POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Multi-outcome feature selection via anomaly detection autoencoders

An application to radiogenomics in breast cancer patients

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

Author: **Alessia Mapelli**

Student ID: 10572745
Advisor: Prof. Francesca Ieva
Co-advisors: Michela Massi, Nicola Rares Franco
Academic Year: 2021-22

Abstract

Whenever external beam radiotherapy is employed on normal tissue to irradiate tumors, side effects may arise as drawbacks of this non-invasive treatment. In particular, toxicities happen when the radiation damages healthy tissue and, in radiosensitive patients, can occur years after radiotherapy impairing quality of life. These complications can depend on several factors, among which are the radiation dose, the volume of the organ irradiated, and the patient's demographics. Coupled with environmental factors, illnesses' impact on individuals can be affected by changes in either one or many of their genes. RADprecise international study aims at personalizing radiotherapy treatment for cancer patients by improving prediction models for the risk of long-term side effects after radiotherapy including innovative biomarkers. Within the RADprecise project, this thesis attempts to include genetic information effects in late-toxicities risk models for breast-cancer patients through an interpretable selection of the most informative genetic variants. Risk models describing radiosensitivity could then be employed by physicians to take more informed individual decisions in cancer treatment.

Several complexities arise from the radiogenomics context: high-dimensionality of the data, unbalancing classes where a minority class of patients presents toxicities, imputation and noise in genomic data collection, and the presence of high-order interaction among genes influencing the toxicities development. Moreover, multivariate analysis is necessary for comprehensive treatment decisions and for feature selection since genetic variants determining inter-individual differences in radiosensitivity are only partly toxicity-specific. The methodology implemented in this work performs a multi-outcome selection of the genetic variants, tackling all the aforementioned complexities, to produce a set of informative features for each of the toxicities measured in the study and general radiosensitivity, accounting for the correlation structure present between the outcomes. The developed model consists of an ensemble method based on anomaly detection autoencoders whose reconstruction error is studied to detect radiosensitive patients and discover the genetic variants correlated with toxicities arising. Each anomaly detection autoencoder within the ensemble is enriched with a denoising technique that robustifies the analysis to the noise of imputed genomic data. The model proprieties are studied in a simulation setting.

Finally, the method is applied to a case study out of REQUITE database provided within the RADprecise project.

Keywords: Radiogenomics, Late toxicity, Breast cancer, Multi-outcome feature selection, Ensamble learning, Anomaly detection autoencoder, Denoising.

Abstract in lingua italiana

Ogni volta che la radioterapia a fasci esterni viene impiegata sui tessuti normali per irradiare i tumori, possono insorgere effetti collaterali come inconvenienti causati dall'irradiazione dei tessuti sani. Queste tossicità possono manifestarsi nei pazienti radiosensibili, anni dopo la radioterapia, compromettendo la loro qualità di vita. L'insorgenza di complicazioni può dipendere da diversi fattori, tra cui la quantità di radiazioni, il volume dell'organo irradiato e le caratteristiche demografiche del paziente. Insieme ai fattori ambientali, l'impatto delle malattie sugli individui può essere influenzato da cambiamenti nel loro DNA di uno o più geni.
Il progetto internazionale RADprecise mira a personalizzare il trattamento radioterapico per i pazienti oncologici migliorando i modelli di previsione del rischio di effetti collaterali a lungo termine dopo la radioterapia includendo biomarcatori innovativi.
Nell'ambito del progetto RADprecise, questa tesi cerca di includere gli effetti genetici nei modelli di rischio di tossicità tardiva per le pazienti affette da cancro al seno, attraverso una selezione interpretabile delle varianti genetiche più informative. I modelli di rischio che descrivono la radiosensibilità del paziente potrebbero quindi essere utilizzati dai medici per prendere decisioni individuali più informate nel trattamento del cancro.

Diverse complessità derivano dal contesto clinico: elevata dimensionalità dei dati, classi non equilibrate in cui una minoranza di pazienti presenta tossicità, imputazione e rumore nella raccolta dei dati genomici e presenza di interazioni di alto ordine tra i geni che influenzano lo sviluppo delle tossicità. Inoltre, è necessario fare inferenza su più effetti collaterali insieme per decisioni terapeutiche complete, e utilizzare tecniche multivariate per la selezione delle caratteristiche, poiché le varianti genetiche che determinano le differenze interindividuali nella radiosensibilità sono solo in parte specifiche della tossicità.
La metodologia implementata in questo lavoro esegue una selezione delle varianti genetiche considerando l'insieme delle tossicità, affrontando tutte le complessità sopra menzionate, per produrre un insieme di variazioni genetiche informative per ciascuna delle tossicità che tengano conto della struttura di correlazione presente tra di esse. La metodologia è basata su un apprendimento ensemble (o d'insieme) che sfrutta come base autoencoder per il rilevamento di anomalie, il cui errore di ricostruzione è studiato per individuare

i pazienti radiosensibili e scoprire le covariate correlate all'insorgenza di effetti collaterali. Ogni autoencoder all'interno dell'ensamble è arricchito con una tecnica di riduzione del rumore che rende l'analisi più robusta rispetto a possibili errori di imputazione. Le proprietà del modello sono studiate in un contesto di simulazione. Infine, il metodo è stato applicato a un caso di studio estratto dal database REQUITE fornito nell'ambito del progetto RADprecise.

Parole chiave: Radiogenomica, Tossicità tardiva, Cancro al seno, Selezione delle covariate multivariata, Apprendiamento di insieme, Autoencoder per il rilevamento di anomalie, Rimozione del rumore.

Contents

Introduction

Thanks to treatments such as radiotherapy, the survival in patients diagnosed with cancer is increasing [1]. However, approximately 5% of patients receiving radiotherapy are particularly sensitive to irradiation and likely to develop side effects after radiotherapy [32]. This thesis work is developed within a large international study, namely, RADprecise [32], aiming at personalizing radiotherapy treatment for cancer patients by improving prediction models for the risk of long-term side effects after radiotherapy including innovative biomarkers [32]. RADprecise project is an extension of REQUITE consortium study that standardized data collection across multiple countries and centers to achieve a unique large database with homogeneous information exploitable in risk model validation [54]. Radiosensitivity is a latent outcome and it is only inferred through measurements of various types of late toxicities. Long term side effects are measured as binary responses or endpoints and multivariate inference is necessary for comprehensive treatment decisions. Normal Tissue Complication Probability (NTCP) is a model-based risk estimate that physicians routinely use to make treatment decisions. Traditional NTCPs model the risk of radio-induced complications in terms of radiation dose (D), and partial volume irradiated (v) [14]. In recent years, new statistical and machine learning methodologies were introduced to expand the set of predictors, including clinical information and biomarkers in risk modeling [6, 35, 40, 41]. Genetic biomarkers are believed to be crucial in predicting late toxicity development [54]. Therefore, their incorporation into NTCPs models may substantially improve personalized treatment planning. A Polygenic Risk Score (PRS) summarizes a patient's genetic predisposition to a disease. In radiogenomics, it is usually computed as the score associated with each patient by a predictive model,such as logistic regression, that links the risk of developing late toxicities to the presence of associated genetic mutations in the patient DNA [33]. The clinical problem for this study can then be rephrased as the need to implement a methodology to incorporate an interpretable PRS into an NTCP logistic model. In general, as with any other classification model, PRS models perform at best when fed with highly influential features that provide intrinsic information and discriminant properties for class separability. Moreover, Features Selection (FS) is fundamental when variables are many and highly correlated. This is typical of genomic studies, where data is highly dimensional (i.e. up to million genetic features) and the curse of dimensionality plays an important role. Indeed the work presented in this thesis is focused on the task of FS for genetic data.

To achieve the goal the dataset designed in REQUITE consortium study is exploited. The information available in the database are both clinical and genetic. Raw genotype data carry challenging characteristics that hinder the applicability of most traditional statistical models for FS: the presence of **imbalance** and the **imputation** in genomic information can seriously bias analysis results [8, 17]. The unbalanced binary traits, determined by the study of rare phenotypical traits (such as late toxicities), pose serious challenges to genomic selection due to a very low case-control ratio that may violate asymptotic assumptions of statistical inference [19]. Imputation methods estimate genotype probabilities at variants not genotyped to achieve completeness in genetic information [17]. Tests of the association of imputed SNPs with the phenotype of interest must be carried out with great care because of their probabilistic nature. Ignoring the genotypic uncertainty and performing analysis with standard statistical tools generally affect the power of the association study [17]. Moreover, the latest radiogenic studies in late-toxicity radiotherapy, reveal that epistasis, or gene-gene interactions, affect polygenic susceptibility to common human diseases, suggesting complex interactions are more important than the effects of any single common genetic variant [22]. The biological relevance of interactions introduces another source of complexity. The introduction of complex interactions in predictive models could effectively discriminate between classes of phenotypes (i.e. cases/controls, etc.). In turn, FS methods need to be able to consider the potentially predictive power of such interactions during selection. However, Traditional FS techniques usually just consider the main effect of covariates in performing the selection and become sub-optimal when the high-order interaction effect is more important then any single genetic variant. Finally, in precision medicine applications, there is an increasing need to model **several endpoints simultaneously**. In fact, risk models developed for all endpoints simultaneously can improve performance by borrowing information from other endpoints and can identify predisposing factors associated with radiosensitivity without explicitly defining it [41].
The presence of challenging characteristics in the data within radiogenomics analysis implies the need to introduce specific techniques to perform a robust and reliable analysis. Most of the above-mentioned complexities have been recently addressed in [37, 39], where the authors develop a Deep Learning-based FS method for imbalanced data. The genetic features selected via their Deep Sparse AutoEncoder Ensemble (DSAEE) are subsequently included in an interaction-aware method for polygenic risk scoring (PRSi) [22]. In brief, the DSAEE FS method exploits Deep Sparse AutoEncoders as weak learners.

