelines for the management of cardiomyopathies — european heart journal — oxford academic. URL https://academic.oup.com/eurheartj/advance-article/doi/10.1093/eurheartj/ehad194/7246608. pages 110, 111, 135

Nhs england » nhs genomic medicine service, 2023. URL https://www.england.nhs.uk/genomics/nhs-genomic-med-service/#. pages 103

Tariq Ahmad, Lars H. Lund, Pooja Rao, Rohit Ghosh, Prashant Warier, Benjamin Vaccaro, Ulf Dahlström, Christopher M. O'Connor, G. Michael Felker, and Nihar R. Desai. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *Journal of the American Heart Association*, 7, 4 2018. ISSN 20479980. doi: 10.1161/JAHA.117.008081. URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.117.008081. pages 75

Ahmed M. Alaa, Thomas Bolton, Emanuele Di Angelantonio, James H.F. Rudd, and Mihaela van der Schaar. Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 uk biobank participants. *PLoS ONE*, 14, 5 2019. ISSN 19326203. doi: 10.1371/JOURNAL.PONE.0213653. URL /pmc/articles/PMC6519796//pmc/articles/PMC6519796/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6519796/. pages 75

Elena Arbelo, Alexandros Protonotarios, Juan R Gimeno, Eloisa Arbustini, Roberto Barriales-Villa,

Cristina Basso, Connie R Bezzina, Elena Biagini, Nico A Blom, Rudolf A de Boer, Tim De Winter, Perry M Elliott, Marcus Flather, Pablo Garcia-Pavia, Kristina H Haugaa, Jodie Ingles, Ruxandra Oana Jurcut, Sabine Klaassen, Giuseppe Limongelli, Bart Loeys, Jens Mogensen, Iacopo Olivotto, Antonis Pantazis, Sanjay Sharma, J Peter Van Tintelen, James S Ware, Juan Pablo Kaski, ESC Scientific Document Group, Philippe Charron, Massimo Imazio, Magdy Abdelhamid, Victor Aboyans, Michael Arad, Folkert W Asselbergs, Riccardo Asteggiano, Zofia Bilinska, Damien Bonnet, Henning Bundgaard, Nuno Miguel Cardim, Jelena Čelutkienė, Maja Cikes, Gaetano Maria De Ferrari, Veronica Dusi, Volkmar Falk, Laurent Fauchier, Estelle Gandjbakhch, Tiina Heliö, Konstantinos Koskinas, Dipak Kotecha, Ulf Landmesser, George Lazaros, Basil S Lewis, Ales Linhart, Maja-Lisa Løchen, Benjamin Meder, Richard Mindham, James Moon, Jens Cosedis Nielsen, Steffen Petersen, Eva Prescott, Mary N Sheppard, Gianfranco Sinagra, Marta Sitges, Jacob Tfelt-Hansen, Rhian Touyz, Rogier Veltrop, Josef Veselka, Karim Wahbi, Arthur Wilde, Katja Zeppenfeld, Brahim Kichou, Hamayak Sisakian, Daniel Scherr, Bernhard Gerber, Alen Džubur, Mariana Gospodinova, Ivo Planinc, Hera Heracleous Moustra, David Zemánek, Morten Steen Kvistholm Jensen, Ahmad Samir, Kairit Palm, Tiina Heliö, Karim Wahbi, Eric Schulze-Bahr, Vlachopoulos Haralambos, Róbert Sepp, Berglind Aalsteinsdóttir, Deirdre Ward, Miry Blich, Gianfranco Sinagra, Afrim Poniku, Olga Lunegova, Ainars Rudzitis, Roland Kassab, Jūratė Barysienė, Steve Huijnen, Tiziana Felice, Eleonora Vataman, Nikola Pavlovic, Nawal Doghmi, Folkert W Asselbergs, Elizabeta Srbinovska Kostovska, Vibeke Marie Almaas, Elżbieta Katarzyna Biernacka, Dulce Brito, Monica Rosca, Marco Zavatta, Arsen Ristic, Eva Goncalvesová, Matjaž Šinkovec, Victoria Cañadas-Godoy, Pyotr G Platonov, Ardan M Saguner, Ahmad Rasheed Al Saadi, Ikram Kammoun, Ahmet Celik, Elena Nesukay, Timur Abdullaev, Eva Prescott, Stefan James, Elena Arbelo, Colin Baigent, Michael A Borger, Sergio Buccheri, Borja Ibanez, Lars Køber, Konstantinos C Koskinas, John William McEvoy, Borislava Mihaylova, Richard Mindham, Lis Neubeck, Jens Cosedis Nielsen, Agnes Pasquet, Amina Rakisheva, Bianca

Rocca, Xavier Rossello, Ilonca Vaartjes, Christiaan Vrints, Adam Witkowski, and Katja Zeppenfeld. 2023 esc guidelines for the management of cardiomyopathies: Developed by the task force on the management of cardiomyopathies of the european society of cardiology (esc). *European Heart Journal*, 44:3503–3626, 10 2023. ISSN 0195-668X. doi: 10.1093/EURHEARTJ/EHAD194. URL https://dx.doi.org/10.1093/eurheartj/ehad194. pages 2, 5

Alessia Argirò, Carolyn Ho, Sharlene M. Day, Jolanda van der Velden, Elisabetta Cerbai, Sara Saberi, Jil C. Tardiff, Neal K. Lakdawala, and Iacopo Olivotto. Sex-related differences in genetic cardiomyopathies. *Journal of the American Heart Association*, 11:24947, 5 2022. ISSN 20479980. doi: 10.1161/JAHA.121.024947. URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.121.024947. pages 109

David Arthur and Sergei Vassilvitskii. How slow is the k-means method? *Proceedings of the Annual Symposium on Computational Geometry*, 2006:144–153, 2006. doi: 10.1145/1137856.1137880. URL https://dl.acm.org/doi/10.1145/1137856.1137880. pages 28

Zachi I. Attia, Suraj Kapa, Francisco Lopez-Jimenez, Paul M. McKie, Dorothy J. Ladewig, Gaurav Satam, Patricia A. Pellikka, Maurice Enriquez-Sarano, Peter A. Noseworthy, Thomas M. Munger, Samuel J. Asirvatham, Christopher G. Scott, Rickey E. Carter, and Paul A. Friedman. Screening for cardiac contractile dysfunction using an artificial intelligence–enabled electrocardiogram. *Nature Medicine 2019 25:1*, 25:70–74, 1 2019. ISSN 1546-170X. doi: 10.1038/s41591-018-0240-2. URL https://www.nature.com/articles/s41591-018-0240-2. pages 20

Joshua D. Backman, Alexander H. Li, Anthony Marcketta, Dylan Sun, Joelle Mbatchou, Michael D. Kessler, Christian Benner, Daren Liu, Adam E. Locke, Suganthi Balasubramanian, Ashish Yadav, Nilanjana Banerjee, Christopher E. Gillies, Amy Damask, Simon Liu, Xiaodong Bai, Alicia Hawes, Evan Maxwell, Lauren Gurski, Kyoko Watanabe, Jack A. Kosmicki, Veera Rajagopal, Jason Mighty, Marcus Jones, Lyndon Mitnaul, Eli Stahl, Giovanni Coppola, Eric Jorgenson, Lukas Habegger,

William J. Salerno, Alan R. Shuldiner, Luca A. Lotta, John D. Overton, Michael N. Cantor, Jeffrey G. Reid, George Yancopoulos, Hyun M. Kang, Jonathan Marchini, Aris Baras, Gonçalo R. Abecasis, and Manuel A.R. Ferreira. Exome sequencing and analysis of 454,787 uk biobank participants. *Nature 2021 599:7886*, 599:628–634, 10 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-04103-z. URL https://www.nature.com/articles/s41586-021-04103-z. pages 39

Adrien Badré, Li Zhang, Wellington Muchero, Justin C. Reynolds, and Chongle Pan. Deep neural network improves the estimation of polygenic risk scores for breast cancer. *Journal of Human Genetics 2020 66:4*, 66:359–369, 10 2020. ISSN 1435-232X. doi: 10.1038/s10038-020-00832-7. URL https://www.nature.com/articles/s10038-020-00832-7. pages 116, 119, 134

Wenjia Bai, Wenzhe Shi, Antonio de Marvao, Timothy J W Dawes, Declan P O'Regan, Stuart A Cook, and Daniel Rueckert. A bi-ventricular cardiac atlas built from 1000+ high resolution mr images of healthy subjects and an analysis of shape and motion. *Medical Image Analysis*, 26:133–145, 2015. ISSN 1361-8423. doi: 10.1016/j.media.2015.08.009. URL http://www.ncbi.nlm.nih.gov/pubmed/26387054. pages 14, 41

Wenjia Bai, Hideaki Suzuki, Chen Qin, Giacomo Tarroni, Ozan Oktay, Paul M. Matthews, and Daniel Rueckert. Recurrent neural networks for aortic image sequence segmentation with sparse annotations. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11073 LNCS:586–594, 8 2018. ISSN 16113349. doi: 10.1007/978-3-030-00937-3_67. URL https://arxiv.org/abs/1808.00273v1. pages 14, 60

Marcel Beetz, Abhirup Banerjee, and Vicente Grau. Multi-domain variational autoencoders for combined modeling of mri-based biventricular anatomy and ecg-based cardiac electrophysiology. *Frontiers in physiology*, 13, 6 2022. ISSN 1664-042X. doi: 10.3389/FPHYS.2022.886723. URL https://pubmed.ncbi.nlm.nih.gov/35755443/. pages 72

Ghalib A Bello, Timothy J W Dawes, Jinming Duan, Carlo Biffi, Antonio de Marvao, Luke S G E Howard, J Simon R Gibbs, Martin R Wilkins, Stuart A Cook, Daniel Rueckert, and Declan P O'Regan. Deep-learning cardiac motion analysis for human survival prediction. *Nature Machine Intelligence*, 1:95–104, 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0019-2. URL https://www.nature.com/articles/s42256-019-0019-2internal-pdf://280/s42256-019-0019-2.html. pages 14, 40

Juan Jose Beunza, Enrique Puertas, E. García-Ovejero, Gema Villalba, Emilia Condes, Gergana Koleva, Cristian Hurtado, and Manuel F. Landecho. Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of Biomedical Informatics*, 97:103257, 9 2019. ISSN 1532-0464. doi: 10.1016/J.JBI.2019.103257. pages 105

Sebastian E. Beyer, Mihir M. Sanghvi, Nay Aung, Alice Hosking, Jackie A. Cooper, José Miguel Paiva, Aaron M. Lee, Kenneth Fung, Elena Lukaschuk, Valentina Carapella, Murray A. Mittleman, Soren Brage, Stefan K. Piechnik, Stefan Neubauer, and Steffen E. Petersen. Prospective association between handgrip strength and cardiac structure and function in uk adults. *PLoS ONE*, 13, 3 2018. ISSN 19326203. doi: 10.1371/JOURNAL.PONE.0193124. URL /pmc/articles/PMC5851546//pmc/articles/PMC5851546/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5851546/. pages 111

Carlo Biffi, Antonio de Marvao, Mark I Attard, Timothy J W Dawes, Nicola Whiffin, Wenjia Bai, Wenzhe Shi, Catherine Francis, Hannah Meyer, Rachel Buchan, Stuart A Cook, Daniel Rueckert, and Declan P O'Regan. Three-dimensional cardiovascular imaging-genetics: a mass univariate framework. *Bioinformatics*, 34:97–103, 2018. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx552. URL https://academic.oup.com/bioinformatics/article/34/1/97/4103396internal-pdf://286/4103396.html. pages 13

Christopher M. Bishop. Pattern recogniton and machine learning. *Informa-*

tion Science and Statistics, page 738, 2006. URL https://www.springer.com/gp/book/9780387310732http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop-PatternRecognitionAndMachineLearning-Springer2006.pdf. pages 26

Rodrigo Bonazzola, Nishant Ravikumar, Rahman Attar, Enzo Ferrante, Tanveer Syeda-Mahmood, and Alejandro F. Frangi. Image-derived phenotype extraction for genetic discovery via unsupervised deep learning in cmr images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 12905 LNCS:699–708, 9 2021. ISSN 16113349. doi: 10.1007/978-3-030-87240-3_67. URL https://link.springer.com/chapter/10.1007/978-3-030-87240-3_67. pages 15

R. Bousseljot, D. Kreiseler, and A. Schnabel. Nutzung der EKG-signaldatenbank CARDIODAT der PTB über das internet. *Biomedizinische Technik/Biomedical Engineering*, 40(s1):317–318, July 1995. doi: 10.1515/bmte.1995.40.s1.317. URL https://doi.org/10.1515/bmte.1995.40.s1.317. pages 59, 68

Eugene Braunwald. Cardiomyopathies: An overview. *Circulation research*, 121:711–721, 9 2017. ISSN 1524-4571. doi: 10.1161/CIRCRESAHA.117.311812. URL https://pubmed.ncbi.nlm.nih.gov/28912178/. pages 2

L. Breiman, Jerome H. Friedman, Richard A. Olshen, and C. J. Stone. Classification and regression trees. In *Classification and Regression Trees*, page 358, New York, 1984. Chapman & Hall. pages 61

Leo. Breiman. Classification and regression trees. page 358, 1984. URL https://books.google.com/books/about/Classification_and_Regression_Trees.html?id=uxPvAAAAMAAJ. pages 29

Leo Breiman. Random forests. *Mach. Learn.*, 45:5–32, 10 2001. ISSN 08856125. doi: 10.1023/a:1010933404324. pages 28, 29, 57

Alexandra Butters, Neal K. Lakdawala, and Jodie Ingles. Sex differences in hypertrophic

cardiomyopathy: Interaction with genetics and environment. *Current Heart Failure Reports*, 18:264, 10 2021. ISSN 15469549. doi: 10.1007/S11897-021-00526-X. URL /pmc/articles/PMC8484093//pmc/articles/PMC8484093/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8484093/. pages 107

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, Adrian Cortes, Samantha Welsh, Alan Young, Mark Effingham, Gil McVean, Stephen Leslie, Naomi Allen, Peter Donnelly, and Jonathan Marchini. The uk biobank resource with deep phenotyping and genomic data. *Nature 2018 562:7726*, 562:203–209, 10 2018. ISSN 1476-4687. doi: 10.1038/s41586-018-0579-z. URL https://www.nature.com/articles/s41586-018-0579-zhttps://www.nature.com/articles/s41586-018-0579-z. pages 39

Christopher C. Chang, Carson C. Chow, Laurent C.A.M. Tellier, Shashaank Vattikuti, Shaun M. Purcell, and James J. Lee. Second-generation plink: Rising to the challenge of larger and richer datasets. *GigaScience*, 4:7, 2 2015. ISSN 2047217X. doi: 10.1186/S13742-015-0047-8/2707533. URL https://dx.doi.org/10.1186/s13742-015-0047-8. pages 10

Shi-Yi Chen, Zhe Feng, and Xiaolian Yi. A general introduction to adjustment for multiple comparisons. *Journal of Thoracic Disease*, 9:1725–1729, 2017. ISSN 2072-1439. doi: 10.21037/jtd.2017.05.34. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506159/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5506159/. pages 47

Shing Wan Choi and Paul F. O'Reilly. Prsice-2: Polygenic risk score software for biobank-scale data. *GigaScience*, 8:1–6, 7 2019. ISSN 2047217X. doi: 10.1093/GIGASCIENCE/GIZ082. URL https://dx.doi.org/10.1093/gigascience/giz082. pages 10

Shing Wan Choi, Timothy Shin Heng Mak, and Paul F. O'Reilly. A guide to performing polygenic

risk score analyses. *Nature protocols*, 15:2759, 9 2020. ISSN 17502799. doi: 10.1038/S41596-020-0353-1. URL /pmc/articles/PMC7612115//pmc/articles/PMC7612115/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7612115/. pages 10, 138

Samuel A. Clark and Julius Van Der Werf. Genomic best linear unbiased prediction (gblup) for the estimation of genomic breeding values. *Methods in Molecular Biology*, 1019:321–330, 2013. ISSN 10643745. doi: 10.1007/978-1-62703-447-0_13/COVER. URL https://link.springer.com/protocol/10.1007/978-1-62703-447-0_13. pages 115