AutoEncoders are trained to learn the normal patterns in the majority class observations and tested on both majority and minority class data, mimicking AutoEncoders' usage in anomaly detection. The FS method in [39] presents three major benefits: the ability to deal with heavy class imbalance the interaction-aware selection and the interpretability of the selection method.

However, this effective algorithm does not account for the multivariate aspect of the LT prediction. In fact, in this, and many similar precision medicine applications, the need to simultaneously model several phenotypic traits or endpoints entails the importance in identify predisposing factors associated with radiosensitivity without explicitly defining it. The main contribution of this work is the improvement of the DSAEE method, generalizing it to a **multiendpoint framework** and widening its applicability to **breast cancer** late toxicities. Specifically, this thesis proposes an innovative methodology capable of performing variable selection in high-dimensional contexts where high-order interactions are of interest and multiple outcomes are simultaneously studied. The multivariate FS is developed specifically to work on genomic data. The algorithm is robust to data imputation and suitable for multiple binary classification problems with high imbalance in the classes of each outcome. As in the case of the work in [22], the selection can be exploited to efficiently include genetic effects in clinical risk models. In this work, each of the fundamental aspects of the feature selection methods is individually studied to properly understand its characteristics: robustness to imputation error is achieved by introducing a denoising procedure in the training process of the ensemble learning autoencoders (Chapter 2 section 2.1) so that the imputation error is considered in the reconstruction error of both groups, leading to an unbiased analysis; selection of the discriminating feature is based on the ensemble learning approximation of the reconstruction error distribution in the groups (Chapter 2 section 2.2). Differences in the distribution over groups can be discovered via a non-parametric test and generalized to the entire population revealing features able to distinguish between them. Finally, multi-outcome feature selection is achieved by the proper definition of the autoencoder control group, accounting for correlation in the endpoints (Chapter 2 section 2.3). Feature selection is performed starting from a unique control sample, extracted from the intersection of each endpoint control group, and an endpoint-depending test sample. Such a selection produces a set of SNPs describing each specific toxicity accounting for the dependency structure in the multivariate outcome. The selected SNPs can then be combined to form an informative set of features correlated with general toxicity and able to distinguish generally radiosensitive patients (Chapter 2 section 2.3).

In conclusion, this thesis offers an end-to-end pipeline for a multi-outcome feature selection method (Chater 2 section 2.4) in genomic applications.

Structure of the Thesis

The structure of this Thesis is composed of four chapters. In **Chapter 1** the motivation behind the work and the specific research question stemming from the RADprecise case study are detailed and discussed. The available data are presented and their preprocessing is described, discussing possible problems that arise from the peculiar context of genomic data. A brief overview of the theoretical background necessary to understand the novel methodology developed in the thesis work is also provided. **Chapter 2** details the feature selection method developed as original contribution. **Chapter 3** presents the applications of the proposed methods to both simulation data and the REQUITE case study. The work is concluded by a **Discussion** based on the results obtained and a few proposals for possible future developments. **Appendix A** presents details about row data. In **Appendix B** complementary information about the simulated analyses are presented. **Appendix C** shows additional results and implementation detail in REQUITE application. Analyses were carried out using Python [53] and R [45]. All code files are available on the GitHub repository: *https://github.com/AlessiaMapelli/Master_ Thesis*

1 | Materials and literature review

1.1. The RADprecise project

1.1.1. General aim of the project

The RADprecise project is researching side effects of treatment in breast and prostate cancer up and beyond 5 years after radiotherapy. Thanks to treatments such as radiotherapy more patients are surviving and thus living longer [1]. About half of all cancer patients receive radiotherapy and approximately 5% of patients are particularly sensitive to irradiation, making them more likely than others to develop side effects after radiotherapy [32]. Side effects happen when the radiation that kills cancer also damages healthy tissue and can occur years after radiotherapy impairing quality of life. RADprecise project is an extension of REQUITE consortium study that standardized data collection across multiple countries and centers to achieve a unique large database with homogeneous information exploitable in risk model validation [54]. RADprecise international study aims at personalizing radiotherapy treatment for cancer patients by improving prediction models for the risk of long-term side effects after radiotherapy including innovative biomarkers. The project focuses on investigating the cellular response to irradiation and creating methods to improve personalized treatment planning and minimize radiation toxicity. One of the cellular response keys to late toxicity development is believed to be found in genetics [54], identifying genetic biomarkers and including their effect in late toxicity risk models could lead to an improvement in prediction and meaningful insight into the magnitude of genetic effects. Identifying predisposing factors can contribute to the prevention of severe late toxicities and their incorporation into treatment planning systems give the chance to individualize radiotherapy treatment for each patient. Predictive models are also important when informing a patient about his chance of toxicity [41]. The RADprecise project is illustrated in Figure 1.1.

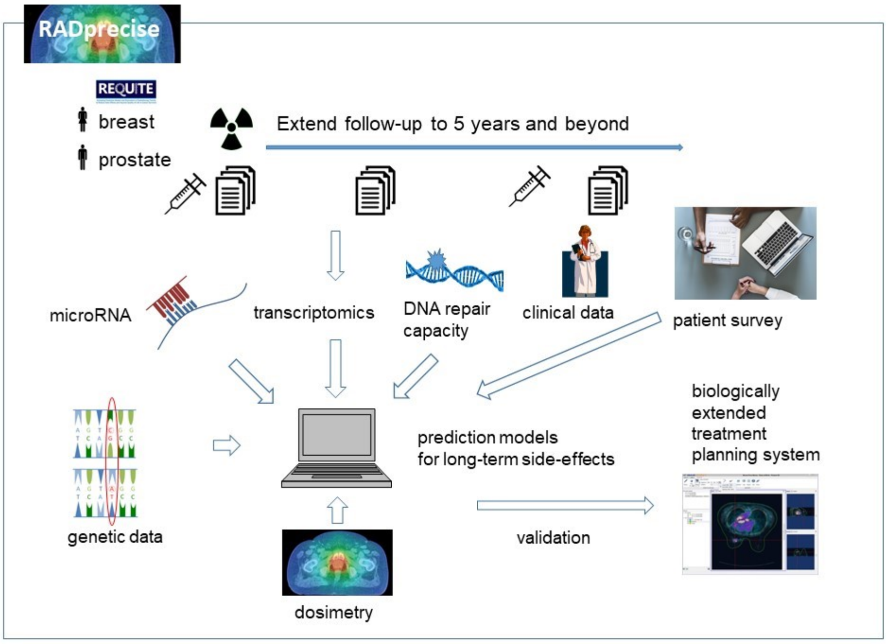

Figure 1.1: **Illustration of RADprecise study**. The RADprecise project is researching side effects of treatment in breast and prostate cancer up and beyond 5 years after radiotherapy. RADprecise's aim is to include genetic data, microRNA, transcriptomics, and DNA repair capacity information, in addition to clinical data to predictive models for long-term side effects. Source: [32]

1.1.2. Clinical models in late toxicity studies

During external beam radiotherapy, normal tissues are irradiated along with the tumor. Radiation therapists try to minimize the dose to normal tissues while delivering a high dose to the target volume. Often this is difficult and complications arise due to the irradiation of normal tissues. These complications, also called toxicities, can depend on several factors among which the more widely known influential covariates are the radiation dose and the volume of the organ irradiated [14]. Normal Tissue Complication Probability (NTCP) is a model-based risk estimate that physicians routinely use to make personalized treatment decisions. Generally, NTCP models attempt to reduce complicated dosimetric and anatomic information to a single risk measure [35].