M. B. Codd, D. D. Sugrue, B. J. Gersh, and L. J. Melton. Epidemiology of idiopathic dilated and hypertrophic cardiomyopathy. a population-based study in olmsted county, minnesota, 1975-1984. *Circulation*, 80:564–572, 1989. ISSN 0009-7322. doi: 10.1161/01.CIR.80.3.564. URL https://pubmed.ncbi.nlm.nih.gov/2766509/https://pubmed.ncbi.nlm.nih.gov/2766509/?dopt=Abstract. pages 5

Ryan L. Collins, Harrison Brand, Konrad J. Karczewski, Xuefang Zhao, Jessica Alföldi, Laurent C. Francioli, Amit V. Khera, Chelsea Lowther, Laura D. Gauthier, Harold Wang, Nicholas A. Watts, Matthew Solomonson, Anne O'Donnell-Luria, Alexander Baumann, Ruchi Munshi, Mark Walker, Christopher W. Whelan, Yongqing Huang, Ted Brookings, Ted Sharpe, Matthew R. Stone, Elise Valkanas, Jack Fu, Grace Tiao, Kristen M. Laricchia, Valentin Ruano-Rubio, Christine Stevens, Namrata Gupta, Caroline Cusick, Lauren Margolin, Jessica Alföldi, Irina M. Armean, Eric Banks, Louis Bergelson, Kristian Cibulskis, Ryan L. Collins, Kristen M. Connolly, Miguel Covarrubias, Beryl Cummings, Mark J. Daly, Stacey Donnelly, Yossi Farjoun, Steven Ferriera, Laurent Francioli, Stacey Gabriel, Laura D. Gauthier, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Konrad J. Karczewski, Kristen M. Laricchia, Christopher Llanwarne, Eric V. Minikel, Ruchi Munshi, Benjamin M. Neale, Sam Novod, Anne H. O'Donnell-Luria, Nikelle Petrillo, Timothy Poterba, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Kaitlin E. Samocha, Molly Schleicher,

Cotton Seed, Matthew Solomonson, Jose Soto, Grace Tiao, Kathleen Tibbetts, Charlotte Tolonen, Christopher Vittal, Gordon Wade, Arcturus Wang, Qingbo Wang, James S. Ware, Nicholas A. Watts, Ben Weisburd, Nicola Whiffin, Carlos A.Aguilar Salinas, Tariq Ahmad, Christine M. Albert, Diego Ardissino, Gil Atzmon, John Barnard, Laurent Beaugerie, Emelia J. Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, John C. Chambers, Juliana C. Chan, Daniel Chasman, Judy Cho, Mina K. Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J.F. Loos, Steven A. Lubitz, Ronald C.W. Ma, Daniel G. MacArthur, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Peter M. Nilsson, Michael C. O'Donovan, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin N.A. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Nazneen Rahman, Anne M. Remes, John D. Rioux, Samuli Ripatti, Dan M. Roden, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Jeremiah Scharf, Heribert Schunkert, Moore B. Shoemaker, Pamela Sklar, Hilkka Soininen, Harry Sokol, Tim Spector, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Yik Ying Teo, Tuomi Tiinamaija, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, Hugh Watkins, Rinse K. Weersma, Maija Wessman, James G. Wilson, Ramnik J. Xavier, Kent D. Taylor, Henry J. Lin, Stephen S. Rich, Wendy S. Post, Yii Der Ida Chen, Jerome I. Rotter, Chad Nusbaum, Anthony Philippakis, Eric Lander, Stacey Gabriel, Sekar Kathiresan, Eric Banks, and Michael E. Talkowski. A structural variation reference for medical

and population genetics. *Nature 2020 581:7809*, 581:444–451, 5 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2287-8. URL https://www.nature.com/articles/s41586-020-2287-8. pages 81

Steven S. Coughlin, James R. Labenberg, and Mariella C. Tefft. Black-white differences in idiopathic dilated cardiomyopathy: the washington dc dilated cardiomyopathy study. *Epidemiology (Cambridge, Mass.)*, 4:165–172, 1993. ISSN 1044-3983. doi: 10.1097/00001648-199303000-00013. URL https://pubmed.ncbi.nlm.nih.gov/8452906/. pages 5

Lara Curran, Antonio de Marvao, Paolo Inglese, Kathryn A McGurk, Pierre-Raphaël Schiratti, Adam Clement, Sean L Zheng, Surui Li, Chee Jian Pua, Mit Shah, Mina Jafari, Pantazis Theotokis, Rachel J Buchan, Sean J Jurgens, Claire E Raphael, Arun John Baksi, Antonis Pantazis, Brian P Halliday, Dudley J Pennell, Wenjia Bai, Calvin W L Chin, Rafik Tadros, Connie R Bezzina, Hugh Watkins, Stuart A Cook, Sanjay K Prasad, James S Ware, and Declan P O. A genotype-phenotype taxonomy of hypertrophic cardiomyopathy. 2023. doi: 10.1101/2023.03.11.23285908. URL https://doi.org/10.1101/2023.03.11.23285908. pages 15, 38

Aldo Córdova-Palomera, Catherine Tcheandjieu, Jason A. Fries, Paroma Varma, Vincent S. Chen, Madalina Fiterau, Ke Xiao, Heliodoro Tejeda, Bernard D. Keavney, Heather J. Cordell, Yosuke Tanigawa, Guhan Venkataraman, Manuel A. Rivas, Christopher Ré, Euan Ashley, and James R. Priest. Cardiac imaging of aortic valve area from 34 287 uk biobank participants reveals novel genetic associations and shared genetic comorbidity with multiple disease phenotypes. *Circulation: Genomic and Precision Medicine*, 13:E003014, 12 2020. ISSN 25748300. doi: 10.1161/CIRCGEN.120.003014. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCGEN.120.003014. pages 143

Eileen O. Dareng, Jonathan P. Tyrer, Daniel R. Barnes, Michelle R. Jones, Xin Yang, Katja K.H. Aben, Muriel A. Adank, Simona Agata, Irene L. Andrulis, Hoda Anton-Culver, Natalia N. Antonenkova,

Gerasimos Aravantinos, Banu K. Arun, Annelie Augustinsson, Judith Balmaña, Elisa V. Bandera, Rosa B. Barkardottir, Daniel Barrowdale, Matthias W. Beckmann, Alicia Beeghly-Fadiel, Javier Benitez, Marina Bermisheva, Marcus Q. Bernardini, Line Bjorge, Amanda Black, Natalia V. Bogdanova, Bernardo Bonanni, Ake Borg, James D. Brenton, Agnieszka Budzilowska, Ralf Butzow, Saundra S. Buys, Hui Cai, Maria A. Caligo, Ian Campbell, Rikki Cannioto, Hayley Cassingham, Jenny Chang-Claude, Stephen J. Chanock, Kexin Chen, Yoke Eng Chiew, Wendy K. Chung, Kathleen B.M. Claes, Sarah Colonna, Linda S. Cook, Fergus J. Couch, Mary B. Daly, Fanny Dao, Eleanor Davies, Miguel de la Hoya, Robin de Putter, Joe Dennis, Allison DePersia, Peter Devilee, Orland Diez, Yuan Chun Ding, Jennifer A. Doherty, Susan M. Domchek, Thilo Dörk, Andreas du Bois, Matthias Dürst, Diana M. Eccles, Heather A. Eliassen, Christoph Engel, Gareth D. Evans, Peter A. Fasching, James M. Flanagan, Renée T. Fortner, Eva Machackova, Eitan Friedman, Patricia A. Ganz, Judy Garber, Francesca Gensini, Graham G. Giles, Gord Glendon, Andrew K. Godwin, Marc T. Goodman, Mark H. Greene, Jacek Gronwald, Eric Hahnen, Christopher A. Haiman, Niclas Håkansson, Ute Hamann, Thomas V.O. Hansen, Holly R. Harris, Mikael Hartman, Florian Heitz, Michelle A.T. Hildebrandt, Estrid Høgdall, Claus K. Høgdall, John L. Hopper, Ruea Yea Huang, Chad Huff, Peter J. Hulick, David G. Huntsman, Evgeny N. Imyanitov, Annemieke H. van der Hout, Claudine Isaacs, Anna Jakubowska, Paul A. James, Ramunas Janavicius, Allan Jensen, Oskar Th Johannsson, Esther M. John, Daehee Kang, Beth Y. Karlan, Anthony Karnezis, Linda E. Kelemen, Elza Khusnutdinova, Lambertus A. Kiemeney, Byoung Gie Kim, Susanne K. Kjaer, Ian Komenaka, Jolanta Kupryjanczyk, Allison W. Kurian, Ava Kwong, Diether Lambrechts, Melissa C. Larson, Conxi Lazaro, Nhu D. Le, Goska Leslie, Jenny Lester, Fabienne Lesueur, Douglas A. Levine, Lian Li, Jingmei Li, Jennifer T. Loud, Karen H. Lu, Jan Lubinski, Phuong L. Mai, Siranoush Manoukian, Jeffrey R. Marks, Rayna Kim Matsuno, Keitaro Matsuo, Taymaa May, Lesley McGuffog, John R. McLaughlin, Iain A. McNeish, Noura Mebirouk, Usha Menon, Austin Miller, Roger L.

Milne, Albina Minlikeeva, Francesmary Modugno, Marco Montagna, Kirsten B. Moysich, Elizabeth Munro, Katherine L. Nathanson, Susan L. Neuhausen, Heli Nevanlinna, Joanne Ngeow Yuen Yie, Henriette Roed Nielsen, Finn C. Nielsen, Liene Nikitina-Zake, Kunle Odunsi, Kenneth Offit, Edith Olah, Siel Olbrecht, Olufunmilayo I. Olopade, Sara H. Olson, Håkan Olsson, Ana Osorio, Laura Papi, Sue K. Park, Michael T. Parsons, Harsha Pathak, Inge Sokilde Pedersen, Ana Peixoto, Tanja Pejovic, Pedro Perez-Segura, Jennifer B. Permuth, Beth Peshkin, Paolo Peterlongo, Anna Piskorz, Darya Prokofyeva, Paolo Radice, Johanna Rantala, Marjorie J. Riggan, Harvey A. Risch, Cristina Rodriguez-Antona, Eric Ross, Mary Anne Rossing, Ingo Runnebaum, Dale P. Sandler, Marta Santamariña, Penny Soucy, Rita K. Schmutzler, V. Wendy Setiawan, Kang Shan, Weiva Sieh, Jacques Simard, Christian F. Singer, Anna P. Sokolenko, Honglin Song, Melissa C. Southey, Helen Steed, Dominique Stoppa-Lyonnet, Rebecca Sutphen, Anthony J. Swerdlow, Yen Yen Tan, Manuel R. Teixeira, Soo Hwang Teo, Kathryn L. Terry, Mary Beth Terry, Liv Cecilie Vestrheim Thomsen, Darcy L. Thull, Marc Tischkowitz, Linda Titus, Amanda E. Toland, Diana Torres, Britton Trabert, Ruth Travis, Nadine Tung, Shelley S. Tworoger, Ellen Valen, Anne M. van Altena, Annemieke H. van der Hout, Els Van Nieuwenhuysen, Elizabeth J. van Rensburg, Ana Vega, Digna Velez Edwards, Robert A. Vierkant, Frances Wang, Barbara Wappenschmidt, Penelope M. Webb, Clarice R. Weinberg, Jeffrey N. Weitzel, Nicolas Wentzensen, Emily White, Alice S. Whittemore, Stacey J. Winham, Alicja Wolk, Yin Ling Woo, Anna H. Wu, Li Yan, Drakoulis Yannoukakos, Katia M. Zavaglia, Wei Zheng, Argyrios Ziogas, Kristin K. Zorn, Zdenek Kleibl, Douglas Easton, Kate Lawrenson, Anna DeFazio, Thomas A. Sellers, Susan J. Ramus, Celeste L. Pearce, Alvaro N. Monteiro, Julie Cunningham, Ellen L. Goode, Joellen M. Schildkraut, Andrew Berchuck, Georgia Chenevix-Trench, Simon A. Gayther, Antonis C. Antoniou, and Paul D.P. Pharoah. Polygenic risk modeling for prediction of epithelial ovarian cancer risk. *European Journal of Human Genetics 2021 30:3*, 30:349–362, 1 2022. ISSN 1476-5438. doi: 10.1038/s41431-021-00987-7. URL

https://www.nature.com/articles/s41431-021-00987-7. pages 116, 119, 134

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. *ACM International Conference Proceeding Series*, 148:233–240, 2006. doi: 10.1145/1143844.1143874. URL https://www.researchgate.net/publication/215721831_The_Relationship_Between_Precision-Recall_and_ROC_Curves. pages 35

Antonio de Marvao, Kathryn A. McGurk, Sean L. Zheng, Marjola Thanaj, Wenjia Bai, Jinming Duan, Carlo Biffi, Francesco Mazzarotto, Ben Statton, Timothy J.W. Dawes, Nicolò Savioli, Brian P. Halliday, Xiao Xu, Rachel J. Buchan, A. John Baksi, Marina Quinlan, Paweł Tokarczuk, Upasana Tayal, Catherine Francis, Nicola Whiffin, Pantazis I. Theotokis, Xiaolei Zhang, Mikyung Jang, Alaine Berry, Antonis Pantazis, Paul J.R. Barton, Daniel Rueckert, Sanjay K. Prasad, Roddy Walsh, Carolyn Y. Ho, Stuart A. Cook, James S. Ware, and Declan P. O'Regan. Phenotypic expression and outcomes in individuals with rare genetic variants of hypertrophic cardiomyopathy. *Journal of the American College of Cardiology*, 78:1097–1110, 9 2021. ISSN 0735-1097. doi: 10.1016/J.JACC.2021.07.017. pages 17, 41

Jinming Duan, Ghalib Bello, Jo Schlemper, Wenjia Bai, Timothy J. W. Dawes, Carlo Biffi, Antonio de Marvao, Georgia Doumoud, Declan P. O'Regan, and Daniel Rueckert. Automatic 3d bi-ventricular segmentation of cardiac images by a shape-refined multi- task deep learning approach. *IEEE Transactions on Medical Imaging*, 38(9):2151–2164, 2019. doi: 10.1109/TMI.2019.2894322. pages 60

Luis Escobar-Lopez, Juan Pablo Ochoa, Jesús G. Mirelis, María Ángeles Espinosa, Marina Navarro, María Gallego-Delgado, Roberto Barriales-Villa, Ainhoa Robles-Mezcua, María Teresa Basurte-Elorz, Laura Gutiérrez García-Moreno, Vicente Climent, Juan Jiménez-Jaimez, María Victoria Mogollón-Jiménez, Javier Lopez, María Luisa Peña-Peña, Ana García-Álvarez, María Brion, Tomas Ripoll-Vera, Julián Palomino-Doza, Coloma Tirón, Uxua Idiazabal, Maria Noël Brögger, Soledad

García-Hernández, María Alejandra Restrepo-Córdoba, Esther Gonzalez-Lopez, Irene Méndez, María Sabater, Eduardo Villacorta, José M. Larrañaga-Moreira, Ana Abecia, Ana Isabel Fernández, José M. García-Pinilla, José F. Rodríguez-Palomares, Juan Ramón Gimeno-Blanes, Antoni Bayes-Genis, Enrique Lara-Pezzi, Fernando Domínguez, and Pablo Garcia-Pavia. Association of genetic variants with outcomes in patients with nonischemic dilated cardiomyopathy. *Journal of the American College of Cardiology*, 78:1682–1699, 10 2021. ISSN 0735-1097. doi: 10.1016/J.JACC.2021.08.039. pages 9

Jack Euesden, Cathryn M. Lewis, and Paul F. O'Reilly. Prsice: Polygenic risk score software. *Bioinformatics*, 31:1466–1468, 5 2015. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTU848. URL https://dx.doi.org/10.1093/bioinformatics/btu848. pages 10

DeLisa Fairweather, Danielle J. Beetler, Nicolas Musigk, Bettina Heidecker, Melissa A. Lyle, Leslie T. Cooper, and Katelyn A. Bruno. Sex and gender differences in myocarditis and dilated cardiomyopathy: An update. *Frontiers in Cardiovascular Medicine*, 10:1129348, 3 2023. ISSN 2297-055X. doi: 10.3389/fcvm.2023.1129348. URL https://www.frontiersin.org/articles/10.3389/fcvm.2023.1129348/full. pages 5, 109

Giulio Falasconi, Luigi Pannone, Massimo Slavich, Alberto Margonato, Gabriele Fragasso, and Roberto Spoladore. Atrial fibrillation in hypertrophic cardiomyopathy: pathophysiology, diagnosis and management. *American Journal of Cardiovascular Disease*, 10:409, 2020. ISSN 2160-200X. URL /pmc/articles/PMC7675160//pmc/articles/PMC7675160/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7675160/. pages 107

Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874, 6 2006. ISSN 0167-8655. doi: 10.1016/J.PATREC.2005.10.010. pages 34

Gherardo Finocchiaro, Michael Papadakis, Gaia Tanzarella, Harshil Dhutia, Chris Miles, Maite

Tome, Elijah R. Behr, Sanjay Sharma, and Mary N. Sheppard. Sudden death can be the first manifestation of hypertrophic cardiomyopathy: Data from a united kingdom pathology registry. *JACC. Clinical electrophysiology*, 5:252–254, 2 2019. ISSN 2405-5018. doi: 10.1016/J.JACEP.2018.11.004. URL https://pubmed.ncbi.nlm.nih.gov/30784699/https://pubmed.ncbi.nlm.nih.gov/30784699/?dopt=Abstract. pages 5

Jerome H. Friedman. Stochastic gradient boosting. *Computational Statistics and Data Analysis*, 38: 367–378, 2 2002. ISSN 01679473. doi: 10.1016/S0167-9473(01)00065-2. URL https://www.researchgate.net/publication/222573328_Stochastic_Gradient_Boosting. pages 30

Tian Ge, Chia Yen Chen, Yang Ni, Yen Chen Anne Feng, and Jordan W. Smoller. Polygenic prediction via bayesian regression and continuous shrinkage priors. *Nature Communications 2019 10:1*, 10: 1–10, 4 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-09718-5. URL https://www.nature.com/articles/s41467-019-09718-5. pages 115, 116

Raoof Gholami and Nikoo Fakhari. *Chapter 27 - Support Vector Machine: Principles, Parameters, and Applications*, pages 515–535. Academic Press, 2017. ISBN 978-0-12-811318-9. URL http://www.sciencedirect.com/science/article/pii/B9780128113189000272internal-pdf://237/B978012811318900027 2.html. pages 31, 32

Marta Gigli, Marco Merlo, Sharon L. Graw, Giulia Barbati, Teisha J. Rowland, Dobromir B. Slavov, Davide Stolfo, Mary E. Haywood, Matteo Dal Ferro, Alessandro Altinier, Federica Ramani, Francesca Brun, Andrea Cocciolo, Ilaria Puggia, Gaetano Morea, William J. McKenna, Francisco G. La Rosa, Matthew R.G. Taylor, Gianfranco Sinagra, and Luisa Mestroni. Genetic risk of arrhythmic phenotypes in patients with dilated cardiomyopathy. *Journal of the American College of Cardiology*, 74:1480–1490, 9 2019. ISSN 0735-1097. doi: 10.1016/J.JACC.2019.06.072. pages 9

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Deep sparse rectifier neural networks. 2011. pages

33

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23), June 2000. doi: 10.1161/01.cir.101.23.e215. URL https://doi.org/10.1161/01.cir.101.23.e215. pages 59

Sanna Gudmundsson, Moriel Singer-Berk, Nicholas A. Watts, William Phu, Julia K. Goodrich, Matthew Solomonson, Heidi L. Rehm, Daniel G. MacArthur, and Anne O'Donnell-Luria. Variant interpretation using population databases: Lessons from gnomad. *Human Mutation*, 43:1012–1030, 8 2022. ISSN 1098-1004. doi: 10.1002/HUMU.24309. URL https://onlinelibrary.wiley.com/doi/full/10.1002/humu.24309https://onlinelibrary.wiley.com/doi/abs/10.1002/humu.24309https://onlinelibrary.wiley.com/doi/10.1002/humu.24309. pages 7

Hanmin Guo, James J. Li, Qiongshi Lu, and Lin Hou. Detecting local genetic correlations with scan statistics. *Nature Communications 2021 12:1*, 12:1–13, 4 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-22334-6. URL https://www.nature.com/articles/s41467-021-22334-6. pages 53

Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1/3):389–422, 2002. doi: 10.1023/a:1012487302797. URL https://doi.org/10.1023/a:1012487302797. pages 58

Trine F. Haland, Nina E. Hasselberg, Vibeke Marie Almaas, Lars A. Dejgaard, Jørg Saberniak, Ida S. Leren, Knut Erik Berge, Kristina H. Haugaa, and Thor Edvardsen. Original research article: The systolic paradox in hypertrophic cardiomyopathy. *Open Heart*, 4, 5 2017. ISSN 20533624. doi: 10.1136/OPENHRT-2016-000571. URL /pmc/articles/PMC5471858//pmc/articles/PMC5471858/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5471858/. pages 136

Brian P. Halliday, Ankur Gulati, Aamir Ali, Simon Newsome, Amrit Lota, Upasana Tayal, Vassilios S. Vassiliou, Monika Arzanauskaite, Cemil Izgi, Kaushiga Krishnathasan, Arvind Singhal, Kayla Chiew, John Gregson, Michael P. Frenneaux, Stuart A. Cook, Dudley J. Pennell, Peter Collins, John G.F. Cleland, and Sanjay K. Prasad. Sex- and age-based differences in the natural history and outcome of dilated cardiomyopathy. *European Journal of Heart Failure*, 20:1392–1400, 10 2018. ISSN 1879-0844. doi: 10.1002/EJHF.1216. URL https://onlinelibrary.wiley.com/doi/full/10.1002/ejhf.1216https://onlinelibrary.wiley.com/doi/abs/10.1002/ejhf.1216https://onlinelibrary.wiley.com/doi/10.1002/ejhf.1216. pages 109

Andrew R. Harper, Anuj Goel, Christopher Grace, Kate L. Thomson, Steffen E. Petersen, Xiao Xu, Adam Waring, Elizabeth Ormondroyd, Christopher M. Kramer, Carolyn Y. Ho, Stefan Neubauer, Paul Kolm, Raymond Kwong, Sarahfaye F. Dolman, Patrice Desvigne-Nickens, John P. Dimarco, Nancy Geller, Dong Yun Kim, Cheng Zhang, William Weintraub, Theodore Abraham, Lisa Anderson, Evan Appelbaum, Camillo Autore, Colin Berry, Elena Biagini, William Bradlow, Chiara Bucciarelli-Ducci, Amedeo Chiribiri, Lubna Choudhury, Andrew Crean, Dana Dawson, Milind Y. Desai, Eleanor Elstein, Andrew Flett, Matthias Friedrich, Stephen Heitner, Adam Helms, Daniel L. Jacoby, Han Kim, Bette Kim, Eric Larose, Masliza Mahmod, Heiko Mahrholdt, Martin Maron, Gerry McCann, Michelle Michels, Saidi Mohiddin, Sherif Nagueh, David Newby, Iacopo Olivotto, Anjali Owens, F. Pierre-Mongeon, Sanjay Prasad, Ornella Rimoldi, Michael Salerno, Jeanette Schulz-Menger, Mark Sherrid, Peter Swoboda, Albert van Rossum, Jonathan Weinsaft, James White, Eric Williamson, Rafik Tadros, James S. Ware, Connie R. Bezzina, Martin Farrall, and Hugh Watkins. Common genetic variants and modifiable risk factors underpin hypertrophic cardiomyopathy susceptibility and expressivity. *Nature genetics*, 53:135–142, 2 2021. ISSN 1546-1718. doi: 10.1038/S41588-020-00764-0. URL https://pubmed.ncbi.nlm.nih.gov/33495597/. pages 11, 108, 110

Ray E. Hershberger, Ana Morales, and Jill D. Siegfried. Clinical and genetic issues in dilated

cardiomyopathy: a review for genetics professionals. *Genetics in medicine : official journal of the American College of Medical Genetics*, 12:655–667, 11 2010. ISSN 1530-0366. doi: 10.1097/GIM.0B013E3181F2481F. URL https://pubmed.ncbi.nlm.nih.gov/20864896/. pages 8

Ray E. Hershberger, Dale J. Hedges, and Ana Morales. Dilated cardiomyopathy: the complexity of a diverse genetic architecture. *Nature Reviews Cardiology 2013 10:9*, 10:531–547, 7 2013. ISSN 1759-5010. doi: 10.1038/nrcardio.2013.105. URL https://www.nature.com/articles/nrcardio.2013.105. pages 3, 5, 7

Ray E Hershberger, Michael M Givertz, Carolyn Y Ho, Daniel P Judge, Paul F Kantor, Kim L McBride, Ana Morales, Matthew R G Taylor, Matteo Vatta, and Stephanie M Ware. Genetic evaluation of cardiomyopathy: a clinical practice resource of the american college of medical genetics and genomics (acmg). *Genetics in Medicine*, 20:899–909, 2018. ISSN 1530-0366. doi: 10.1038/s41436-018-0039-z. URL https://www.nature.com/articles/s41436-018-0039-zinternal-pdf://337/s41436-018-0039-z.html. pages 21

Carolyn Y. Ho, Begoña López, Otavio R. Coelho-Filho, Neal K. Lakdawala, Allison L. Cirino, Petr Jarolim, Raymond Kwong, Arantxa González, Steven D. Colan, J.G. Seidman, Javier Díez, and Christine E. Seidman. Myocardial fibrosis as an early manifestation of hypertrophic cardiomyopathy. *New England Journal of Medicine*, 363:552–563, 8 2010. ISSN 0028-4793. doi: 10.1056/NEJMOA1002659/SUPPL_FILE/NEJMOA1002659_DISCLOSURES.PDF. URL https://www.nejm.org/doi/full/10.1056/nejmoa1002659. pages 18

Carolyn Y. Ho, Philippe Charron, Pascale Richard, Francesca Girolami, Karin Y. Van Spaendonck-Zwarts, and Yigal Pinto. Genetic advances in sarcomeric cardiomyopathies: state of the art. *Cardiovascular research*, 105:397–408, 4 2015. ISSN 1755-3245. doi: 10.1093/CVR/CVV025. URL https://pubmed.ncbi.nlm.nih.gov/25634555/. pages 6, 8

Carolyn Y. Ho, Sharlene M. Day, Euan A. Ashley, Michelle Michels, Alexandre C. Pereira, Daniel Jacoby, Allison L. Cirino, Jonathan C. Fox, Neal K. Lakdawala, James S. Ware, Colleen A. Caleshu, Adam S. Helms, Steven D. Colan, Francesca Girolami, Franco Cecchi, Christine E. Seidman, Gautam Sajeev, James Signorovitch, Eric M. Green, and Iacopo Olivotto. Genotype and lifetime burden of disease in hypertrophic cardiomyopathy: Insights from the sarcomeric human cardiomyopathy registry (share). *Circulation*, 138:1387, 10 2018. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.117.033200. URL /pmc/articles/PMC6170149//pmc/articles/PMC6170149/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6170149/. pages 5

Thomas Hofmann, Bernhard Schölkopf, and Alexander J Smola. Kernel methods in machine learning. *Annals of Statistics*, 36:1171–1220, 2008. ISSN 0090-5364, 2168-8966. doi: 10.1214/009053607000000677. URL https://projecteuclid.org/euclid.aos/1211819561internal-pdf://349/1211819561.html. pages 32

G. Hommel. A stagewise rejective multiple test procedure based on a modified bonferroni test. *Biometrika*, 75:383–386, 6 1988. ISSN 0006-3444. doi: 10.1093/BIOMET/75.2.383. URL https://dx.doi.org/10.1093/biomet/75.2.383. pages 27, 47

Md Maruf Hossain, Md Shahin Ali, Md Mahfuz Ahmed, Md Rakibul Hasan Rakib, Moutushi Akter Kona, Sadia Afrin, Md Khairul Islam, Md Manjurul Ahsan, Sheikh Md Razibul Hasan Raj, and Md Habibur Rahman. Cardiovascular disease identification using a hybrid cnn-lstm model with explainable ai. *Informatics in Medicine Unlocked*, page 101370, 10 2023. ISSN 2352-9148. doi: 10.1016/J.IMU.2023.101370. URL https://linkinghub.elsevier.com/retrieve/pii/S2352914823002162. pages 104

Cristopher V Van Hout, Ioanna Tachmazidou, Joshua D Backman, Joshua X Hoffman, Bin Ye, Ashutosh K Pandey, Claudia Gonzaga-Jauregui, Shareef Khalid, Daren Liu, Nilanjana Banerjee, Alexander H

Li, O'Dushlaine Colm, Anthony Marcketta, Jeffrey Staples, Claudia Schurmann, Alicia Hawes, Evan Maxwell, Leland Barnard, Alexander Lopez, John Penn, Lukas Habegger, Andrew L Blumenfeld, Ashish Yadav, Kavita Praveen, Marcus Jones, William J Salerno, Wendy K Chung, Ida Surakka, Cristen J Willer, Kristian Hveem, Joseph B Leader, David J Carey, David H Ledbetter, Lon Cardon, George D Yancopoulos, Aris Economides, Giovanni Coppola, Alan R Shuldiner, Suganthi Balasubramanian, Michael Cantor, Matthew R Nelson, John Whittaker, Jeffrey G Reid, Jonathan Marchini, John D Overton, Robert A Scott, Gonçalo Abecasis, Laura Yerges-Armstrong, and Aris Baras. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the uk biobank. *bioRxiv*, page 572347, 2019. doi: 10.1101/572347. URL `http://biorxiv.org/content/early/2019/03/09/572347.abstract`. pages 40

Te-Cheng Hsu and Che Lin. Generative adversarial networks for robust breast cancer prognosis prediction with limited data size. volume 2020-July, pages 5669–5672. IEEE, 7 2020. ISBN 978-1-7281-1990-8. doi: 10.1109/EMBC44109.2020.9175736. URL `https://ieeexplore.ieee.org/document/9175736/`. pages 139

Kishore Jaganathan, Sofia Kyriazopoulou Panagiotopoulou, Jeremy F. McRae, Siavash Fazel Darbandi, David Knowles, Yang I. Li, Jack A. Kosmicki, Juan Arbelaez, Wenwu Cui, Grace B. Schwartz, Eric D. Chow, Efstathios Kanterakis, Hong Gao, Amirali Kia, Serafim Batzoglou, Stephan J. Sanders, and Kyle Kai How Farh. Predicting splicing from primary sequence with deep learning. *Cell*, 176:535–548.e24, 1 2019. ISSN 10974172. doi: 10.1016/j.cell.2018.12.015. URL `http://www.cell.com/article/S0092867418316295/fulltexthttp://www.cell.com/article/S0092867418316295/abstracthttps://www.cell.com/cell/abstract/S0092-8674(18)31629-5`. pages 81

Harshit Jindal, Sarthak Agrawal, Rishabh Khera, al, Daniel F Polan, Samuel L Brady, Robert A Kaufman, Rimpy Morya, Someet Singh, Madhumita Pal, and Smita Parija. Prediction of heart diseases using random forest. *Journal of Physics: Conference Series*,

1817:012009, 3 2021. ISSN 1742-6596. doi: 10.1088/1742-6596/1817/1/012009. URL https://iopscience.iop.org/article/10.1088/1742-6596/1817/1/012009https://iopscience.iop.org/article/10.1088/1742-6596/1817/1/012009/meta. pages 105

Elizabeth Jordan, Laiken Peterson, Tomohiko Ai, Babken Asatryan, Lucas Bronicki, Emily Brown, Rudy Celeghin, Matthew Edwards, Judy Fan, Jodie Ingles, Cynthia A. James, Olga Jarinova, Renee Johnson, Daniel P. Judge, Najim Lahrouchi, Ronald H. Lekanne Deprez, R. Thomas Lumbers, Francesco Mazzarotto, Argelia Medeiros Domingo, Rebecca L. Miller, Ana Morales, Brittney Murray, Stacey Peters, Kalliopi Pilichou, Alexandros Protonotarios, Christopher Semsarian, Palak Shah, Petros Syrris, Courtney Thaxton, J. Peter Van Tintelen, Roddy Walsh, Jessica Wang, James Ware, and Ray E. Hershberger. Evidence-based assessment of genes in dilated cardiomyopathy. *Circulation*, 144:7–19, 7 2021. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.120.053033. URL https://pubmed.ncbi.nlm.nih.gov/33947203/. pages 7