"Dosimetric" predictors are those variables that relate specifically to the delivery of radiation while "non-dosimetric" or "clinical" predictors include all other variables, such as age, sex, and histology [29]. Traditional NTCP modeling is based on the Lyman-Kutcher-Burman power law model for calculating normal tissue response with regard to nonuniform arbitrary organ fractions. The base of the model is an error function that connects the three variables of interest considered, normal tissue complication probability (NTCP), dose (D), and partial volume (v) [14]. The resultant model is parameterized by the dose-volume histogram (DHV) reduced to a scalar equivalent. NTCPs have conventionally focused on using dosimetric predictors alone but the necessity of using an additional pool of non-dosimetric predictors was highlighted in QUANTEC [35], a group formed to update existing models and address new modeling issues in radiation oncology. Consequently, new statistical and machine learning methodologies were introduced to expand the set of considered predictors. The normal tissue toxicities following treatment can be represented as multiple binary responses or endpoints and can be predicted via classification methods [31]. One of the most commonly used is the logistic regression model which combines both dosimetric and non-dosimetric covariates in a unique probability prediction.

This thesis attempts to include genetic information effects in NTCP models summarizing it in a Polygenic Risk Score (PRS). A PRS summarizes a patient's genetic predisposition to a disease. In radiogenomics, it is usually computed as the score associated with each patient by a predictive model, as logistic regression, that links the risk of developing late toxicities to the presence of associated genetic mutations in the patient DNA [33]. Illnesses' impact on individuals can be affected by changes in either one or many of their genes, frequently coupled with environmental factors. Researchers are studying genomic variants to understand the role that they play in diseases across different populations by comparing the genomes of individuals with and without those diseases calculating which variants tend to be found more frequently in groups of people with a given disease [42]. Radiogenic studies in late-toxicity radiotherapy, are finding an increasing number of common genetic variants, called single nucleotide polymorphisms (SNPs). A polymorphism is a genetic variation leading to the presence, in the population of a species, of multiple alleles of the same gene. A SNP is a genetic polymorphism in which a given gene exhibits, in different individuals, sequence variations that are due to a single base of the polynucleotide chain. The presence of SNPs in a given gene can change its structure, its expression level, or the function of the encoded protein, making it unique to that individual. There is also increasing awareness that epistasis, or SNP-SNP interactions, affect polygenic susceptibility to common human diseases suggesting complex interactions are more important than the effects of any single common genetic variant [22]. Standard weighted PRS estimation relies on Genome-Wide Association Study (GWAS) summary statistics obtained on one or more discovery cohorts modeling the independent effect of individual SNPs on the outcome. These PRSs exploit SNP-specific odds ratios or effect sizes to weigh the contribution of the risk alleles on the disease risk or outcome. The set of SNPs to be included in this estimation may of course affect the score's predictive power significantly. Some approaches include all SNPs, with the risk of incorporating useless or redundant information, while others retain a subset of SNPs based on predefined criteria (e.g., those passing an arbitrary p-value threshold in the GWAS results)[38]. In this thesis, the process to define and evaluate PRS score instead followed the methodologies introduced in [22]. In the paper, a novel methodology to compute PRS is developed to include high-order interactions between genetic variations. The algorithm once defined the most important features, defines the PRS as the score of a logistic regression trained to predict the presence or absence of the selected late toxicity. This methodology will be further discussed in Section 1.3.5. PRS evaluation can then be incorporated into NTCP models to include individual patient genetic data in personalized treatment.

1.1.3. Statistical modeling for multiple endpoints in late toxicity studies

While current guidelines generally recommend single endpoints for primary analyses of confirmatory clinical trials, it is recognized that certain settings require inference on multiple endpoints for comprehensive conclusions on treatment effects [46]. Factors determining inter-individual differences in radiosensitivity outcomes are only partly toxicity-specific [30]. Therefore, combining treatment effect estimates from several outcome measures can increase the statistical power of the resulting models. Such an efficient use of resources is of special relevance for trials in small populations [46]. Prediction models developed for all endpoints simultaneously can improve the predictions of a given endpoint by borrowing information from other endpoints and are able to identify predisposing factors associated with radiosensitivity without explicitly defining/quantifying radiosensitivity [41]. Consequently, it is vital to consider multiple endpoints in prediction models, designing new approaches for modeling overall radiosensitivity and predicting multiple toxicity endpoints [41]. Main multi-endpoint analysis examples in the literature concern multi-endpoint testing, a review can be found in [43, 46] or single-endpoint multivariate logistic regression models with modified maximum likelihood estimation of the logistic coefficients [41]. An efficient methodology to perform genetic variant selection in a multi-outcome setting prior to their inclusion in a classification model could lead to higher performance and interpretability.

1.2. Materials

1.2.1. Data presentation

This thesis work is based on the REQUITE database. The REQUITE project included over 4400 patients and the data were collected in the same way in 26 hospitals in 80 countries establishing a standardized data collection across multiple centers. The importance of collecting the same information from different countries is to create a unique database with homogeneous information even when radiotherapy has been delivered in different ways [54]. Patients were recruited prior to radiotherapy (baseline) and followed up to five years after treatment. The following standardized data were collected prospectively by case report forms (CRFs): demographics, co-morbidity, treatment (with comprehensive information on radiotherapy regimens and dose to organs at risk), physics, longitudinal standardized radiotherapy toxicity (following Common Terminology Criteria for Adverse Events (CTCAE v4.0 [27]), quality-of-life, and treatment outcome. Late toxicity data were collected at 12, 24, 36, 48 and 60 months from the initial cancer treatment via patient-reported outcome (PRO) forms, together with quality-of-life information [54]. All patients donated at least two blood samples prior to the start of radiotherapy. Samples were exploited for Single Nucleotide Polymorphisms (SNP) genotyping and to retrieve microRNA [54]. The REQUITE CRFs, questionnaires and omic data were stored generating a huge database that can be exploited for long-term side effects prediction and study of biomarkers and radiosensitivity relation. Details on the REQUITE population are published in [49].

Summarizing, the REQUITE project database includes:

- patient's clinical history, demographics, co-morbidity and patient's habits;
- patient's comprehensive information on cancer treatment including tumor information, surgery information, radiotherapy and chemotherapy information;
- patient's late radiotherapy toxicity information for each follow-up visit;
- patient's genetic information including Single Nucleotide Polymorphisms (SNP) genotyping.

Further details about the dataset can be found in [49] and in Appendix A.

To study the genetic correlation with different late toxicities a subset of the genotyped SNPs was considered. Only SNPs previously linked to radiotherapy sensitivity in breast cancer were considered [4, 9–11, 20, 34, 48, 50–52].

Late toxicities summary at 3y follow-up visit

	Nipple retraction	Oedema	Arm lymphodema	Telangiectasia	Skin induration
0	512	508	537	520	294
1	39	41	14	35	188
2	3	7	5	2	63
3	0	1	1	0	9
NA's	45	42	42	42	45

Table 1.1: **Late toxicities summary at 3 years follow-up visit.** 5 late toxicities are considered in this thesis and their numerosities in the different classes are reported in the table. The last row counts the missing data for each toxicity.

Late toxicities summary at baseline visit

	Nipple retraction	Oedema	Arm lymphodema	Telangiectasia	Skin induration
0	487	507	541	519	234
1	58	41	14	35	272
2	4	7	1	2	42
3	4	1	0	0	4
NA's	46	43	43	43	47

Table 1.2: **Late toxicities summary at the baseline visit.** 5 late toxicities are considered in this thesis and their numerosities in the different classes are reported in the table. The last row counts the missing data for each toxicity.

1.2.2. Data preprocessing

This thesis project focuses on breast cancer patients with a documented follow-up visit three years after the initial cancer treatment. The cohort considered involves 599 REQUITE patients treated with radiation therapy for breast cancer. Information up to 3 years after the first visit is provided. The clinical goal of the thesis project is to improve the prediction of risk models for late toxicity in breast cancer by introducing genetic information specific to each patient.
The late toxicities are defined starting from PRO forms collected 36 months after the initial cancer treatment, where patients scored each of them between 0 and 3 according to the toxicity gravity. A summary of the late toxicity values is reported in Table 1.1.

Information about the initial toxicity conditions at baseline visit for each patient is provided and summarized in Table 1.2

In clinical trials, two or more binary responses obtained by dichotomizing continuous or

categorical responses are often employed as multiple primary endpoints [28].
Late toxicity at 3y follow-up visit are dichotomized based on suggestions from the clinical counterpart in the RADprecise project, to create six endpoints included as final endpoints in the project:

- $y_i^1 = \{1 \text{ if skin induration of the } i^{th} \text{ patient} \geq 1, \, 0 \text{ otherwise}\}$ for $i \in \{1, ..., 599\}$
- $y_i^2 = \{1 \text{ if skin induration of the } i^{th} \text{ patient} \geq 2, \, 0 \text{ otherwise}\}$ for $i \in \{1, ..., 599\}$
- $y_i^3 = \{1 \text{ if nipple retraction of the } i^{th} \text{ patient} \geq 1, \, 0 \text{ otherwise}\}$ for $i \in \{1, ..., 599\}$
- $y_i^4 = \{1 \text{ if oedema of the } i^{th} \text{ patient} \geq 1, \, 0 \text{ otherwise}\}$ for $i \in \{1, ..., 599\}$
- $y_i^5 = \{1 \text{ if telangiectasia of the } i^{th} \text{ patient} \geq 1, \, 0 \text{ otherwise}\}$ for $i \in \{1, ..., 599\}$
- $y_i^6 = \{1 \text{ if arm lymphodema of the } i^{th} \text{ patient} \geq 1, \, 0 \text{ otherwise}\}$ for $i \in \{1, ..., 599\}$

Each endpoint is then corrected based on its baseline score: if the level of toxicity remains unchanged or decreases between baseline and long-term follow-up then the endpoint for the patient is considered 0; otherwise, the value reported at the follow-up visit is considered.

Endpoints frequency in the sample ranges between 1% and 9% except for y^1 with an occurrence of approximately 47%. Table 1.3 present a summary of the endpoints' occurrence in the sample. Note that the setting of the analysis is strongly unbalanced, stressing the need for techniques able to tackle such issues.

Endpoints

	total	occurrences (perc)	occurrences	nan
y^1	554	47%	260	45
y^2	554	1%	72	45
y^3	554	8%	42	45
y^4	557	9%	49	42
y^5	557	7%	37	42
y^6	557	4%	20	42

Table 1.3: **Summary of the binary endpoints' occurrence in the dataset**, exploited in the project as outcomes of the analysis. The columns present the sample number available for each outcome, the occurrence of the endpoint in the sample with the related percentage, and the subsample for which the endpoint is not available

Endpoints' occurrences in the group receiving the boost dose

		Boost group (%)	
	TOTAL	Endpoint present	Endpoint not present
		(%)	(%)
y^1	554	438 (79%)	
		216 (49%)	222 (51%)
y^2	554	438 (79%)	
		58 (13%)	380 (87%)
y^3	554	439 (79%)	
		31 (7%)	408 (93%)
y^4	557	441 (79%)	
		45 (10%)	396 (90%)
y^5	557	441 (79%)	
		34 (8%)	407 (92%)
y^6	557	441 (79%)	
		20 (5%)	421 (95%)

Table 1.4: **Summary of the endpoints' frequencies in the patient group who received the additional boost dose of radiotherapy**. For each endpoint, the total numerosity and the group numerosity are presented. Additional information on the occurrence of the endpoint in the group is also reported.

Genetic information about the patients is included through Single-Nucleotide Polymorphisms (SNPs). 122 literature-identified SNPs are considered in the analysis. SNPs are usually recorded as a trichotomic categorical variable with values 0, 1, or 2 indicating absence, heterozygosity, or homozygosity of the considered minor allele. Genomic data were imputed to estimate probabilities at variants not genotyped and achieve completeness in genetic information. To better address possible imputation errors, two different datasets are created referred to in the following as $\tilde{\mathbf{D}}$ and $\mathbf{D}$, the first including imputed SNPs as continuous and the second as categorical, rounding imputed SNPs to the closest integer.

In the analysis conducted, two groups of patients are considered based on whether or not they received an additional dose of radiation therapy. Table 1.4 and Table 1.5 describe the numerosity of the two groups across different endpoints including the occurrences of each endpoint in the single groups.

The final datasets resulting from the data preprocessing contain N = 599 triplets (genome, boost, endpoints)

$$\mathbf{D} = \{(\underline{x}_1, z_1, \underline{y}_1), \ldots, (\underline{x}_N, z_N, \underline{y}_N)\}$$

Endpoints' occurrences in the group not receiving the boost dose

		No boost group (%)	
	TOTAL	**Endpoint present**	**Endpoint not present**
		(%)	**(%)**
y^1	554	116 (21%)	
		44 (38%)	72 (62%)
y^2	554	116 (21%)	
		14 (12%)	102 (88%)
y^3	554	115 (21%)	
		11 (10%)	104 (90%)
y^4	557	116 (21%)	
		4 (3%)	112 (97%)
y^5	557	116 (21%)	
		3 (3%)	113 (97%)
y^6	557	116 (21%)	
		0 (0%)	116 (100%)

Table 1.5: **Summary of the endpoints' frequencies in the patient group who didn't receive the additional boost dose of radiotherapy.** For each endpoint the total numerosity and the group numerosity are presented. Additional information on the occurrence of the endpoint in the group is also reported.

and

$$\tilde{\mathbf{D}} = \{(\underline{\tilde{x}}_1, z_1, \underline{y}_1), \ldots, (\underline{\tilde{x}}_N, z_N, \underline{y}_N)\}$$

where for every patient i belonging to $\{1, \ldots, 599\}$, $\underline{\tilde{x}_i}$ belonging to R^{122} is the vector containing values in [0,2] of the 122 SNPs considered; $\underline{x_i}$ belonging to $\{0, 1, 2\}^{122}$ is the vector containing rounded values of the 122 SNPs considered; $\underline{y_i}$ belongs to $\{0, 1\}^6$ and it's the vector describing the presence or absence for each of the 6 endpoints evaluated, while z_i belonging to $\{0, 1\}$ describes the delivery or not of the boost dose.

In the case of analyses conducted separately for each endpoint and separately for each patient group, the input datasets to the model present the form

$$\mathbf{D} = \{(\underline{x}_1, Y_1), \ldots, (\underline{x}_N, Y_N)\}$$

and

$$\tilde{\mathbf{D}} = \{(\underline{\tilde{x}}_1, Y_1), \ldots, (\underline{\tilde{x}}_N, Y_N)\}$$

where for every patient i belonging to $\{1, \ldots, 599\}$, $\tilde{x}_i$ belongs to R^{122} is the vector containing values in [0,2] of the 122 SNPs considered; x_i belonging to $\{0, 1, 2\}^{122}$ is the vector containing rounded values of the 122 SNPs considered; Y_i belonging to $\{0, 1\}$ describes the presence or absence of the specific endpoint considered. Note that since the dataset contains some NAs, for each endpoint, only the observations that are defined were selected.

In the clinical setting, an important need is to model also several endpoints simultaneously. The models considering each output separately give information limited to the possible development of the toxicity selected. It is also useful to have a wider model describing the probability of the presence or absence of generic toxicity considering more than one endpoint as output to the model.

1.2.3. Data complexities in Genome-wide association study

Imbalance

Genome-wide association study (GWAS) has been widely witnessed as a powerful tool for revealing suspicious loci from various diseases. However, real-world GWAS tasks always suffer from the data imbalance problem of sufficient control samples and limited case samples. Unbalance datasets are present more and more in genome association studies as technologies have enabled the collection of thousands of phenotypes from large cohorts, in particular for diseases with low prevalence [19]. The unbalanced binary traits pose serious challenges to traditional statistical methods in terms of both disease prediction [19] and genomic selection causing serious biases to the result and thus leading to losses of significance for true causal markers [8]. More specifically, a very low case-control ratio in GWAS data may violate asymptotic assumptions of statistical inferences. If the number of cases is drastically smaller than the number of controls, these cases may be viewed as outliers in most statistical models, and hence it leads to a higher variation for the estimation of coefficients. As a result, it shrinks the absolute value of test statistics and yields a larger p-value, which makes a truly influential variant insignificant [19]. This results in decreasing the discovery power of conventional GWAS. Specific methodologies need to be implemented to deal with the imbalance and perform robust and interpretable analysis.

Data imputation in genomics studies

Genome-Wide Association Studies (GWAS) are usually performed with DNA array-based approaches that permit the collection of thousands of low-cost genotypes. However, this process comes at the expense of resolution and completeness as SNP chip technologies are ultimately limited by SNPs chosen during array development. An alternative low-cost approach is low-pass whole genome sequencing (WGS) followed by imputation. Rather than relying on high levels of genotype confidence at a set of select loci, low-pass WGS and imputation rely on the combined information from millions of randomly sampled low-confidence genotypes. Results in humans demonstrate that low-pass WGS and imputation provide more accurate genotypes than those imputed using array data, leading to increased power for genome-wide association studies (GWAS) and more accurate risk model prediction [13, 26]. Imputation methods estimate genotype probabilities at variants not genotyped by identifying segments of alleles in an organism that are inherited together from a single parent common to the individuals studied and reference populations of more densely typed individuals [17]. However, the presence of imputed data within association analysis implies the need to introduce specific techniques to deal with errors imputation may cause. In the absence of strategies for study design and data processing, the probability of poor performance and misleading results is unknown. Overall, the imputation process led to decreased significance levels, suggesting that imputation errors may cause statistical significance to be lost for certain experimental configurations [17].

Correct analysis of imputed data calls for the implementation of specific methods which take genotype imputation uncertainty into account. Several techniques can be applied to the analysis of such "uncertain" data. One approach would be to use the genotype with the highest posterior probability for analysis as if it were directly typed but such a procedure could result in biased estimates of the effects. Another approach is based on maximum likelihood: the likelihood can be computed using the total probability formula in which summation is performed over the genotypes, whose true values are not known, but whose posterior probabilities can be estimated given the data. This approach is computationally demanding in the numerical maximization of the likelihood function. Alternatively, a regression approach in which the posterior genotypic probabilities are used as predictors can be applied [5]. Imputation error in the data can be also considered as noise in the input data and tacked via signal processing techniques.