Konrad J. Karczewski, Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, Kristen M. Laricchia, Andrea Ganna, Daniel P. Birnbaum, Laura D. Gauthier, Harrison Brand, Matthew Solomonson, Nicholas A. Watts, Daniel Rhodes, Moriel Singer-Berk, Eleina M. England, Eleanor G. Seaby, Jack A. Kosmicki, Raymond K. Walters, Katherine Tashman, Yossi Farjoun, Eric Banks, Timothy Poterba, Arcturus Wang, Cotton Seed, Nicola Whiffin, Jessica X. Chong, Kaitlin E. Samocha, Emma Pierce-Hoffman, Zachary Zappala, Anne H. O'Donnell-Luria, Eric Vallabh Minikel, Ben Weisburd, Monkol Lek, James S. Ware, Christopher Vittal, Irina M. Armean, Louis Bergelson, Kristian Cibulskis, Kristen M. Connolly, Miguel Covarrubias, Stacey Donnelly, Steven Ferriera, Stacey Gabriel, Jeff Gentry, Namrata Gupta, Thibault Jeandet, Diane Kaplan, Christopher Llanwarne, Ruchi Munshi, Sam Novod, Nikelle Petrillo, David Roazen, Valentin Ruano-Rubio, Andrea Saltzman, Molly Schleicher, Jose Soto, Kathleen Tibbetts, Charlotte Tolonen, Gordon Wade, Michael E. Talkowski, Carlos A. Aguilar Salinas, Tariq Ahmad, Christine M. Albert,

Diego Ardissino, Gil Atzmon, John Barnard, Laurent Beaugerie, Emelia J. Benjamin, Michael Boehnke, Lori L. Bonnycastle, Erwin P. Bottinger, Donald W. Bowden, Matthew J. Bown, John C. Chambers, Juliana C. Chan, Daniel Chasman, Judy Cho, Mina K. Chung, Bruce Cohen, Adolfo Correa, Dana Dabelea, Mark J. Daly, Dawood Darbar, Ravindranath Duggirala, Josée Dupuis, Patrick T. Ellinor, Roberto Elosua, Jeanette Erdmann, Tõnu Esko, Martti Färkkilä, Jose Florez, Andre Franke, Gad Getz, Benjamin Glaser, Stephen J. Glatt, David Goldstein, Clicerio Gonzalez, Leif Groop, Christopher Haiman, Craig Hanis, Matthew Harms, Mikko Hiltunen, Matti M. Holi, Christina M. Hultman, Mikko Kallela, Jaakko Kaprio, Sekar Kathiresan, Bong Jo Kim, Young Jin Kim, George Kirov, Jaspal Kooner, Seppo Koskinen, Harlan M. Krumholz, Subra Kugathasan, Soo Heon Kwak, Markku Laakso, Terho Lehtimäki, Ruth J.F. Loos, Steven A. Lubitz, Ronald C.W. Ma, Daniel G. MacArthur, Jaume Marrugat, Kari M. Mattila, Steven McCarroll, Mark I. McCarthy, Dermot McGovern, Ruth McPherson, James B. Meigs, Olle Melander, Andres Metspalu, Benjamin M. Neale, Peter M. Nilsson, Michael C. O'Donovan, Dost Ongur, Lorena Orozco, Michael J. Owen, Colin N.A. Palmer, Aarno Palotie, Kyong Soo Park, Carlos Pato, Ann E. Pulver, Nazneen Rahman, Anne M. Remes, John D. Rioux, Samuli Ripatti, Dan M. Roden, Danish Saleheen, Veikko Salomaa, Nilesh J. Samani, Jeremiah Scharf, Heribert Schunkert, Moore B. Shoemaker, Pamela Sklar, Hilkka Soininen, Harry Sokol, Tim Spector, Patrick F. Sullivan, Jaana Suvisaari, E. Shyong Tai, Yik Ying Teo, Tuomi Tiinamaija, Ming Tsuang, Dan Turner, Teresa Tusie-Luna, Erkki Vartiainen, Hugh Watkins, Rinse K. Weersma, Maija Wessman, James G. Wilson, and Ramnik J. Xavier. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature 2020 581:7809*, 581:434–443, 5 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2308-7. URL `https://www.nature.com/articles/s41586-020-2308-7`. pages 81

Melpomeni Kasapi, Kexin Xu, Timothy M.D. Ebbels, Declan P. O'Regan, James S. Ware, and Joram M. Posma. Lavaset: Latent variable stochastic ensemble of trees. a novel ensemble method

for correlated datasets. *bioRxiv*, page 2023.10.20.563223, 10 2023. doi: 10.1101/2023.10.20.563223. URL https://www.biorxiv.org/content/10.1101/2023.10.20.563223v1https://www.biorxiv.org/content/10.1101/2023.10.20.563223v1.abstract. pages 58

Mahendra Khened, Varghese Alex, and Ganapathy Krishnamurthi. Densely connected fully convolutional network for short-axis cardiac cine mr image segmentation and heart diagnosis using random forest. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10663 LNCS:140–151, 2018. ISSN 16113349. doi: 10.1007/978-3-319-75541-0_15/FIGURES/4. URL https://link.springer.com/chapter/10.1007/978-3-319-75541-0_15. pages 17

Sun Bin Kim, Joon Ho Kang, and Dong Jun Kim. Snprs:stacked neural network for predicting polygenic risk score. 2023. doi: 10.21203/rs.3.rs-3076268/v1. URL https://doi.org/10.21203/rs.3.rs-3076268/v1. pages 119

Akinori Kimura. Molecular genetics and pathogenesis of cardiomyopathy. *Journal of Human Genetics 2016 61:1*, 61:41–50, 7 2015. ISSN 1435-232X. doi: 10.1038/jhg.2015.83. URL https://www.nature.com/articles/jhg201583. pages 7

Paul Kligfield, Leonard S. Gettes, James J. Bailey, Rory Childers, Barbara J. Deal, E. William Hancock, Gerard Van Herpen, Jan A. Kors, Peter Macfarlane, David M. Mirvis, Olle Pahlm, Pentti Rautaharju, and Galen S. Wagner. Recommendations for the standardization and interpretation of the electrocardiogram. *Circulation*, 115:1306–1324, 3 2007. ISSN 00097322. doi: 10.1161/CIRCULATIONAHA.106.180200. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.106.180200. pages 18

Wei Yin Ko, Konstantinos C. Siontis, Zachi I. Attia, Rickey E. Carter, Suraj Kapa, Steve R. Ommen, Steven J. Demuth, Michael J. Ackerman, Bernard J. Gersh, Adelaide M. Arruda-Olson, Jeffrey B.

Geske, Samuel J. Asirvatham, Francisco Lopez-Jimenez, Rick A. Nishimura, Paul A. Friedman, and Peter A. Noseworthy. Detection of hypertrophic cardiomyopathy using a convolutional neural network-enabled electrocardiogram. *Journal of the American College of Cardiology*, 75:722–733, 2 2020. ISSN 0735-1097. doi: 10.1016/J.JACC.2019.12.030. pages 19

Stephen C. Kolwicz and Rong Tian. Glucose metabolism and cardiac hypertrophy. *Cardiovascular Research*, 90:194, 5 2011. ISSN 00086363. doi: 10.1093/CVR/CVR071. URL /pmc/articles/PMC3078804//pmc/articles/PMC3078804/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3078804/. pages 108

J. A. Kors, G. van Herpen, A. C. Sittig, and J. H. van Bemmel. Reconstruction of the frank vectorcardiogram from standard electrocardiographic leads: diagnostic comparison of different methods. *European Heart Journal*, 11(12):1083–1092, December 1990. doi: 10.1093/oxfordjournals.eurheartj.a059647. URL https://doi.org/10.1093/oxfordjournals.eurheartj.a059647. pages 59

Christopher M Kramer. The role of cmr in cardiomyopathies. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, 56:39S–45S, 2015. ISSN 0161-5505. doi: 10.2967/jnumed.114.142729. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465292/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4465292/. pages 4

William H. Kruskal and W. Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47:583–621, 1952. ISSN 1537274X. doi: 10.1080/01621459.1952.10483441. pages 27

Miron B Kursa and Witold R Rudnicki. Feature selection with the boruta package. *Journal of Statistical Software*, 36:1–13, 2010. doi: 10.18637/jss.v036.i11. URL https://www.jstatsoft.org/index.php/jss/article/view/v036i11. pages 57

Melissa J Landrum, Jennifer M Lee, George R Riley, Wonhee Jang, Wendy S Rubinstein, Deanna M

Church, and Donna R Maglott. Clinvar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42:D980–985, 2014. ISSN 1362-4962. doi: 10.1093/nar/gkt1113. URL http://www.ncbi.nlm.nih.gov/pubmed/24234437. pages 82

Melissa J. Landrum, Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Wonhee Jang, Karen Karapetyan, Kenneth Katz, Chunlei Liu, Zenith Maddipatla, Adriana Malheiro, Kurt McDaniel, Michael Ovetsky, George Riley, George Zhou, J. Bradley Holmes, Brandi L. Kattman, and Donna R. Maglott. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Research*, 46:D1062–D1067, 1 2018. ISSN 0305-1048. doi: 10.1093/NAR/GKX1153. URL https://dx.doi.org/10.1093/nar/gkx1153. pages 82

Cathryn M. Lewis and Evangelos Vassos. Polygenic risk scores: From research tools to clinical instruments. *Genome Medicine*, 12:1–11, 5 2020. ISSN 1756994X. doi: 10.1186/S13073-020-00742-5/TABLES/2. URL https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00742-5. pages 10

Jing Li, Yue yue Feng, Chun ai Hu, Yan Zhao, Ju pan Hou, Hui Xu, Li na Dou, Ming Lou, and Bing Han. Predictive value for mortality of left ventricular wall thickness in dilated cardiomyopathy. *ESC Heart Failure*, 2023. ISSN 2055-5822. doi: 10.1002/EHF2.14534. URL https://onlinelibrary.wiley.com/doi/full/10.1002/ehf2.14534https://onlinelibrary.wiley.com/doi/abs/10.1002/ehf2.14534https://onlinelibrary.wiley.com/doi/10.1002/ehf2.14534. pages 16, 111

Silvia Liverani, David I Hastie, Michail Papathomas, and Sylvia Richardson. Premium: An r package for profile regression mixture models using dirichlet processes. *JSS Journal of Statistical Software*, 64, 2015. URL http://www.jstatsoft.org/. pages 53

Luke R. Lloyd-Jones, Jian Zeng, Julia Sidorenko, Loïc Yengo, Gerhard Moser, Kathryn E. Kemper, Huanwei Wang, Zhili Zheng, Reedik Magi, Tõnu Esko, Andres Metspalu, Naomi R. Wray, Michael E. Goddard, Jian Yang, and Peter M. Visscher. Improved polygenic prediction by bayesian multiple regression on summary statistics. *Nature Communications 2019 10:1*, 10:1–11, 11 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12653-0. URL https://www.nature.com/articles/s41467-019-12653-0. pages 11

Massimiliano Lorenzini, Gabrielle Norrish, Ella Field, Juan Pablo Ochoa, Marcos Cicerchia, Mohammed M. Akhtar, Petros Syrris, Luis R. Lopes, Juan Pablo Kaski, and Perry M. Elliott. Penetrance of hypertrophic cardiomyopathy in sarcomere protein mutation carriers. *Journal of the American College of Cardiology*, 76:550–559, 8 2020. ISSN 15583597. doi: 10.1016/J.JACC.2020.06.011/SUPPL_FILE/MMC1.DOCX. URL https://www.jacc.org/doi/10.1016/j.jacc.2020.06.011. pages 9, 18

R. Thomas Lumbers, Sonia Shah, Honghuang Lin, Tomasz Czuba, Albert Henry, Daniel I. Swerdlow, Anders Mälarstig, Charlotte Andersson, Niek Verweij, Michael V. Holmes, Johan Ärnlöv, Per Svensson, Harry Hemingway, Neneh Sallah, Peter Almgren, Krishna G. Aragam, Geraldine Asselin, Joshua D. Backman, Mary L. Biggs, Heather L. Bloom, Eric Boersma, Jeffrey Brandimarto, Michael R. Brown, Hans Peter Brunner-La Rocca, David J. Carey, Mark D. Chaffin, Daniel I. Chasman, Olympe Chazara, Xing Chen, Xu Chen, Jonathan H. Chung, William Chutkow, John G.F. Cleland, James P. Cook, Simon de Denus, Abbas Dehghan, Graciela E. Delgado, Spiros Denaxas, Alexander S. Doney, Marcus Dörr, Samuel C. Dudley, Gunnar Engström, Tõnu Esko, Ghazaleh Fatemifar, Stephan B. Felix, Chris Finan, Ian Ford, Francoise Fougerousse, René Fouodjio, Mohsen Ghanbari, Sahar Ghasemi, Vilmantas Giedraitis, Franco Giulianini, John S. Gottdiener, Stefan Gross, Daníel F. Gubjartsson, Hongsheng Gui, Rebecca Gutmann, Christopher M. Haggerty, Pim van der Harst, Åsa K. Hedman, Anna Helgadottir, Hans Hillege, Craig L. Hyde, Jaison Jacob, J. Wouter Jukema,

Frederick Kamanu, Isabella Kardys, Maryam Kavousi, Kay Tee Khaw, Marcus E. Kleber, Lars Køber, Andrea Koekemoer, Bill Kraus, Karoline Kuchenbaecker, Claudia Langenberg, Lars Lind, Cecilia M. Lindgren, Barry London, Luca A. Lotta, Ruth C. Lovering, Jian'an Luan, Patrik Magnusson, Anubha Mahajan, Douglas Mann, Kenneth B. Margulies, Nicholas A. Marston, Winfried März, John J.V. McMurray, Olle Melander, Giorgio Melloni, Ify R. Mordi, Michael P. Morley, Andrew D. Morris, Andrew P. Morris, Alanna C. Morrison, Michael W. Nagle, Christopher P. Nelson, Christopher Newton-Cheh, Alexander Niessner, Teemu Niiranen, Christoph Nowak, Michelle L. O'Donoghue, Anjali T. Owens, Colin N.A. Palmer, Guillaume Paré, Markus Perola, Louis Philippe Lemieux Perreault, Eliana Portilla-Fernandez, Bruce M. Psaty, Kenneth M. Rice, Paul M. Ridker, Simon P.R. Romaine, Carolina Roselli, Jerome I. Rotter, Christian T. Ruff, Marc S. Sabatine, Perttu Salo, Veikko Salomaa, Jessica van Setten, Alaa A. Shalaby, Diane T. Smelser, Nicholas L. Smith, Kari Stefansson, Steen Stender, David J. Stott, Garar Sveinbjörnsson, Mari Liis Tammesoo, Jean Claude Tardif, Kent D. Taylor, Maris Teder-Laving, Alexander Teumer, Gumundur Thorgeirsson, Unnur Thorsteinsdottir, Christian Torp-Pedersen, Stella Trompet, Danny Tuckwell, Benoit Tyl, Andre G. Uitterlinden, Felix Vaura, Abirami Veluchamy, Peter M. Visscher, Uwe Völker, Adriaan A. Voors, Xiaosong Wang, Nicholas J. Wareham, Peter E. Weeke, Raul Weiss, Harvey D. White, Kerri L. Wiggins, Heming Xing, Jian Yang, Yifan Yang, Laura M. Yerges-Armstrong, Bing Yu, Faiez Zannad, Faye Zhao, Jemma B. Wilk, Hilma Holm, Naveed Sattar, Steven A. Lubitz, David E. Lanfear, Svati Shah, Michael E. Dunn, Quinn S. Wells, Folkert W. Asselbergs, Aroon D. Hingorani, Marie Pierre Dubé, Nilesh J. Samani, Chim C. Lang, Thomas P. Cappola, Patrick T. Ellinor, Ramachandran S. Vasan, and J. Gustav Smith. The genomics of heart failure: design and rationale of the hermes consortium. *ESC heart failure*, 8:5531–5541, 12 2021. ISSN 2055-5822. doi: 10.1002/EHF2.13517. URL `https://pubmed.ncbi.nlm.nih.gov/34480422/`. pages 12, 44, 117

Aurore Lyon, Rina Ariga, Ana Mincholé, Masliza Mahmod, Elizabeth Ormondroyd, Pablo Laguna,

Nando de Freitas, Stefan Neubauer, Hugh Watkins, and Blanca Rodriguez. Distinct ecg phenotypes identified in hypertrophic cardiomyopathy using machine learning associate with arrhythmic risk markers. *Frontiers in Physiology*, 9:328675, 3 2018. ISSN 1664042X. doi: 10.3389/FPHYS.2018.00213/BIBTEX. pages 19