1.3. Literature review: anomaly detection autoencoders and feature selection methods

This section contains the groundwork for the techniques exploited in the thesis work, starting from a general overview of autoencoders to specifics about the methodologies employed to perform feature selection.

1.3.1. Autoencoders general review

A neural network models input-output relations, in terms of functions containing various adaptive parameters whose values are determined using the data in training. It is possible to write such functions in the form $y = y(\underline{x}; \mathbf{W})$ where $\mathbf{W}$ is the parameters vector representing the weights of the network and $\underline{x}$ is the network input. Neural networks offer a very powerful general scheme for representing nonlinear relationships between many input variables and many output variables [15]. A neural network is typically represented by a network diagram in which single units called neurons are structured in layers and each layer output is the input of the next one. To describe neural network functioning, it is convenient to start describing mathematically each neuron. A neuron output can be represented in terms of a non-linear function of a linear combination of the neuron inputs or hidden functions $z_j(\underline{x})$. Suppose we are computing the output of the k^{th} neuron in the layer given that the previous layer contains M neurons:

$$y_k(\underline{x}) = f\left(\sum_{j=1}^{M} w_{kj} z_j(\underline{x})\right) \tag{1.1}$$

where usually $z_0 = 1$ and w_{k0} is called bias. An illustration can be found in Figure 1.2.

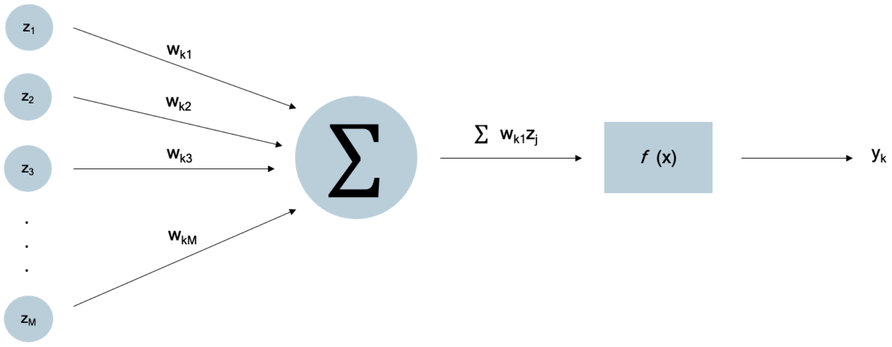

Figure 1.2: **Illustration of the neuron mathematics.** Given the M input, their linear combination with trainable weights is computed and a non-linear function is applied to the weighted sum.

The output of a layer will be the vector of the M' neurons contained output in the layer. The choice of neuron number in each layer depends on the complexity of the problem being treated and on the actual size of the input space when discarding correlated inputs, called the intrinsic size. The final output of one hidden-layer neural network with M hidden neurons, d inputs ($\underline{x}$) and one output (y), illustrated in Figure 1.3, is described as:

$$y(\underline{x}) = f\left(\sum_{k=0}^{M} \tilde{w}_k g\left(\sum_{j=0}^{d} w_{kj} x_j\right)\right) \tag{1.2}$$

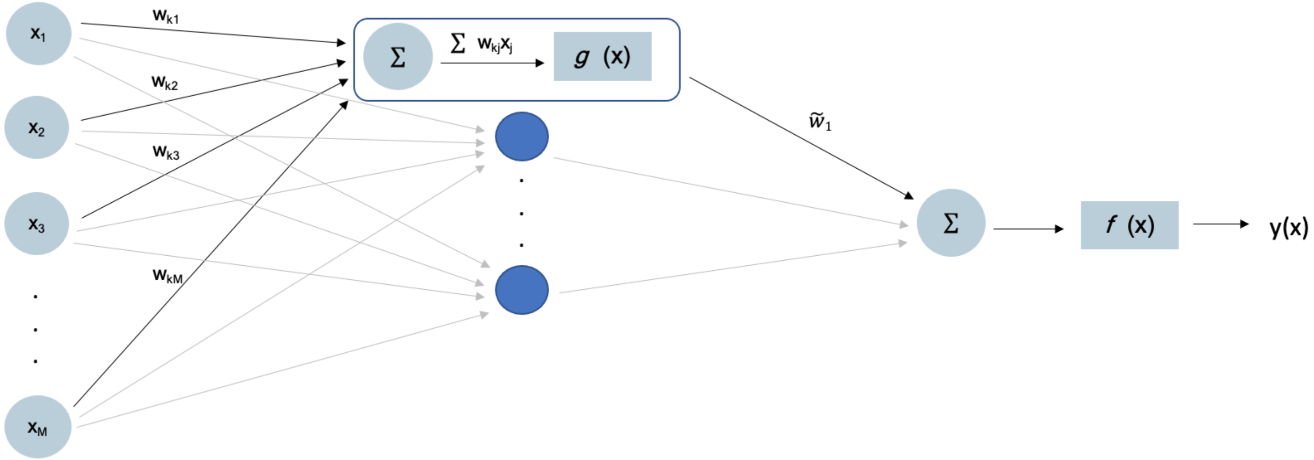

Figure 1.3: **Illustration of one hidden-layer neural network.** For each hidden neuron the output is computed as in 1.1 and the same formula is applied to compute the output layer neuron result considering as input the neurons' outputs of the previous (hidden) layer

The weights $\mathbf{W}$ are adjusted over the training to best represent the input-output relation, based on a context-dependent loss which allows quantifying the goodness of the network and therefore have a response on the validity of the set weights. Training a Neural Network means finding the appropriate weights that minimize the loss function. Tools of infinitesimal calculus are employed to find the minimum of the error function in the Gradient Descent propagation algorithm, which is usually exploited in the training.

An Autoencoder is a neural network trained to copy its input to its output. Autoencoders are used for data reconstruction in unsupervised learning. Let the matrix $\mathbf{X} \in R^{NxJ}$ be the input data, $\mathbf{X} = \{\underline{x}_1, ..., \underline{x}_N\}$ set of N training vectors $\underline{x}_i$ $(i \in \{1, ..., N\})$, characterized by J features. The simplest version of an AE is constituted by a single hidden layer with H neurons between the input and output layers which count J neurons each. The hidden layer has usually fewer neurons than the input and output layer $(H <<< J)$ generating a latent representation of the input in a space of reduced dimension. Mimicking the identity function, the Autoencoder learns an encoded version of the data compressing and aggregating information in input, in the best way for the network to reconstruct the original information from the latent representation.

The network can be seen as constituted by two parts: an encoder and a decoder. The encoder and decoder functions can be represented, with reference to neural network definition as: $\underline{h}_i = f(\mathbf{W}\underline{x}_i + \underline{b})$ in the encoder that map each input vector $\underline{x}_i$ into an encoded version of itself of size H; and $\underline{\overline{x}_i} = g(\mathbf{W}'\underline{h}_i + \underline{b}')$ in the decoder that maps back the latent representation vector to the J-dimensional space. An illustration of the one hidden-layer autoencoder can be found in Figure 1.4.

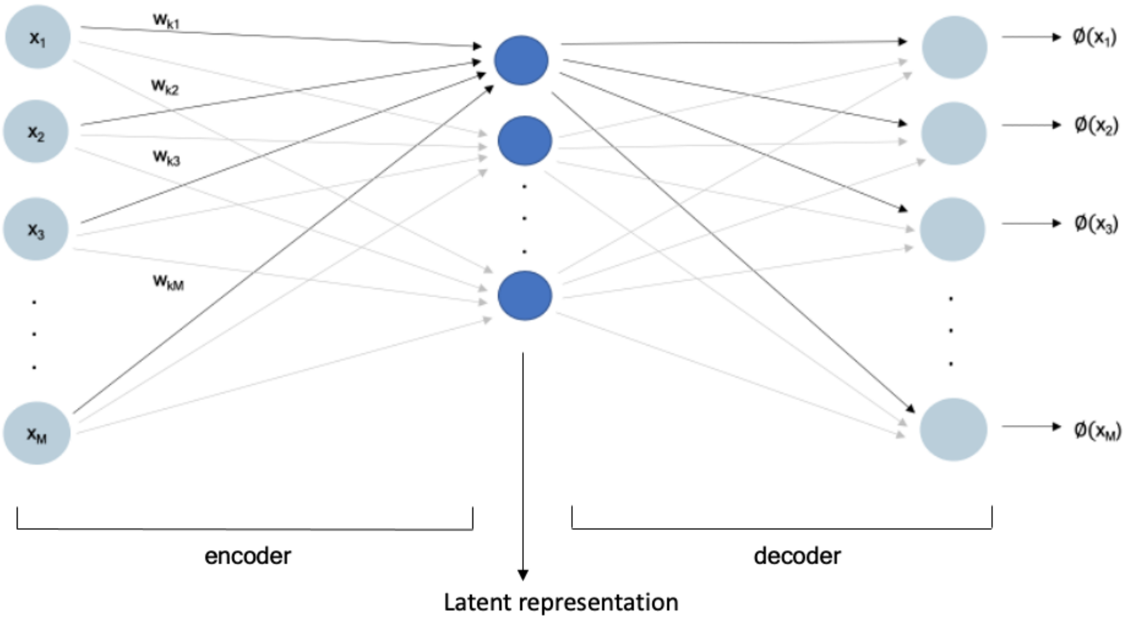

Figure 1.4: **Illustration of a one hidden layer Autoencoder.** The encoder maps each input vector x_i into an encoded version of itself of size H; and the decoder maps back the latent representation vector to the J-dimensional space.

The model is trained through gradient descent of the loss function $L(\underline{x}, \underline{\overline{x}})$; where L is typically the mean squared reconstruction error (MSRE) for continuous features, that is, the mean squared Euclidean distance between the input values and the reconstructed values for each observation, and L is typically a cross-entropy for categorical variables that measure the difference between input and reconstructed values probability distributions. Minimizing the cross-entropy corresponds to maximizing the likelihood with respect to parameter $\mathbf{W}$.

In general, we can define an AE as a map $\phi(\underline{x}_i) : R^J \to R^J$, such that

$$\phi(\underline{x}_i) = g(\mathbf{W}' f(\mathbf{W}\underline{x}_i + \underline{b}) + \underline{b}')$$

and the weights are optimized so that the reconstruction $\underline{\overline{x}_i} = \phi(\underline{x}_i)$ is as close as possible, considering the loss, to $\underline{x}_i$. Suppose now that the first K features are continuous while the last J-K are binary, the optimal representation of $\underline{x}_i, \underline{\overline{x}_i}$, is given by:

$$\overline{\phi(\underline{x}_i)} = argmin_{\phi} L(\underline{x}_i, \phi(\underline{x}_i))$$

$$L(\underline{x}_i, \phi(\underline{x}_i)) = \left(\sum_{j=1}^{K} (x_{ij} - \phi(x_{ij}))^2 + \sum_{j=K+1}^{J} (x_{ij} * log(\phi(x_{ij})) + (1 - x_{ij}) log(1 - \phi(x_{ij})) \right) \tag{1.3}$$

$$\tag{1.4}$$

Autoencoders typically do not provide exact reconstruction since $H <<< J$ but the latent representation is expected to be meaningful and a compact representation of the input. [39]. Better representations can be achieved using multiple hidden layers and constraints that force autoencoders not only to replicate the input but to learn effective representations of such input in the hidden layer. A common way to achieve this goal is to introduce sparsity in the most internal layer of the deep architecture by adding a regularization term in the loss function (1.4) :

$$L^S(\underline{x}_i, \phi(\underline{x}_i)) = L(\underline{x}_i, \phi(\underline{x}_i)) + \omega(h(l)) \tag{1.5}$$

The regularization can take various forms. In a deep sparse autoencoder (DSAE), the L_1 penalization is applied on the activation of the most internal hidden layer h(l) and it is controlled by a parameter λ, that is:

$$L^S(\underline{x}_i, \phi(\underline{x}_i)) = L(\underline{x}_i, \phi(\underline{x}_i)) + \lambda |h(l)| \tag{1.6}$$

The parameter λ is usually optimized through grid search. This penalization term forces the model to activate the minimum number of hidden nodes to reconstruct the input reducing the need for tailored choices or expensive optimization to define the proper architecture and incrementing generalization propriety of the model [39].

1.3.2. Anomaly detection autoencoders

Autoencoders are used for learning data representations, dimensionality reduction, and anomaly detection. An anomaly is a data point that is significantly different from the remaining data and arouses suspicions that it was generated by a different mechanism [25]. Among many anomaly detection methods, spectral anomaly detection techniques try to find the lower dimensional embeddings of the original data where anomalies and normal data are expected to be separated from each other. After finding those lower dimensional embeddings, they are brought back to the original data space. Reconstructed data are expected to contain only the true nature of the data, without uninteresting features and noise. Given a set of data, an autoencoder learns to minimize the reconstruction error on this set and, thanks to its generalization probability, it should be able to reconstruct effortlessly points close to its original input. Autoencoder-based anomaly detection methods are deviation-based. That is, in a semisupervised learning setting, they exploit the reconstruction error as the anomaly score. In particular, one-class detection AEs, are trained exclusively on normal observations so that the AE will reconstruct normal data very well while failing to do so with anomaly data that has not been encountered before. Data points with high loss are considered to be anomalies. A pseudo-algorithm of an autoencoder-based anomaly detection procedure is presented in Algorithm 1.1 [2].

Algorithm 1.1 Autoencoder-based anomaly detection

INPUTS

- **X**: Training dataset;
- x_i $i = \{1, ..., N\}$ Test dataset;
- Threshold Δ.

```
$\phi$, **W**, **W'** $\leftarrow$ train an autoencoder using the normal dataset **X**
for $i = 1, ...., N$ do
    $RE_i$ $\leftarrow$ evaluate reconstruction error as in (1.4)
    if $RE_i > \Delta$ then
        $x_i$ is an anomaly
    else
        $x_i$ is not an anomaly
    end if
end for
```

In the algorithm, the autoencoder is trained only on normal data and tested on an independent population containing both normal and anomaly data. For each of the test set samples, the reconstruction error of the autoencoder is evaluated and, given a user-defined threshold, the sample is classified as an anomaly or not.

1.3.3. Denosing autoencoders

An additional consideration should be pointed out for the context of uncertain data, such as that of imputed genotype. Imputation error can be also considered as noise in the input data and tacked via signal processing techniques. Noise reduction techniques are needed in all those analyses where real-world data are corrupted by noise and outliers. Denoising autoencoders are trained to reconstruct noise-free corrupted versions of their inputs. They can be viewed either as a regularization option or as robust autoencoders which can be used for error correction [7]. In these architectures, the input is disrupted by some noise (e.g., additive white Gaussian noise) and the autoencoder is expected to reconstruct the clean version of the input [7], as illustrated in Figure 1.5

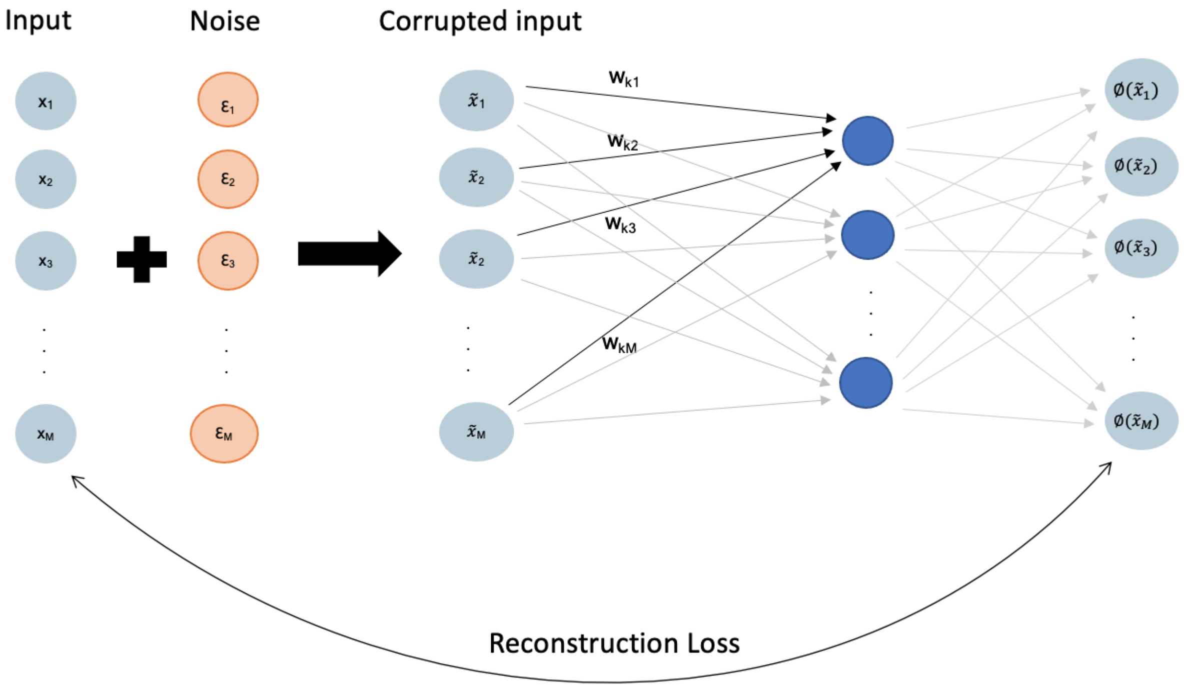

Figure 1.5: **Illustration of a one-layer denoising autoencoder**. The input $\underline{x}$ is disrupted by some noise $\underline{\epsilon}$ and the autoencoder is expected to reconstruct the original input.