Daniel Mak, Kaitlin A. Ryan, and Joan C. Han. Review of insulin resistance in dilated cardiomyopathy and implications for the pediatric patient short title: Insulin resistance dcm and pediatrics. *Frontiers in Pediatrics*, 9:756593, 10 2021. ISSN 22962360. doi: 10.3389/FPED.2021.756593/BIBTEX. pages 109

Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic Epidemiology*, 41:469–480, 9 2017. ISSN 1098-2272. doi: 10.1002/GEPI.22050. URL https://onlinelibrary.wiley.com/doi/full/10.1002/gepi.22050https://onlinelibrary.wiley.com/doi/abs/10.1002/gepi.22050https://onlinelibrary.wiley.com/doi/10.1002/gepi.22050. pages 115

H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *https://doi.org/10.1214/aoms/1177730491*, 18: 50–60, 3 1947. ISSN 0003-4851. doi: 10.1214/AOMS/1177730491. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-18/issue-1/On-a-Test-of-Whether-one-of-Two-Random-Variables/10.1214/aoms/1177730491.fullhttps://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-18/issue-1/On-a-Test-of-Whether-one-of-Two-Random-Variables/10.1214/aoms/1177730491.short. pages 26

Qi Mao, Li Wang, Ivor W. Tsang, and Yijun Sun. Principal graph and structure learning based on reversed graph embedding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39: 2227–2241, 11 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2635657. pages 53

Ali J. Marian and Eugene Braunwald. Hypertrophic cardiomyopathy. *Circulation Research*, 121: 749–770, 9 2017. ISSN 15244571. doi: 10.1161/CIRCRESAHA.117.311059. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCRESAHA.117.311059. pages 103, 108

Barry J. Maron, Julius M. Gardin, John M. Flack, Samuel S. Gidding, Tom T. Kurosaki, and Diane E. Bild. Prevalence of hypertrophic cardiomyopathy in a general population of young adults. echocardiographic analysis of 4111 subjects in the cardia study. coronary artery risk development in (young) adults. *Circulation*, 92:785–789, 8 1995. ISSN 0009-7322. doi: 10.1161/01.CIR.92.4.785. URL https://pubmed.ncbi.nlm.nih.gov/7641357/. pages 5, 107

Barry J. Maron, Iacopo Olivotto, Paolo Spirito, Susan A. Casey, Pietro Bellone, Thomas E. Gohman, Kevin J. Graham, David A. Burton, and Franco Cecchi. Epidemiology of hypertrophic cardiomyopathy-related death: revisited in a large non-referral-based patient population. *Circulation*, 102:858–864, 8 2000. ISSN 1524-4539. doi: 10.1161/01.CIR.102.8.858. URL https://pubmed.ncbi.nlm.nih.gov/10952953/https://pubmed.ncbi.nlm.nih.gov/10952953/?dopt=Abstract. pages 5

Antonio De Marvao, Timothy J.W. Dawes, Wenzhe Shi, Giuliana Durighel, Daniel Rueckert, Stuart A. Cook, and Declan P. O'Regan. Precursors of hypertensive heart phenotype develop in healthy adults: A high-resolution 3d mri study. *JACC: Cardiovascular Imaging*, 8:1260–1269, 11 2015. ISSN 1936-878X. doi: 10.1016/J.JCMG.2015.08.007. pages 14, 41

Francesco Mazzarotto, Upasana Tayal, Rachel J. Buchan, William Midwinter, Alicja Wilk, Nicola Whiffin, Risha Govind, Erica Mazaika, Antonio De Marvao, Timothy J.W. Dawes, Leanne E. Felkin, Mian Ahmad, Pantazis I. Theotokis, Elizabeth Edwards, Alexander Y. Ing, Kate L. Thomson, Laura L.H. Chan, David Sim, A. John Baksi, Antonis Pantazis, Angharad M. Roberts, Hugh Watkins, Birgit Funke, Declan P. O'regan, Iacopo Olivotto, Paul J.R. Barton, Sanjay K. Prasad, Stuart A. Cook, James S. Ware, and Roddy Walsh. Reevaluating the genetic contribution of

monogenic dilated cardiomyopathy. *Circulation*, 141:387–398, 2 2020. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.119.037661. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.119.037661. pages 103

Kathryn A. McGurk, Melpomeni Kasapi, and James S. Ware. Effect of taurine administration on symptoms, severity, or clinical outcome of dilated cardiomyopathy and heart failure in humans: a systematic review. *Wellcome Open Research*, 7:9, 7 2022. doi: 10.12688/WELLCOMEOPENRES.17505.3. pages 12

Kathryn A. McGurk, Xiaolei Zhang, Pantazis Theotokis, Kate Thomson, Andrew Harper, Rachel J. Buchan, Erica Mazaika, Elizabeth Ormondroyd, William T. Wright, Daniela Macaya, Chee Jian Pua, Birgit Funke, Daniel G. MacArthur, Sanjay K. Prasad, Stuart A. Cook, Mona Allouba, Yasmine Aguib, Magdi H. Yacoub, Declan P. O'Regan, Paul J.R. Barton, Hugh Watkins, Leonardo Bottolo, and James S. Ware. The penetrance of rare variants in cardiomyopathy-associated genes: A cross-sectional approach to estimating penetrance for secondary findings. *American Journal of Human Genetics*, 110:1482–1495, 9 2023. ISSN 15376605. doi: 10.1016/j.ajhg.2023.08.003. URL http://www.cell.com/article/S0002929723002793/fulltexthttp://www.cell.com/article/S0002929723002793/abstracthttps://www.cell.com/ajhg/abstract/S0002-9297(23)00279-3. pages 9

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861. pages 64

Leland Mcinnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020. pages 28

William J. McKenna and Daniel P. Judge. Epidemiology of the inherited cardiomyopathies. *Nature*

reviews. Cardiology, 18:22–36, 1 2021. ISSN 1759-5010. doi: 10.1038/S41569-020-0428-2. URL https://pubmed.ncbi.nlm.nih.gov/32895535/. pages 4, 5

William McLaren, Laurent Gil, Sarah E. Hunt, Harpreet Singh Riat, Graham R.S. Ritchie, Anja Thormann, Paul Flicek, and Fiona Cunningham. The ensembl variant effect predictor. *Genome Biology*, 17: 1–14, 6 2016. ISSN 1474760X. doi: 10.1186/S13059-016-0974-4/TABLES/8. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4. pages 81

Hannah V. Meyer, Timothy J.W. Dawes, Marta Serrani, Wenjia Bai, Paweł Tokarczuk, Jiashen Cai, Antonio de Marvao, Albert Henry, R. Thomas Lumbers, Jakob Gierten, Thomas Thumberger, Joachim Wittbrodt, James S. Ware, Daniel Rueckert, Paul M. Matthews, Sanjay K. Prasad, Maria L. Costantino, Stuart A. Cook, Ewan Birney, and Declan P. O'Regan. Genetic and functional insights into the fractal structure of the heart. *Nature 2020 584:7822*, 584:589–594, 8 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2635-8. URL https://www.nature.com/articles/s41586-020-2635-8. pages 16

Manish Motwani, Roy Jogiya, Sebastian Kozerke, John P. Greenwood, and Sven Plein. Advanced cardiovascular magnetic resonance myocardial perfusion imaging high-spatial resolution versus 3-dimensional whole-heart coverage. *Circulation: Cardiovascular Imaging*, 6:339–348, 3 2013. ISSN 19419651. doi: 10.1161/CIRCIMAGING.112.000193. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCIMAGING.112.000193. pages 54

Zlatan Mujagic, Melpomeni Kasapi, Daisy MAE Jonkers, Isabel Garcia-Perez, Lisa Vork, Zsa Zsa R.M. Weerts, Jose Ivan Serrano-Contreras, Alexandra Zhernakova, Alexander Kurilshikov, Jamie Scotcher, Elaine Holmes, Cisca Wijmenga, Daniel Keszthelyi, Jeremy K Nicholson, Joram M Posma, and Ad AM Masclee. Integrated fecal microbiome–metabolome signatures reflect stress and serotonin metabolism in irritable bowel syndrome. *Gut Microbes*, 14(1):2063016, 2022. doi: 10.1080/19490976.2022.2063016. URL https://doi.org/10.1080/19490976.2022.2063016.

PMID: 35446234. pages 53

Kazuo Murakami, Yuji Shigematsu, Mareomi Hamada, and Jitsuo Higaki. Insulin resistance in patients with hypertrophic cardiomyopthy. *Circulation Journal*, 68:650–655, 7 2004. ISSN 1346-9843. doi: 10.1253/CIRCJ.68.650. pages 108

Vinod Nair and Geoffrey E Hinton. Rectified linear units improve restricted boltzmann machines. pages

Stefan Neubauer, Paul Kolm, Carolyn Y. Ho, Raymond Y. Kwong, Milind Y. Desai, Sarahfaye F. Dolman, Evan Appelbaum, Patrice Desvigne-Nickens, John P. DiMarco, Matthias G. Friedrich, Nancy Geller, Andrew R. Harper, Petr Jarolim, Michael Jerosch-Herold, Dong Yun Kim, Martin S. Maron, Jeanette Schulz-Menger, Stefan K. Piechnik, Kate Thomson, Cheng Zhang, Hugh Watkins, William S. Weintraub, and Christopher M. Kramer. Distinct subgroups in hypertrophic cardiomyopathy in the nhlbi hcm registry. *Journal of the American College of Cardiology*, 74:2333–2345, 11 2019. ISSN 0735-1097. doi: 10.1016/J.JACC.2019.08.1057. pages 8

Sarah B Ng, Emily H Turner, Peggy D Robertson, Steven D Flygare, Abigail W Bigham, Choli Lee, Tristan Shaffer, Michelle Wong, Arindam Bhattacharjee, Evan E Eichler, Michael Bamshad, Deborah A Nickerson, and Jay Shendure. Targeted capture and massively parallel sequencing of twelve human exomes. *Nature*, 461:272–276, 2009. ISSN 0028-0836. doi: 10.1038/nature08250. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2844771/https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2844771/`. pages 21

Jean-Michel Nguyen, Pascal Jézéquel, Pierre Gillois, Luisa Silva, Faouda Ben Azzouz, Sophie Lambert-Lacroix, Philippe Juin, Mario Campone, Aurélie Gaultier, Alexandre Moreau-Gaudry, and Daniel Antonioli. Random forest of perfect trees: concept, performance, applications and perspectives. *Bioinformatics*, 37:2165–2174, 8 2021. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTAB074. URL `https://academic.oup.com/bioinformatics/article/37/15/2165/6125385`.

pages 22, 57

Kristin K. Nicodemus, James D. Malley, Carolin Strobl, and Andreas Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11:1–13, 2 2010. ISSN 14712105. doi: 10.1186/1471-2105-11-110/FIGURES/6. URL https://link.springer.com/articles/10.1186/1471-2105-11-110https://link.springer.com/article/10.1186/1471-2105-11-110. pages 57

Jens Nußberger, Frederic Boesel, Stefan Lenz, Harald Binder, and Moritz Hess. Synthetic observations from deep generative models and binary omics data with limited sample size. *Briefings in Bioinformatics*, 22:1–12, 7 2021. ISSN 14774054. doi: 10.1093/BIB/BBAA226. URL https://dx.doi.org/10.1093/bib/bbaa226. pages 139

Monica Patten, Simon Pecha, and Ali Aydin. Atrial fibrillation in hypertrophic cardiomyopathy: Diagnosis and considerations for management. *Journal of Atrial Fibrillation*, 10, 2 2018. ISSN 19416911. doi: 10.4022/JAFIB.1556. URL /pmc/articles/PMC6006972//pmc/articles/PMC6006972/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6006972/. pages 107

Steffen E. Petersen, Bjarke Jensen, Nay Aung, Matthias G. Friedrich, Colin J. McMahon, Saidi A. Mohiddin, Ricardo H. Pignatelli, Fabrizio Ricci, Robert H. Anderson, and David A. Bluemke. Excessive trabeculation of the left ventricle: Jacc: Cardiovascular imaging expert panel paper. *Jacc. Cardiovascular Imaging*, 16:408, 3 2023. ISSN 18767591. doi: 10.1016/J.JCMG.2022.12.026. URL /pmc/articles/PMC9988693//pmc/articles/PMC9988693/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9988693/. pages 16

J. Pfanzagl and O. Sheynin. Studies in the history of probability and statistics xliv a forerunner of the t-distribution. *Biometrika*, 83:891–898, 12 1996. ISSN 0006-3444. doi: 10.1093/BIOMET/83.4.891. URL https://dx.doi.org/10.1093/biomet/83.4.891. pages 27

Yigal M. Pinto, Perry M. Elliott, Eloisa Arbustini, Yehuda Adler, Aris Anastasakis, Michael Böhm, Denis Duboc, Juan Gimeno, Pascal De Groote, Massimo Imazio, Stephane Heymans, Karin Klingel, Michel Komajda, Giuseppe Limongelli, Ales Linhart, Jens Mogensen, James Moon, Petronella G. Pieper, Petar M. Seferovic, Stephan Schueler, Jose L. Zamorano, Alida L.P. Caforio, and Philippe Charron. Proposal for a revised definition of dilated cardiomyopathy, hypokinetic non-dilated cardiomyopathy, and its implications for clinical practice: a position statement of the esc working group on myocardial and pericardial diseases. *European Heart Journal*, 37:1850–1858, 6 2016. ISSN 0195-668X. doi: 10.1093/EURHEARTJ/EHV727. URL https://dx.doi.org/10.1093/eurheartj/ehv727. pages 137

James P. Pirruccello, Alexander Bick, Minxian Wang, Mark Chaffin, Samuel Friedman, Jie Yao, Xiuqing Guo, Bharath Ambale Venkatesh, Kent D. Taylor, Wendy S. Post, Stephen Rich, Joao A.C. Lima, Jerome I. Rotter, Anthony Philippakis, Steven A. Lubitz, Patrick T. Ellinor, Amit V. Khera, Sekar Kathiresan, and Krishna G. Aragam. Analysis of cardiac magnetic resonance imaging in 36,000 individuals yields genetic insights into dilated cardiomyopathy. *Nature Communications 2020 11:1*, 11:1–10, 5 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15823-7. URL https://www.nature.com/articles/s41467-020-15823-7. pages 12, 115, 134

D M W Powers and Ailab. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. 10 2020. URL https://arxiv.org/abs/2010.16061v1. pages 34, 35

Florian Privé, Bjarni J Vilhjálmsson, Hugues Aschard, and Michael G B Blum. Making the most of clumping and thresholding for polygenic scores. doi: 10.1016/j.ajhg.2019.11.001. URL https://doi.org/10.1016/j.ajhg.2019.11.001. pages 52

Esther Puyol-Anton, Bram Ruijsink, Wenjia Bai, Helene Langet, Mathieu De Craene, Julia A. Schnabel, Paolo Piro, Andrew P. King, and Matthew Sinclair. Fully automated myocardial strain estimation

from cine mri using convolutional neural networks. *Proceedings - International Symposium on Biomedical Imaging*, 2018-April:1139–1143, 5 2018. ISSN 19458452. doi: 10.1109/ISBI.2018.8363772. pages 15

Adityanarayanan Radhakrishnan, Sam F. Friedman, Shaan Khurshid, Kenney Ng, Puneet Batra, Steven A. Lubitz, Anthony A. Philippakis, and Caroline Uhler. Cross-modal autoencoder framework learns holistic representations of cardiovascular state. *Nature Communications 2023 14:1*, 14:1–12, 4 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38125-0. URL https://www.nature.com/articles/s41467-023-38125-0. pages 72

Zahra Raisi-Estabragh, Nicholas C. Harvey, Stefan Neubauer, and Steffen E. Petersen. Cardiovascular magnetic resonance imaging in the uk biobank: a major international health research resource. *European Heart Journal - Cardiovascular Imaging*, 22:251–258, 2 2021. ISSN 2047-2404. doi: 10.1093/EHJCI/JEAA297. URL https://dx.doi.org/10.1093/ehjci/jeaa297. pages 16, 40

Jay Ramchand, Mathew Wallis, Ivan Macciocca, Elly Lynch, Omar Farouque, Melissa Martyn, Dean Phelan, Belinda Chong, Siobhan Lockwood, Robert Weintraub, Tina Thompson, Alison Trainer, Dominica Zentner, Jitendra Vohra, Michael Chetrit, David L. Hare, and Paul James. Prospective evaluation of the utility of whole exome sequencing in dilated cardiomyopathy. *Journal of the American Heart Association: Cardiovascular and Cerebrovascular Disease*, 9, 1 2020. ISSN 20479980. doi: 10.1161/JAHA.119.013346. URL /pmc/articles/PMC7033851//pmc/articles/PMC7033851/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7033851/. pages 22