General noise reduction techniques in autoencoders and machine learning problems are detailed in [24].

1.3.4. Feature selection in unbalanced datasets with covariates characterized by high-order interactions

To perform a solid and effective classification and consequently, create robust prediction models, it is essential to consider only highly influential features that provide intrinsic information and discriminant property for class separability. Feature selection aims at

doing so, while decreasing computational costs, aiding inference, and giving a better understanding of the model representation.
Features selection is key in clinical cases where the variables are often many and correlated with each other, and became essential within genomic studies where the covariates are often several hundred and the curse of dimensionality plays an important role. Oftentimes genetic studies, in addition to a large number of covariates, are characterized by a relatively limited amount of data. High variance and overfitting are major concerns in this setting since not enough information is present in the relatively small number of samples to efficiently estimate a high-dimensional covariance matrix or a large number of model parameters. As a result, feature selection and model penalization need to be introduced. However, feature selection in genetic applications is made difficult by the presence of unbalanced classes and the importance of identifying the discriminant characteristics of the minority class, where an inaccurate feature selection can lead to an inaccurate diagnosis. On top of the limited amount of data and unbalanced setting, there is also increasing awareness that epistasis, or gene-gene complex interactions, are more important than the effects of any single genetic variant so the feature selection method needs to take into account the effect of high-order interactions in selecting important covariates [22]. Finally, domain experts are oftentimes interested in understanding which specific features should be kept under control, requiring a simple and interpretable feature selection model [39].

Traditional FS techniques become sub-optimal or even prejudicial to classification effectiveness when the classes are strongly imbalanced and usually just consider the main effect of covariates in performing the selection.

Autoencoders have been recently exploited for reconstruction-based FS in many approaches. Some examples can be found in [18, 21, 47, 55]. Autoencoders models for feature selection are a mixture of two different statistical problems: density estimation, the objective of which is to model the unconditioned distribution of data in an unsupervised manner, and classification, the objective of which is to model the conditional distribution of target classes based on inputs in a supervised way. The goal is to produce a data-based model approximating the distribution of data conditioned to classification labels. Typical autoencoder-based FS share an unsupervised setting and have demonstrated their potential as feature selectors against other state-of-the-art techniques. On the other hand, they usually are approaches designed for balanced classification. This balanced selection of features was argued as potentially harmful in strongly imbalanced settings.
Given the clinical setting, it is fundamental that the FS process is not affected by these problems.

1.3.5. Application of anomaly detection autoencoder to feature selection in imbalanced settings and interaction-aware PRS algorithm

This thesis starts from an innovative feature selection methodology for binary outcome in imbalance settings presented in [39] and aims at widening its applicability to breast cancer late toxicities and generalizing it to a multi-endpoint framework. The FS presented in [39] is a combination of anomaly detection autoencoders and FS autoencoders and was developed specifically to be efficient in the case of unbalanced classes. To achieve FS advantages in this setting, it has been proposed a filtering FS algorithm, ranking feature importance based on autoencoder reconstruction error. Exploiting the ability of the autoencoder to differentiate between samples generated from different mechanisms, likewise anomaly detection autoencoders, the method's final aim is selecting features able to distinguish between the different classes.

The FS method presented in [39] was assembled in a wider method presented in [22] whose purpose was to analyze prostate cancer patients within the REQUITE study to identify the effect of different single nucleotide polymorphisms (SNPs) and their interactions on the risk of post-radiotherapy (RT) toxicity; its methodological aim was to propose a new method for polygenetic risk scoring that incorporates SNP-SNP high order interactions (hiPRS) to maximize the predictive power of risk models and performing an interpretable FS, characterizing the most important SNPs also basing on their synergy. Further details on hiPRS algorithm can be found in [22].

In this section, the work within [22] is briefly explained by presenting the methodology pipeline and main innovations, focusing on the FS section.

The pipeline can be divided into two parts. The first part exploits a Deep Sparse Autoencoder Ensemble to perform feature selection, and the second part presents hiPRS algorithm in which the most important high-order interactions are selected and the PRS is defined.

In the first part of the pipeline, a Deep Sparse AutoEncoder Ensamble (DSAEE) is exploited to perform feature selection and create a subset of influential SNPs related to the specific endpoint in consideration. Each DSAE is trained to characterize the set of patients not presenting the endpoint. The set of patients presenting the toxicity will instead be referred to as minority or case sample. This set will be referred to in the following as the majority class or control set. Accordingly, the training set consists only of patients without toxicity (overrepresented class), and the model is tested on a mixed independent population that includes patients with and without toxicity (underrepresented class). The algorithm ranks feature importance based on the Reconstruction Error of the test set. From the analysis of the aggregated Reconstruction Error, the features where the minority class presents a different distribution of values w.r.t. the overrepresented one are identified, thus selecting the most relevant features to discriminate between the two. The DSAE pseudo-algorithm is presented in Algorithm 1.2

Algorithm 1.2 DSAEE for Minority Class Feature Selection

INPUTS

- **X**: Majority class covariate dataset with K features;
- **X'**: Minority class covariate dataset with K features;
- B: Enasamble iterations;
- α: Threshold (quantile).

OUTPUT Feature set F.

```
minority sample size ← number of **X'** rows
Set $\mathbf{Q}$ // empty array
**for** $i = 1, ...., B$ **do**
    $\mathbf{X_{test}} \leftarrow$ Sample minority sample size observations from $\mathbf{X}$
    $\mathbf{X_{train}} \leftarrow \{\underline{x}_i \in \mathbf{X} | \underline{x}_i \notin \mathbf{X_{test}}\}$
    $\mathbf{X_{test}} \leftarrow concatenate(\mathbf{X_{test}}, \mathbf{X'})$
    Define $\underline{y}_{test}$ containing class belongings of $\mathbf{X_{test}}$
    $\phi$, **W**, **W'** ← train a autoencoder using the normal dataset $\mathbf{X_{train}}$ considering (1.6)
    $\mathbf{R} \leftarrow$ evaluate reconstruction error on $\mathbf{X_{test}}$ as in (1.4)
    Label $\mathbf{R}$ with $\underline{y}_{test}$
    $\mathbf{Q} \leftarrow concatenate(\mathbf{Q}, \mathbf{R})$
**end for**
$\mathbf{Q_{maj}}, \mathbf{Q_{min}} \leftarrow \{re_j \in \mathbf{Q} | y_{test,j} = 0\}, \{re_j \in \mathbf{Q} | y_{test,j} = 1\}$
$mean\_re_{maj}$ , $mean\_re_{min}$ ← Column mean of $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$
$\Delta \leftarrow mean\_re_{min}$ - $mean\_re_{maj}$
F ← $\{k | \Delta_k > \Delta_\alpha, k \in \{1, ..., K\}\}$
```

This methodology is exploited because of the inherent hierarchical structure of the autoencoder where each layer performs a non-linear combination of previous ones. This is particularly suited to mimic the complex dependencies within the data. The approach uses a representation learning technique to obtain the best representation of the majority class (healthy patients in this dataset) and consequently identify which SNPs distinguish the minority class (unhealthy patients) from the majority class.

This algorithm presents three major benefits: the first is the ability to deal with heavy class imbalance and robustify the selection thanks to its ensemble approach [37, 39]; the second one is the interaction consideration, with autoencoder intrinsic hierarchical structure; the third one lies in the increased interpretability of the subsequent algorithms and results. Indeed, identifying features that are the most informative, w.r.t. a target class within a dataset is insightful information by itself in many application contexts [39].

In the second part of the pipeline hiPRS algorithm is presented. It is based on a simple and interpretable model that succeeds in including high-level interactions among SNPs in the calculation of the Polygenic Risk Score. The procedure is displayed in Figure 1.6.