Margaret M. Redfield, Steven J. Jacobsen, John C. Burnett, Douglas W. Mahoney, Kent R. Bailey, and Richard J. Rodeheffer. Burden of systolic and diastolic ventricular dysfunction in the community: appreciating the scope of the heart failure epidemic. *JAMA*, 289:194–202, 1 2003. ISSN 0098-7484. doi: 10.1001/JAMA.289.2.194. URL https://pubmed.ncbi.nlm.nih.gov/12517230/https://pubmed.ncbi.nlm.nih.gov/12517230/?dopt=Abstract. pages 5

Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, Karl Voelkerding, and Heidi L. Rehm. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the american college of medical genetics and genomics and the association for molecular pathology. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17:405, 5 2015. ISSN 15300366. doi: 10.1038/GIM.2015.30. URL /pmc/articles/PMC4544753//pmc/articles/PMC4544753/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4544753/. pages 21

Greg Ridgeway. Generalized boosted models: A guide to the gbm package. 2020. pages 30, 31

Josefina Robertson, Martin Lindgren, Maria Schaufelberger, Martin Adiels, Lena Björck, Christina E. Lundberg, Naveed Sattar, Annika Rosengren, and Maria Åberg. Body mass index in young women and risk of cardiomyopathy: A long-term follow-up study in sweden. *Circulation*, 141: 520–529, 2 2020. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.119.044056. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.119.044056. pages 110

Ethan J. Rowin, Barry J. Maron, Richard T. Carrick, Parth P. Patel, Ben Koethe, Sophie Wells, and Martin S. Maron. Outcomes in patients with hypertrophic cardiomyopathy and left ventricular systolic dysfunction. *Journal of the American College of Cardiology*, 75:3033–3043, 6 2020. ISSN 15583597. doi: 10.1016/J.JACC.2020.04.045. URL https://www.jacc.org/doi/10.1016/j.jacc.2020.04.045. pages 135

D. Rueckert, LI Sonoda, C. Hayes, DL Hill, MO Leach, and DJ Hawkes. Nonrigid registration using free-form deformations: application to breast mr images. *IEEE transactions on medical imaging*, 18: 712–721, 1999. ISSN 0278-0062. doi: 10.1109/42.796284. URL https://pubmed.ncbi.nlm.nih.gov/10534053/https://pubmed.ncbi.nlm.nih.gov/10534053/?dopt=Abstract. pages 15

M Sabater-Molina, I Pérez-Sánchez, J P Hernández del Rincón, and J R Gimeno. Genetics of hypertrophic cardiomyopathy: A review of current state. *Clinical Genetics*, 93:3–14, 2018. ISSN 1399-0004. doi: 10.1111/cge.13027. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cge.13027internal-pdf://319/cge.html. pages 2, 6

Arunashis Sau, Antonio H. Ribeiro, Kathryn A. McGurk, Libor Pastika, Nikesh Bajaj, Maddalena Ardissino, Jun Yu Chen, Huiyi Wu, Xili Shi, Katerina Hnatkova, Sean Zheng, Annie Britton, Martin Shipley, Irena Andršová, Tomáš Novotný, Ester Sabino, Luana Giatti, Sandhi M Barreto, Jonathan W. Waks, Daniel B. Kramer, Danilo Mandic, Nicholas S. Peters, Declan P. O'Regan, Marek Malik, James S. Ware, Antonio Luiz P. Ribeiro, and Fu Siong Ng. Neural network-derived electrocardiographic features have prognostic significance and important phenotypic and genotypic associations. *medRxiv*, page 2023.06.15.23291428, 6 2023. doi: 10.1101/2023.06.15.23291428. URL https://www.medrxiv.org/content/10.1101/2023.06.15.23291428v1https://www.medrxiv.org/content/10.1101/2023.06.15.23291428v1.abstract. pages 20

Tabea Schoeler, Doug Speed, Eleonora Porcu, Nicola Pirastu, Jean Baptiste Pingault, and Zoltán Kutalik. Participation bias in the uk biobank distorts genetic associations and downstream analyses. *Nature Human Behaviour 2023 7:7*, 7:1216–1227, 4 2023. ISSN 2397-3374. doi: 10.1038/s41562-023-01579-9. URL https://www.nature.com/articles/s41562-023-01579-9. pages 109

Heinz Peter Schultheiss, De Lisa Fairweather, Alida L.P. Caforio, Felicitas Escher, Ray E. Hershberger, Steven E. Lipshultz, Peter P. Liu, Akira Matsumori, Andrea Mazzanti, John McMurray, and Silvia G. Priori. Dilated cardiomyopathy. *Nature Reviews Disease Primers 2019 5:1*, 5:1–19, 5 2019. ISSN 2056-676X. doi: 10.1038/s41572-019-0084-1. URL https://www.nature.com/articles/s41572-019-0084-1. pages 2

Eleanor G. Seaby, Reuben J. Pengelly, and Sarah Ennis. Exome sequencing explained: a practical guide

to its clinical application. *Briefings in Functional Genomics*, 15:374–384, 9 2016. ISSN 2041-2649. doi: 10.1093/BFGP/ELV054. URL https://dx.doi.org/10.1093/bfgp/elv054. pages 21

Sara B. Seidelmann, Olga Laur, John Hwa, Eugene Depasquale, Lavanya Bellumkonda, Lissa Sugeng, Pawel Pomianowski, Jeffrey Testani, Michael Chen, William McKenna, and Daniel Jacoby. Familial dilated cardiomyopathy diagnosis is commonly overlooked at the time of transplant listing. *The Journal of heart and lung transplantation : the official publication of the International Society for Heart Transplantation*, 35:474–480, 4 2016. ISSN 1557-3117. doi: 10.1016/J.HEALUN.2015.12.002. URL https://pubmed.ncbi.nlm.nih.gov/26852066/https://pubmed.ncbi.nlm.nih.gov/26852066/?dopt=Abstract. pages 5

Devansh Shah, Samir Patel, and Santosh Kumar Bharti. Heart disease prediction using machine learning techniques. *SN Computer Science*, 1:1–6, 11 2020. ISSN 26618907. doi: 10.1007/S42979-020-00365-Y/METRICS. URL https://link.springer.com/article/10.1007/s42979-020-00365-y. pages 105

Ravi A. Shah, Babken Asatryan, Ghaith Sharaf Dabbagh, Nay Aung, Mohammed Y. Khanji, Luis R. Lopes, Stefan Van Duijvenboden, Anthony Holmes, Daniele Muser, Andrew P. Landstrom, Aaron Mark Lee, Pankaj Arora, Christopher Semsarian, Virend K. Somers, Anjali T. Owens, Patricia B. Munroe, Steffen E. Petersen, and C. Anwar A. Chahal. Frequency, penetrance, and variable expressivity of dilated cardiomyopathy-associated putative pathogenic gene variants in uk biobank participants. *Circulation*, 146:110–124, 7 2022. ISSN 15244539. doi: 10.1161/CIRCULATIONAHA.121.058143. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCULATIONAHA.121.058143. pages 19

R. J. Simes. An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73:751–754, 12 1986. ISSN 0006-3444. doi: 10.1093/BIOMET/73.3.751. URL https://dx.doi.org/10.1093/biomet/73.3.751. pages 47

GCM Siontis, I Tzoulaki, KC Siontis, JPA Ioannidis Bmj, and undefined 2012. Comparisons of established risk prediction models for cardiovascular disease: systematic review. *bmj.comGCM Siontis, I Tzoulaki, KC Siontis, JPA IoannidisBmj, 2012•bmj.com*, 344, 6 2012. doi: 10.1136/bmj.e3318. URL `https://www.bmj.com/content/344/bmj.e3318.abstract`. pages 75

Konstantinos C. Siontis, Steve R. Ommen, and Jeffrey B. Geske. Sex, survival, and cardiomyopathy: Differences between men and women with hypertrophic cardiomyopathy. *Journal of the American Heart Association*, 8:14448, 11 2019. ISSN 20479980. doi: 10.1161/JAHA.119.014448. URL `https://www.ahajournals.org/doi/abs/10.1161/JAHA.119.014448`. pages 107

Konstantinos C. Siontis, Sean Abreau, Zachi I. Attia, Joshua P. Barrios, Thomas A. Dewland, Priyanka Agarwal, Aarthi Balasubramanyam, Yunfan Li, Steven J. Lester, Ahmad Masri, Andrew Wang, Amy J. Sehnert, Jay M. Edelberg, Theodore P. Abraham, Paul A. Friedman, Jeffrey E. Olgin, Peter A. Noseworthy, and Geoffrey H. Tison. Patient-level artificial intelligence–enhanced electrocardiography in hypertrophic cardiomyopathy: Longitudinal treatment and clinical biomarker correlations. *JACC: Advances*, 2:100582, 10 2023. ISSN 2772-963X. doi: 10.1016/J.JACADV.2023.100582. pages 19

Gerard Snaauw, Dong Gong, Gabriel Maicas, Anton Van Den Hengel, Wiro J. Niessen, Johan Verjans, and Gustavo Carneiro. End-to-end diagnosis and segmentation learning from cardiac magnetic resonance imaging. *Proceedings - International Symposium on Biomedical Imaging*, 2019-April: 802–805, 4 2019. ISSN 19458452. doi: 10.1109/ISBI.2019.8759276. pages 17

Natasha T. Strande, Erin Rooney Riggs, Adam H. Buchanan, Ozge Ceyhan-Birsoy, Marina DiStefano, Selina S. Dwight, Jenny Goldstein, Rajarshi Ghosh, Bryce A. Seifert, Tam P. Sneddon, Matt W. Wright, Laura V. Milko, J. Michael Cherry, Monica A. Giovanni, Michael F. Murray, Julianne M. O'Daniel, Erin M. Ramos, Avni B. Santani, Alan F. Scott, Sharon E. Plon, Heidi L. Rehm, Christa L. Martin, and Jonathan S. Berg. Evaluating the clinical validity of gene-disease associations: An

evidence-based framework developed by the clinical genome resource. *American journal of human genetics*, 100:895–906, 6 2017. ISSN 1537-6605. doi: 10.1016/J.AJHG.2017.04.015. URL https://pubmed.ncbi.nlm.nih.gov/28552198/. pages 6

Daniel Svozil, Vladimír Kvasnička, and Jiří Pospíchal. Introduction to multi-layer feed-forward neural networks. *Chemometrics and Intelligent Laboratory Systems*, 39:43–62, 11 1997. ISSN 0169-7439. doi: 10.1016/S0169-7439(97)00061-0. pages 33

Rafik Tadros, Catherine Francis, Xiao Xu, Alexa M.C. Vermeer, Andrew R. Harper, Roy Huurman, Ken Kelu Bisabu, Roddy Walsh, Edgar T. Hoorntje, Wouter P. te Rijdt, Rachel J. Buchan, Hannah G. van Velzen, Marjon A. van Slegtenhorst, Jentien M. Vermeulen, Joost Allard Offerhaus, Wenjia Bai, Antonio de Marvao, Najim Lahrouchi, Leander Beekman, Jacco C. Karper, Jan H. Veldink, Elham Kayvanpour, Antonis Pantazis, A. John Baksi, Nicola Whiffin, Francesco Mazzarotto, Geraldine Sloane, Hideaki Suzuki, Deborah Schneider-Luftman, Paul Elliott, Pascale Richard, Flavie Ader, Eric Villard, Peter Lichtner, Thomas Meitinger, Michael W.T. Tanck, J. Peter van Tintelen, Andrew Thain, David McCarty, Robert A. Hegele, Jason D. Roberts, Julie Amyot, Marie Pierre Dubé, Julia Cadrin-Tourigny, Geneviève Giraldeau, Philippe L. L'Allier, Patrick Garceau, Jean Claude Tardif, S. Matthijs Boekholdt, R. Thomas Lumbers, Folkert W. Asselbergs, Paul J.R. Barton, Stuart A. Cook, Sanjay K. Prasad, Declan P. O'Regan, Jolanda van der Velden, Karin J.H. Verweij, Mario Talajic, Guillaume Lettre, Yigal M. Pinto, Benjamin Meder, Philippe Charron, Rudolf A. de Boer, Imke Christiaans, Michelle Michels, Arthur A.M. Wilde, Hugh Watkins, Paul M. Matthews, James S. Ware, and Connie R. Bezzina. Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. *Nature genetics*, 53:128–134, 2 2021. ISSN 1546-1718. doi: 10.1038/S41588-020-00762-2. URL https://pubmed.ncbi.nlm.nih.gov/33495596/. pages 12, 115, 134

Rafik Tadros, Sean L Zheng, Christopher Grace, Paloma Jordà, Catherine Francis, Sean J Jurgens,

Kate L Thomson, Andrew R Harper, Elizabeth Ormondroyd, Dominique M West, Xiao Xu, Pantazis I Theotokis, Rachel J Buchan, Kathryn A McGurk, Francesco Mazzarotto, Beatrice Boschi, Elisabetta Pelo, Michael Lee, Michela Noseda, Amanda Varnava, Alexa MC Vermeer, Roddy Walsh, Ahmad S Amin, Marjon A van Slegtenhorst, Nicole Roslin, Lisa J Strug, Erika Salvi, Chiara Lanzani, Antonio de Marvao, Hypergenes InterOmics Collaborators, Jason D Roberts, Maxime Tremblay-Gravel, Genevieve Giraldeau, Julia Cadrin-Tourigny, Philippe L L'Allier, Patrick Garceau, Mario Talajic, Yigal M Pinto, Harry Rakowski, Antonis Pantazis, John Baksi, Brian P Halliday, Sanjay K Prasad, Paul JR Barton, Declan P O'Regan, Stuart A Cook, Rudolf A de Boer, Imke Christiaans, Michelle Michels, Christopher M Kramer, Carolyn Y Ho, Stefan Neubauer, HCMR Investigators, Paul M Matthews, Arthur A Wilde, Jean-Claude Tardif, Iacopo Olivotto, Arnon Adler, Anuj Goel, James S Ware, Connie R Bezzina, and Hugh Watkins. Large scale genome-wide association analyses identify novel genetic loci and mechanisms in hypertrophic cardiomyopathy. *medRxiv*, 27:2023.01.28.23285147, 2 2023. doi: 10.1101/2023.01.28.23285147. URL `https://www.medrxiv.org/content/10.1101/2023.01.28.23285147v2https://www.medrxiv.org/content/10.1101/2023.01.28.23285147v2.abstract`. pages 11, 44, 115, 117

Upasana Tayal, Sanjay Prasad, and Stuart A Cook. Genetics and genomics of dilated cardiomyopathy and systolic heart failure. *Genome Medicine*, 9:20, 2017. ISSN 1756-994X. doi: 10.1186/s13073-017-0410-8. URL `https://doi.org/10.1186/s13073-017-0410-8internal-pdf://304/s13073-017-0410-8.html`. pages 7

Upasana Tayal, Job A.J. Verdonschot, Mark R. Hazebroek, James Howard, John Gregson, Simon Newsome, Ankur Gulati, Chee Jian Pua, Brian P. Halliday, Amrit S. Lota, Rachel J. Buchan, Nicola Whiffin, Lina Kanapeckaite, Resham Baruah, Julian W.E. Jarman, Declan P. O'Regan, Paul J.R. Barton, James S. Ware, Dudley J. Pennell, Bouke P. Adriaans, Sebastiaan C.A.M.

Bekkers, Jackie Donovan, Michael Frenneaux, Leslie T. Cooper, James L. Januzzi, John G.F. Cleland, Stuart A. Cook, Rahul C. Deo, Stephane R.B. Heymans, and Sanjay K. Prasad. Precision phenotyping of dilated cardiomyopathy using multidimensional data. *Journal of the American College of Cardiology*, 79:2219, 6 2022. ISSN 15583597. doi: 10.1016/J.JACC.2022.03.375. URL /pmc/articles/PMC9168440//pmc/articles/PMC9168440/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9168440/. pages 53

Marjola Thanaj, Johanna Mielke, Kathryn A. McGurk, Wenjia Bai, Nicolò Savioli, Antonio de Marvao, Hannah V. Meyer, Lingyao Zeng, Florian Sohler, R. Thomas Lumbers, Martin R. Wilkins, James S. Ware, Christian Bender, Daniel Rueckert, Aidan MacNamara, Daniel F. Freitag, and Declan P. O'Regan. Genetic and environmental determinants of diastolic heart function. *Nature Cardiovascular Research 2022 1:4*, 1:361–371, 4 2022. ISSN 2731-0590. doi: 10.1038/s44161-022-00048-2. URL https://www.nature.com/articles/s44161-022-00048-2. pages 15

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58:267–288, 1 1996. ISSN 2517-6161. doi: 10.1111/J.2517-6161.1996.TB02080.X. URL https://onlinelibrary.wiley.com/doi/full/10.1111/j.2517-6161.1996.tb02080.xhttps://onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02080.xhttps://rss.onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x. pages 70

Geoffrey H. Tison, Jeffrey Zhang, Francesca N. Delling, and Rahul C. Deo. Automated and interpretable patient ecg profiles for disease detection, tracking, and discovery. *Circulation: Cardiovascular Quality and Outcomes*, 12, 9 2019. ISSN 19417705. doi: 10.1161/CIRCOUTCOMES.118.005289. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.118.005289. pages 20

Diem H. Tran and Zhao V. Wang. Glucose metabolism in cardiac hypertrophy and heart failure. *Journal of the American Heart Association*, 8, 6 2019. ISSN 20479980. doi: 10.1161/JAHA.119.012673.

URL https://www.ahajournals.org/doi/abs/10.1161/JAHA.119.012673. pages 108

Jeffrey Tsao, Peter Boesiger, and Klaas P. Pruessmann. k-t blast and k-t sense: Dynamic mri with high frame rate exploiting spatiotemporal correlations. *Magnetic Resonance in Medicine*, 50:1031–1042, 11 2003. ISSN 1522-2594. doi: 10.1002/MRM.10611. URL https://onlinelibrary.wiley.com/doi/full/10.1002/mrm.10611https://onlinelibrary.wiley.com/doi/abs/10.1002/mrm.10611https://onlinelibrary.wiley.com/doi/10.1002/mrm.10611. pages 54

Job A.J. Verdonschot, Mark R. Hazebroek, James S. Ware, Sanjay K. Prasad, and Stephane R.B. Heymans. Role of targeted therapy in dilated cardiomyopathy: The challenging road toward a personalized approach. *Journal of the American Heart Association*, 8, 6 2019. ISSN 2047-9980. doi: 10.1161/JAHA.119.012514. URL https://pubmed.ncbi.nlm.nih.gov/31433726/. pages 4

Job A.J. Verdonschot, Mark R. Hazebroek, Ingrid P.C. Krapels, Michiel T.H.M. Henkens, Anne Raafs, Ping Wang, Jort J. Merken, Godelieve R.F. Claes, Els K. Vanhoutte, Arthur Van Den Wijngaard, Stephane R.B. Heymans, and Han G. Brunner. Implications of genetic testing in dilated cardiomyopathy. *Circulation: Genomic and Precision Medicine*, 13:476–487, 10 2020. ISSN 25748300. doi: 10.1161/CIRCGEN.120.003031. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCGEN.120.003031. pages 105

Julia Vietheer, Lena Lehmann, Claudia Unbehaun, Ulrich Fischer-Rasokat, Jan Sebastian Wolter, Steffen Kriechbaum, Maren Weferling, Beatrice von Jeinsen, Andreas Hain, Christoph Liebetrau, Christian W. Hamm, Till Keller, and Andreas Rolf. Cmr-derived myocardial strain analysis differentiates ischemic and dilated cardiomyopathy-a propensity score-matched study. *The international journal of cardiovascular imaging*, 38:863–872, 4 2021. ISSN 1875-8312. doi: 10.1007/S10554-021-02469-9. URL https://pubmed.ncbi.nlm.nih.gov/34839396/. pages 138

Bjarni J. Vilhjálmsson, Jian Yang, Hilary K. Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po Ru Loh, Gaurav Bhatia, Ron Do, Tristan Hayeck, Hong Hee Won, Benjamin M. Neale, Aiden Corvin, James T.R. Walters, Kai How Farh, Peter A. Holmans, Phil Lee, Brendan Bulik-Sullivan, David A. Collier, Hailiang Huang, Tune H. Pers, Ingrid Agartz, Esben Agerbo, Margot Albus, Madeline Alexander, Farooq Amin, Silviu A. Bacanu, Martin Begemann, Richard A. Belliveau, Judit Bene, Sarah E. Bergen, Elizabeth Bevilacqua, Tim B. Bigdeli, Donald W. Black, Richard Bruggeman, Nancy G. Buccola, Randy L. Buckner, William Byerley, Wiepke Cahn, Guiqing Cai, Dominique Campion, Rita M. Cantor, Vaughan J. Carr, Noa Carrera, Stanley V. Catts, Kimberly D. Chambert, Raymond C.K. Chan, Ronald Y.L. Chen, Eric Y.H. Chen, Wei Cheng, Eric F.C. Cheung, Siow Ann Chong, C. Robert Cloninger, David Cohen, Nadine Cohen, Paul Cormican, Nick Craddock, James J. Crowley, David Curtis, Michael Davidson, Kenneth L. Davis, Franziska Degenhardt, Jurgen Del Favero, Lynn E. Delisi, Ditte Demontis, Dimitris Dikeos, Timothy Dinan, Srdjan Djurovic, Gary Donohoe, Elodie Drapeau, Jubao Duan, Frank Dudbridge, Naser Durmishi, Peter Eichhammer, Johan Eriksson, Valentina Escott-Price, Laurent Essioux, Ayman H. Fanous, Martilias S. Farrell, Josef Frank, Lude Franke, Robert Freedman, Nelson B. Freimer, Marion Friedl, Joseph I. Friedman, Menachem Fromer, Lyudmila Georgieva, Elliot S. Gershon, Ina Giegling, Paola Giusti-Rodrguez, Stephanie Godard, Jacqueline I. Goldstein, Vera Golimbet, Srihari Gopal, Jacob Gratten, Jakob Grove, Lieuwe De Haan, Christian Hammer, Marian L. Hamshere, Mark Hansen, Thomas Hansen, Vahram Haroutunian, Annette M. Hartmann, Frans A. Henskens, Stefan Herms, Joel N. Hirschhorn, Per Hoffmann, Albert Hofman, Mads V. Hollegaard, David M. Hougaard, Masashi Ikeda, Inge Joa, Antonio Julia, Rene S. Kahn, Luba Kalaydjieva, Sena Karachanak-Yankova, Juha Karjalainen, David Kavanagh, Matthew C. Keller, Brian J. Kelly, James L. Kennedy, Andrey Khrunin, Yunjung Kim, Janis Klovins, James A. Knowles, Bettina Konte, Vaidutis Kucinskas, Zita Ausrele Kucinskiene, Hana Kuzelova-Ptackova, Anna K. Kahler, Claudine Laurent, Jimmy

Lee Chee Keong, S. Hong Lee, Sophie E. Legge, Bernard Lerer, Miaoxin Li, Tao Li, Kung Yee Liang, Jeffrey Lieberman, Svetlana Limborska, Carmel M. Loughland, Jan Lubinski, Jouko Lnnqvist, Milan Macek, Patrik K.E. Magnusson, Brion S. Maher, Wolfgang Maier, Jacques Mallet, Sara Marsal, Manuel Mattheisen, Morten Mattingsdal, Robert W. McCarley, Colm McDonald, Andrew M. McIntosh, Sandra Meier, Carin J. Meijer, Bela Melegh, Ingrid Melle, Raquelle I. Mesholam-Gately, Andres Metspalu, Patricia T. Michie, Lili Milani, Vihra Milanova, Younes Mokrab, Derek W. Morris, Ole Mors, Preben B. Mortensen, Kieran C. Murphy, Robin M. Murray, Inez Myin-Germeys, Bertram Mller-Myhsok, Mari Nelis, Igor Nenadic, Deborah A. Nertney, Gerald Nestadt, Kristin K. Nicodemus, Liene Nikitina-Zake, Laura Nisenbaum, Annelie Nordin, Eadbhard O'Callaghan, Colm O'Dushlaine, F. Anthony O'Neill, Sang Yun Oh, Ann Olincy, Line Olsen, Jim Van Os, Christos Pantelis, George N. Papadimitriou, Sergi Papiol, Elena Parkhomenko, Michele T. Pato, Tiina Paunio, Milica Pejovic-Milovancevic, Diana O. Perkins, Olli Pietilinen, Jonathan Pimm, Andrew J. Pocklington, John Powell, Alkes Price, Ann E. Pulver, Shaun M. Purcell, Digby Quested, Henrik B. Rasmussen, Abraham Reichenberg, Mark A. Reimers, Alexander L. Richards, Joshua L. Roffman, Panos Roussos, Douglas M. Ruderfer, Veikko Salomaa, Alan R. Sanders, Ulrich Schall, Christian R. Schubert, Thomas G. Schulze, Sibylle G. Schwab, Edward M. Scolnick, Rodney J. Scott, Larry J. Seidman, Jianxin Shi, Engilbert Sigurdsson, Teimuraz Silagadze, Jeremy M. Silverman, Kang Sim, Petr Slominsky, Jordan W. Smoller, Hon Cheong So, Chris C.A. Spencer, Eli A. Stahl, Hreinn Stefansson, Stacy Steinberg, Elisabeth Stogmann, Richard E. Straub, Eric Strengman, Jana Strohmaier, T. Scott Stroup, Mythily Subramaniam, Jaana Suvisaari, Dragan M. Svrakic, Jin P. Szatkiewicz, Erik Sderman, Srinivas Thirumalai, Draga Toncheva, Paul A. Tooney, Sarah Tosato, Juha Veijola, John Waddington, Dermot Walsh, Dai Wang, Qiang Wang, Bradley T. Webb, Mark Weiser, Dieter B. Wildenauer, Nigel M. Williams, Stephanie Williams, Stephanie H. Witt, Aaron R. Wolen, Emily H.M. Wong, Brandon K. Wormley, Jing Qin Wu, Hualin Simon Xi, Clement C. Zai, Xuebin Zheng, Fritz

Zimprich, Naomi R. Wray, Kari Stefansson, Rolf Adolfsson, Ole A. Andreassen, Peter M. Visscher, Douglas H.R. Blackwood, Elvira Bramon, Joseph D. Buxbaum, Anders D. Børglum, Sven Cichon, Ariel Darvasi, Enrico Domenici, Hannelore Ehrenreich, Tonu Esko, Pablo V. Gejman, Michael Gill, Hugh Gurling, Christina M. Hultman, Nakao Iwata, Assen V. Jablensky, Erik G. Jonsson, Kenneth S. Kendler, George Kirov, Jo Knight, Todd Lencz, Douglas F. Levinson, Qingqin S. Li, Jianjun Liu, Anil K. Malhotra, Andrew McQuillin, Jennifer L. Moran, Bryan J. Mowry, Markus M. Nthen, Roel A. Ophoff, Michael J. Owen, Aarno Palotie, Carlos N. Pato, Tracey L. Petryshen, Danielle Posthuma, Marcella Rietschel, Brien P. Riley, Dan Rujescu, Pak C. Sham, Pamela Sklar, David St. Clair, Daniel R. Weinberger, Jens R. Wendland, Thomas Werge, Mark J. Daly, Patrick F. Sullivan, Michael C. O'Donovan, Peter Kraft, David J. Hunter, Muriel Adank, Habibul Ahsan, Kristiina Aittomäki, Laura Baglietto, Sonja Berndt, Carl Blomquist, Federico Canzian, Jenny Chang-Claude, Stephen J. Chanock, Laura Crisponi, Kamila Czene, Norbert Dahmen, Isabel Dos Santos Silva, Douglas Easton, A. Heather Eliassen, Jonine Figueroa, Olivia Fletcher, Montserrat Garcia-Closas, Mia M. Gaudet, Lorna Gibson, Christopher A. Haiman, Per Hall, Aditi Hazra, Rebecca Hein, Brian E. Henderson, John L. Hopper, Astrid Irwanto, Mattias Johansson, Rudolf Kaaks, Muhammad G. Kibriya, Peter Lichtner, Eiliv Lund, Enes Makalic, Alfons Meindl, Hanne Meijers-Heijboer, Bertram Müller-Myhsok, Taru A. Muranen, Heli Nevanlinna, Petra H. Peeters, Julian Peto, Ross L. Prentice, Nazneen Rahman, María José Sánchez, Daniel F. Schmidt, Rita K. Schmutzler, Melissa C. Southey, Rulla Tamimi, Ruth Travis, Clare Turnbull, Andre G. Uitterlinden, Rob B. Van Der Luijt, Quinten Waisfisz, Zhaoming Wang, Alice S. Whittemore, Rose Yang, Wei Zheng, Sekar Kathiresan, Michele Pato, Carlos Pato, Eli Stahl, Noah Zaitlen, Bogdan Pasaniuc, Eimear E. Kenny, Mikkel H. Schierup, Philip De Jager, Nikolaos A. Patsopoulos, Steve McCarroll, Mark Daly, Shaun Purcell, Daniel Chasman, Michael Goddard, Nick Patterson, and Alkes L. Price. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *American journal of*

human genetics, 97:576–592, 1 2015. ISSN 1537-6605. doi: 10.1016/J.AJHG.2015.09.001. URL https://pubmed.ncbi.nlm.nih.gov/26430803/. pages 10, 115

Lucas F Voges, Lukas C Jarren, and Stephan Seifert. Exploitation of surrogate variables in random forests for unbiased analysis of mutual impact and importance of features. *Bioinformatics*, 39(8), July 2023. doi: 10.1093/bioinformatics/btad471. URL https://doi.org/10.1093/bioinformatics/btad471. pages 71

Roddy Walsh, Kate L. Thomson, James S. Ware, Birgit H. Funke, Jessica Woodley, Karen J. McGuire, Francesco Mazzarotto, Edward Blair, Anneke Seller, Jenny C. Taylor, Eric V. Minikel, Daniel G. MacArthur, Martin Farrall, Stuart A. Cook, and Hugh Watkins. Reassessment of mendelian gene pathogenicity using 7,855 cardiomyopathy cases and 60,706 reference samples. *Genetics in Medicine*, 19:192–203, 2 2017. ISSN 15300366. doi: 10.1038/gim.2016.90. URL http://www.gimjournal.org/article/S1098360021015112/fulltexthttp://www.gimjournal.org/article/S1098360021015112/abstracthttps://www.gimjournal.org/article/S1098-3600(21)01511-2/abstract. pages 7, 42

Andrew Wang. Hypertension and hypertrophic cardiomyopathy. *Hypertrophic Cardiomyopathy*, pages 221–230, 2019. doi: 10.1007/978-3-319-92423-6_16. URL https://link.springer.com/chapter/10.1007/978-3-319-92423-6_16. pages 108

Thomas J. Wang, Jane C. Evans, Emelia J. Benjamin, Daniel Levy, Elizabeth C. LeRoy, and Ramachandran S. Vasan. Natural history of asymptomatic left ventricular systolic dysfunction in the community. *Circulation*, 108:977–982, 8 2003. ISSN 1524-4539. doi: 10.1161/01.CIR.0000085166.44904.79. URL https://pubmed.ncbi.nlm.nih.gov/12912813/https://pubmed.ncbi.nlm.nih.gov/12912813/?dopt=Abstract. pages 5

Hugh Watkins, Houman Ashrafian, and Charles Redwood. Inherited cardiomyopathies. *New*

England Journal of Medicine, 364:1643–1656, 2011. ISSN 0028-4793. doi: 10.1056/NEJMra0902923. URL https://doi.org/10.1056/NEJMra0902923/Users/mk218/Zotero/storage/YVJEZ78R/NEJMra0902923.html. pages 14

Stephen F. Weng, Jenna Reps, Joe Kai, Jonathan M. Garibaldi, and Nadeem Qureshi. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS ONE*, 12: e0174944, 4 2017. ISSN 1932-6203. doi: 10.1371/JOURNAL.PONE.0174944. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0174944. pages 104

Nicola Whiffin, Roddy Walsh, Risha Govind, Matthew Edwards, Mian Ahmad, Xiaolei Zhang, Upasana Tayal, Rachel Buchan, William Midwinter, Alicja E Wilk, Hanna Najgebauer, Catherine Francis, Sam Wilkinson, Thomas Monk, Laura Brett, Declan P O'Regan, Sanjay K Prasad, Deborah J Morris-Rosendahl, Paul J R Barton, Elizabeth Edwards, James S Ware, and Stuart A Cook. Cardioclassifier: disease- and gene-specific computational decision support for clinical genome interpretation. *Genetics in Medicine*, 20:1246–1254, 2018. ISSN 1530-0366. doi: 10.1038/gim.2017.258. URL https://www.nature.com/articles/gim2017258internal-pdf://308/gim2017258.html. pages 21, 81, 82

Kitchener D Wilson, Peidong Shen, Eula Fung, Ioannis Karakikes, Angela Zhang, Kolsoum InanlooRahatloo, Justin Odegaard, Karim Sallam, Ronald W Davis, George K Lui, Euan A Ashley, Curt Scharfe, and Joseph C Wu. A rapid, high-quality, cost-effective, comprehensive and expandable targeted next-generation sequencing assay for inherited heart diseases. *Circulation Research*, 117:603–611, 2015. ISSN 1524-4571. doi: 10.1161/CIRCRESAHA.115.306723. URL http://www.ncbi.nlm.nih.gov/pubmed/26265630. pages 21

Jelmer M. Wolterink, Tim Leiner, Max A. Viergever, and Ivana Išgum. Automatic segmentation and disease classification using cardiac cine mr images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10663 LNCS:

101–110, 2018. ISSN 16113349. doi: 10.1007/978-3-319-75541-0_11/FIGURES/5. URL https://link.springer.com/chapter/10.1007/978-3-319-75541-0_11. pages 17

Yun Xu and Royston Goodacre. On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing*, 2:249, 7 2018. ISSN 25094696. doi: 10.1007/S41664-018-0068-2. URL /pmc/articles/PMC6373628//pmc/articles/PMC6373628/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6373628/. pages 135

Joseph Yeboah, Carlos J. Rodriguez, Brandon Stacey, Joao A. Lima, Songtao Liu, J. Jeffrey Carr, W. Gregory Hundley, and David M. Herrington. Prognosis of individuals with asymptomatic left ventricular systolic dysfunction in the multi-ethnic study of atherosclerosis (mesa). *Circulation*, 126:2713–2719, 12 2012. ISSN 1524-4539. doi: 10.1161/CIRCULATIONAHA.112.112201. URL https://pubmed.ncbi.nlm.nih.gov/23124035/https://pubmed.ncbi.nlm.nih.gov/23124035/?dopt=Abstract. pages 5

Yaohan Yu, Sisi Yu, Xuepei Tang, Haibo Ren, Shuhao Li, Qian Zou, Fakui Xiong, Tian Zheng, and Lianggeng Gong. Evaluation of left ventricular strain in patients with dilated cardiomyopathy. *The Journal of International Medical Research*, 45:2092, 12 2017. ISSN 14732300. doi: 10.1177/0300060517712164. URL /pmc/articles/PMC5805211//pmc/articles/PMC5805211/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC5805211/. pages 138

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, December 2005. doi: 10.1111/j.1467-9868.2005.00532.x. URL https://doi.org/10.1111/j.1467-9868.2005.00532.x. pages 70

Marina Zaromytidou and Konstantinos Savvatis. The weight of obesity in hypertrophic cardiomyopa-

thy. *Clinical Medicine*, 23:357, 7 2023. ISSN 14734893. doi: 10.7861/CLINMED.2023-0194. URL /pmc/articles/PMC10541040//pmc/articles/PMC10541040/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC10541040/. pages 110

Kan Zhang, Richard Sheu, Nicole M. Zimmerman, Andrej Alfirevic, Shiva Sale, A. Marc Gillinov, and Andra E. Duncan. A comparison of global longitudinal, circumferential, and radial strain to predict outcomes after cardiac surgery. *Journal of cardiothoracic and vascular anesthesia*, 33:1315–1322, 5 2019. ISSN 1532-8422. doi: 10.1053/J.JVCA.2018.10.031. URL https://pubmed.ncbi.nlm.nih.gov/30581109/. pages 137

Yiliang Zhang, Qiongshi Lu, Yixuan Ye, Kunling Huang, Wei Liu, Yuchang Wu, Xiaoyuan Zhong, Boyang Li, Zhaolong Yu, Brittany G. Travers, Donna M. Werling, James J. Li, and Hongyu Zhao. Supergnova: local genetic correlation analysis reveals heterogeneous etiologic sharing of complex traits. *Genome Biology 2021 22:1*, 22:1–30, 9 2021. ISSN 1474-760X. doi: 10.1186/S13059-021-02478-W. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02478-w. pages 53

Sean L. Zheng, Albert Henry, Douglas Cannie, Michael Lee, David Miller, Kathryn A McGurk, Isabelle Bond, Xiao Xu, Hanane Issa, Catherine Francis, Antonio De Marvao, Pantazis I Theotokis, Rachel J Buchan, Doug Speed, Erik Abner, Lance Adams, Krishna G. Aragam, Johan Arnlov, Anna Axelsson Raja, Joshua D. Backman, John Baksi, Paul JR Barton, Kiran J. Biddinger, Eric Boersma, Jeffrey Brandimarto, Soren Brunak, Henning Bundgaard, David J. Carey, Philippe Charron, James P. Cook, Stuart A Cook, Spiros Denaxas, Jean-Francois Deleuze, Alexander S. Doney, Perry Elliot, Christian Erikstrup, Tonu Esko, Eric H. Farber-Eger, Chris Finan, Sophie Garnier, Jonas Ghouse, Vilmantas Giedraitis, Daniel F. Gudbjartsson, Christopher M. Haggerty, Brian P. Halliday, Anna Helgadottir, Harry Hemingway, Hans Hillege, Isabella Kardys, Lars Lind, Cecilia M. Lindgren, Brandon D. Lowery, Charlotte Manisty, Kenneth B. Margulies, Ify R. Mordi,

Michael P. Morley, Andrew D. Morris, Andrew P. Morris, Lori Morton, Mahdad Noursadeghi, Sisse R Ostrowski, Anjali T. Owens, Colin N. A. Palmer, Antonis Pantazis, Ole Birger Pedersen, Sanjay K. Prasad, Akshay Shekhar, Diane T. Smelser, Sundararajan Srinivasan, Kari Stefansson, Gardar Sveinbjornsson, Petros Syrris, Mari-Liis Tammesoo, Upasana Tayal, Maris Teder-Laving, Gudmundur Thorgeirsson, Unnur Thorsteinsdottir, Vinicius Tragante, David-Alexandre Tregouet, Thomas A. Treibel, Henrik Ullum, Ana Valdes, Jessica van Setten, Marion van Vugt, Abirami Veluchamy, Monique Verschuuren, Eric Villard, Yifan Yang, COVIDsortium, DBDS Genomic Consortium, Genomics England Research Consortium, HERMES Consortium, Folkert W. Asselbergs, Thomas P. Cappola, Marie-Pierre Dube, Michael E. Dunn, Patrick T. Ellinor, Aroon D. Hingorani, Chim C. Lang, Nilesh J. Samani, Svati Shah, Gustav Smith, Ramachandran S. Vasan, Declan P. O'Regan, Hilma Holm, Michela Noseda, Quinn S. Wells, James S. Ware, and R. Thomas Lumbers. Genome-wide association analysis reveals insights into the molecular aetiology underlying dilated cardiomyopathy. *medRxiv*, 36:2023.09.28.23295408, 9 2023a. doi: 10.1101/2023.09.28.23295408. URL `https://www.medrxiv.org/content/10.1101/2023.09.28.23295408v1https://www.medrxiv.org/content/10.1101/2023.09.28.23295408v1.abstract`. pages 13, 53, 103, 115, 116, 145

Sean L Zheng, Sean J Jurgens, Kathryn A McGurk, Xiao Xu, Chris Grace, Pantazis I Theotokis, Rachel J Buchan, Catherine Francis, Antonio de Marvao, Lara Curran, Wenjia Bai, Chee Jian Pua, Tang Hak Chiaw, Paloma Jorda, Marjon A van Slegtenhorst, Judith MA Verhagen, Andrew R Harper, Elizabeth Ormondroyd, Calvin WL Chin, Genomics England Research Consortium, HCM GWAS Collaborators, Antonis Pantazis, John Baksi, Brian P Halliday, Paul Matthews, Yigal M Pinto, Roddy Walsh, Ahmad S Amin, Arthur AM Wilde, Stuart A Cook, Sanjay K Prasad, Paul JR Barton, Declan P O'Regan, RT Lumbers, Anuj Goel, Rafik Tadros, Michelle Michels, Hugh Watkins, Connie R Bezzina, and James S Ware. Evaluation of polygenic score for hypertrophic cardiomyopathy in the general

population and across clinical settings. *medRxiv*, 2023b. doi: 10.1101/2023.03.14.23286621. URL https://www.medrxiv.org/content/early/2023/03/16/2023.03.14.23286621. pages 11, 53, 103, 116, 125, 134, 136, 145

Xiaopu Zhou, Yu Chen, Fanny C. F. Ip, Yuanbing Jiang, Han Cao, Ge Lv, Huan Zhong, Jiahang Chen, Tao Ye, Yuewen Chen, Yulin Zhang, Shuangshuang Ma, Ronnie M. N. Lo, Estella P. S. Tong, Michael W. Weiner, Paul Aisen, Ronald Petersen, Clifford R. Jack, William Jagust, John Q. Trojanowski, Arthur W. Toga, Laurel Beckett, Robert C. Green, Andrew J. Saykin, John Morris, Leslie M. Shaw, Zaven Khachaturian, Greg Sorensen, Lew Kuller, Marcus Raichle, Steven Paul, Peter Davies, Howard Fillit, Franz Hefti, David Holtzman, Marek M. Mesulam, William Potter, Peter Snyder, Adam Schwartz, Tom Montine, Ronald G. Thomas, Michael Donohue, Sarah Walter, Devon Gessert, Tamie Sather, Gus Jiminez, Danielle Harvey, Matthew Bernstein, Paul Thompson, Norbert Schuff, Bret Borowski, Jeff Gunter, Matt Senjem, Prashanthi Vemuri, David Jones, Kejal Kantarci, Chad Ward, Robert A. Koeppe, Norm Foster, Eric M. Reiman, Kewei Chen, Chet Mathis, Susan Landau, Nigel J. Cairns, Erin Householder, Lisa Taylor-Reinwald, Virginia Lee, Magdalena Korecka, Michal Figurski, Karen Crawford, Scott Neu, Tatiana M. Foroud, Steven G. Potkin, Li Shen, Kelley Faber, Sungeun Kim, Kwangsik Nho, Leon Thal, Neil Buckholtz, Marylyn Albert, Richard Frank, John Hsiao, Jeffrey Kaye, Joseph Quinn, Betty Lind, Raina Carter, Sara Dolen, Lon S. Schneider, Sonia Pawluczyk, Mauricio Beccera, Liberty Teodoro, Bryan M. Spann, James Brewer, Helen Vanderswag, Adam Fleisher, Judith L. Heidebrink, Joanne L. Lord, Sara S. Mason, Colleen S. Albers, David Knopman, Kris Johnson, Rachelle S. Doody, Javier Villanueva-Meyer, Munir Chowdhury, Susan Rountree, Mimi Dang, Yaakov Stern, Lawrence S. Honig, Karen L. Bell, Beau Ances, Maria Carroll, Sue Leon, Mark A. Mintun, Stacy Schneider, Angela Oliver, Daniel Marson, Randall Griffith, David Clark, David Geldmacher, John Brockington, Erik Roberson, Hillel Grossman, Effie Mitsis, Leyla de Toledo-Morrell, Raj C. Shah, Ranjan Duara, Daniel Varon, Maria T. Greig, Peggy Roberts,

Chiadi Onyike, Daniel D'Agostino, Stephanie Kielb, James E. Galvin, Brittany Cerbone, Christina A. Michel, Henry Rusinek, Mony J. de Leon, Lidia Glodzik, Susan De Santi, P. Murali Doraiswamy, Jeffrey R. Petrella, Terence Z. Wong, Steven E. Arnold, Jason H. Karlawish, David Wolk, Charles D. Smith, Greg Jicha, Peter Hardy, Partha Sinha, Elizabeth Oates, Gary Conrad, Oscar L. Lopez, MaryAnn Oakley, Donna M. Simpson, Anton P. Porsteinsson, Bonnie S. Goldstein, Kim Martin, Kelly M. Makino, M. Saleem Ismail, Connie Brand, Ruth A. Mulnard, Gaby Thai, Catherine McAdams-Ortiz, Kyle Womack, Dana Mathews, Mary Quiceno, Ramon Diaz-Arrastia, Richard King, Myron Weiner, Kristen Martin-Cook, Michael DeVous, Allan I. Levey, James J. Lah, Janet S. Cellar, Jeffrey M. Burns, Heather S. Anderson, Russell H. Swerdlow, Liana Apostolova, Kathleen Tingus, Ellen Woo, Daniel H. S. Silverman, Po H. Lu, George Bartzokis, Neill R. Graff-Radford, Francine Parfitt, Tracy Kendall, Heather Johnson, Martin R. Farlow, Ann Marie Hake, Brandy R. Matthews, Scott Herring, Cynthia Hunt, Christopher H. van Dyck, Richard E. Carson, Martha G. MacAvoy, Howard Chertkow, Howard Bergman, Chris Hosein, Ging-Yuek Robin Hsiung, Howard Feldman, Benita Mudge, Michele Assaly, Charles Bernick, Donna Munic, Andrew Kertesz, John Rogers, Dick Trost, Diana Kerwin, Kristine Lipowski, Chuang-Kuo Wu, Nancy Johnson, Carl Sadowsky, Walter Martinez, Teresa Villena, Raymond Scott Turner, Kathleen Johnson, Brigid Reynolds, Reisa A. Sperling, Keith A. Johnson, Gad Marshall, Meghan Frey, Barton Lane, Allyson Rosen, Jared Tinklenberg, Marwan N. Sabbagh, Christine M. Belden, Sandra A. Jacobson, Sherye A. Sirrel, Neil Kowall, Ronald Killiany, Andrew E. Budson, Alexander Norbash, Patricia Lynn Johnson, Joanne Allard, Alan Lerner, Paula Ogrocki, Leon Hudson, Evan Fletcher, Owen Carmichael, John Olichney, Charles DeCarli, Smita Kittur, Michael Borrie, T-Y. Lee, Rob Bartha, Sterling Johnson, Sanjay Asthana, Cynthia M. Carlsson, Adrian Preda, Dana Nguyen, Pierre Tariot, Stephanie Reeder, Vernice Bates, Horacio Capote, Michelle Rainka, Douglas W. Scharre, Maria Kataki, Anahita Adeli, Earl A. Zimmerman, Dzintra Celmins, Alice D. Brown, Godfrey D. Pearlson, Karen Blank, Karen

Anderson, Robert B. Santulli, Tamar J. Kitzmiller, Eben S. Schwartz, Kaycee M. Sink, Jeff D. Williamson, Pradeep Garg, Franklin Watkins, Brian R. Ott, Henry Querfurth, Geoffrey Tremont, Stephen Salloway, Paul Malloy, Stephen Correia, Howard J. Rosen, Bruce L. Miller, Jacobo Mintzer, Kenneth Spicer, David Bachman, Stephen Pasternak, Irina Rachinsky, Dick Drost, Nunzio Pomara, Raymundo Hernando, Antero Sarrael, Susan K. Schultz, Laura L. Boles Ponto, Hyungsub Shim, Karen Elizabeth Smith, Norman Relkin, Gloria Chaing, Lisa Raudin, Amanda Smith, Kristin Fargher, Balebail Ashok Raj, Thomas Neylan, Jordan Grafman, Melissa Davis, Rosemary Morrison, Jacqueline Hayes, Shannon Finley, Karl Friedl, Debra Fleischman, Konstantinos Arfanakis, Olga James, Dino Massoglia, J. Jay Fruehling, Sandra Harding, Elaine R. Peskind, Eric C. Petrie, Gail Li, Jerome A. Yesavage, Joy L. Taylor, Ansgar J. Furst, Vincent C. T. Mok, Timothy C. Y. Kwok, Qihao Guo, Kin Y. Mok, Maryam Shoai, John Hardy, Lei Chen, Amy K. Y. Fu, and Nancy Y. Ip. Deep learning-based polygenic risk analysis for alzheimer's disease prediction. *Communications Medicine 2023 3:1*, 3:1–20, 4 2023. ISSN 2730-664X. doi: 10.1038/s43856-023-00269-x. URL `https://www.nature.com/articles/s43856-023-00269-x`. pages 116, 119