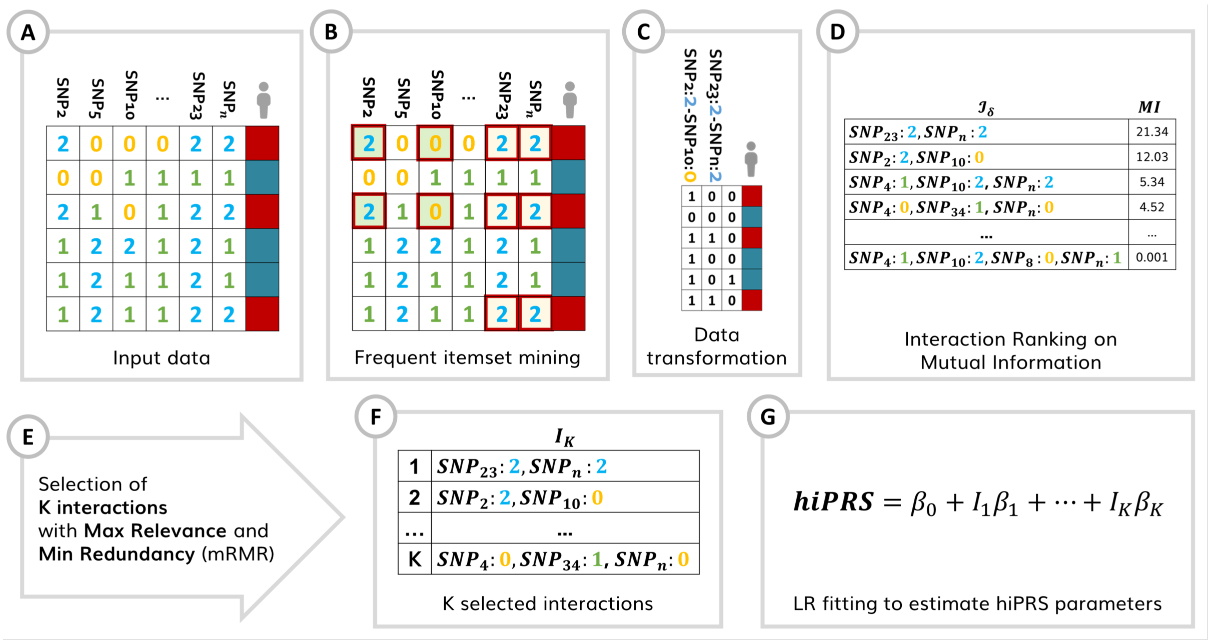
A
SNP2
SNP5
SNP10
...
SNP23
SNPn
Input data
B
Frequent itemset mining
C
SNP23:2-SNPn:2
SNP2:2-SNP10:0
Data transformation
D
$\mathcal{I}_\delta$
MI
SNP_{23}: 2, SNP_n: 2 — 21.34
SNP_2: 2, SNP_{10}: 0 — 12.03
SNP_4: 1, SNP_{10}: 2, SNP_n: 2 — 5.34
SNP_4: 0, SNP_{34}: 1, SNP_n: 0 — 4.52
...
SNP_4: 1, SNP_{10}: 2, SNP_8: 0, SNP_n: 1 — 0.001
Interaction Ranking on Mutual Information
E
Selection of K interactions with Max Relevance and Min Redundancy (mRMR)
F
I_K
1 SNP_{23}: 2, SNP_n: 2
2 SNP_2: 2, SNP_{10}: 0
...
K SNP_4: 0, SNP_{34}: 1, SNP_n: 0
K selected interactions
G
$hiPRS = \beta_0 + I_1\beta_1 + \cdots + I_K\beta_K$
LR fitting to estimate hiPRS parameters

Figure 1.6: **Illustration of hiPRS algorithm workflow.** Starting with a predefined set of SNPs of interest *(A)*, hiPRS exploits FIM (Frequent Itemset Mining) routines to create a list of possible significant interactions *(B)*. These terms are then ranked according to their relationship to the endpoint in terms of Mutual Information (MI) *(D)*. A limited number of user-defined interactions are then selected via Minimum Redundancy Maximum Relevance (mRMR) algorithm *(E)* for inclusion in the final PRS *(F and G)*. Source: [38]

hiPRS treats data at the genotype level and, starting with a predefined set of SNPs of interest *(A)*, exploits FIM (Frequent Itemset Mining) routines to create a list of possible significant interactions. This is achieved considering only patients with toxicity and evaluating interactions that, within the class, have a high empirical frequency *(B)*. These terms are then ranked according to their relationship to the endpoint in terms of Mutual Information (MI) *(D)*. A limited number of user-defined interactions are then selected for inclusion in the final PRS *(F and G)*. To this end, hiPRS exploits a selection algorithm similar to Minimum Redundancy Maximum Relevance (mRMR): the algorithm selects terms through greedy optimization of the ratio between MI (relevance) and a suitable similarity measure for interactions (redundancy) *(E)*. This leads to a set of predictive, yet diverse, interactions that are used to define the score. In the end, the latter is built by weighting the contribution of each interaction term accordingly to the weights obtained when fitting a logistic regression model with the selected endpoint as outcome *(G)* [38].

1.4. Aim of the study

This thesis proposes an innovative methodology capable of performing variable selection in high-dimensional contexts where high-order interactions are of interest and multiple outcomes are simultaneously studied. The feature selection accounts for intra-outcome correlation structure and properly defines a set of significant features for a comprehensive endpoint that summarizes the multivariate response, aggregating individual results. The multivariate FS is developed specifically to work on genomic data. The algorithm is robust to data imputation and suitable for multiple binary classification problems with high imbalance in the classes of each outcome. The selection can be exploited to efficiently include genetic effects in clinical risk models.

To this aim techniques of Machine Learning, Deep Learning, signal processing method, and non-parametric statistics are exploited.

Each of the fundamental aspects of the feature selection methods is individually tackled, developing a specific methodology and simulation setting to properly understand its characteristics. A final comprehensive methodology that encapsulates the single methodologies developed is presented and applied to the case study of REQUITE data.

2 | Methodologies

This Chapter presents the methodologies developed in order to achieve the aim of the study and tackle all the problems described in Chapter 1. Starting from the method reported in Section 1.3.5 in order to achieve robustness to imputation error signal processing methods are assimilated to deep learning techniques: a denoising procedure is introduced in the training process of the DSAE so that the imputation error is considered in the reconstruction error of both groups, leading to an unbiased analysis. Selection of the discriminating feature between the classes is achieved by comparing the reconstruction error distributions available from the ensemble learning procedure. In particular, a non-parametric test is exploited to assess the distribution difference in the reconstruction error between the classes. This methodology selects all the features able to distinguish between them. Multi-outcome feature selection is achieved by the proper definition of DSAEE groups, accounting for correlation in the endpoints. Feature selection is performed starting from a unique control sample, extracted from the intersection of each endpoint majority class, and an endpoint-depending case sample. The selection is able to produce a set of SNPs describing each specific toxicity accounting for the dependency structure in the multivariate outcome, increasing the statistical power of the model.
Finally, a comprehensive methodology that implements a multi-outcome feature selection specifically developed in the context of clinical risk models and genomic data is presented. The selected SNPs for each outcome can then be combined to form an informative set of features correlated with general toxicity and able to distinguish generally radiosensitive patients.

2.1. Anomaly detection autoencoders for feature selection with imputed data

The first part of the chapter is devoted to the application of a denoising procedure to the DSAE to achieve robustness to imputation error and enable an unbiased analysis. The presence of imputed data within association analysis implies the need to introduce specific techniques to avoid poor performance and misleading results. The technique introduced in this section is inspired by denoising autoencoders. The hypothesis is that, as autoencoders are able to reconstruct corrupted input error, similarly it is possible to force them to reconstruct noisy continuous input data into accurate categorical data.

A binary supervised learning setup is considered with an available set of N (input, target) pairs

$$\tilde{\mathbf{D}} = \{(\underline{\tilde{x}_1}, Y_1), ..., (\underline{\tilde{x}_N}, Y_N)\}$$

where Y_i is the target that takes values in $\{0,\ 1\}$ and $\underline{\tilde{x}_i} \in R^J$ with $i = 1, ..., N$ is the input feature vector of imputed data or, in general, noisy data. Suppose that $\underline{x_i}$ true categorical feature vector is known for each patient present in the training set and that a fixed number of M categories is available for each feature. Therefore a second dataset is available with N (input, target) pairs

$$\mathbf{D} = \{(\underline{x_1}, Y_1), ..., (\underline{x_N}, Y_N)\}$$

where Y_i is the target that takes values in $\{0,\ 1\}$ and $\underline{x_i} \in \{1, ..., M\}^J$ with $i = 1, ..., N$ is the input feature vector of categorical data. If the true categorical feature vector is unknown, it is possible to simply round each imputation to the closest integer and consider it the categorical representation of the data. The accuracy of this representation is not fundamental, since the algorithm works so that the possible inaccuracy is considered. Finally, suppose that imbalance in the class is present, with a minority class $Y = 1$ (case class) and a majority class $Y = 0$ (control class). The rationale of the DSAEE follows in general the one presented in [38] with the intention of including imputation noise in the reconstruction process. The procedure rationale is detailed in the following and schematized in Algorithm 2.1.

In particular, **X** is defined as the set of categorical features related to the majority class and $\tilde{\mathbf{X}}$ is defined as the set of continuous features related to the majority class. Analogously, **X'** is defined as the set of categorical features related to the minority class, and $\tilde{\mathbf{X}}'$ is defined as the set of continuous features related to the minority class.
In each ensemble iteration b $\in \{1, ..., B\}$ a test set containing 2 * O data points, where O is the minority class numerosity, is constructed by concatenating all the minority class patients with a random sample of the same size from the majority class. The remaining observations of the majority class are included in the training set. Note that for both the training and the test set two datasets are available, respectively the one with continuous and categorical features. This train-test set structure allows training each DSAE learner in an unsupervised fashion only on the overrepresented population, and to test its performance when facing both majority and minority class examples, so that comparison of

the RE in two populations is possible.

Algorithm 2.1 DSAEE for Minority Class Feature Selection with imputed data

INPUTS

- $\mathbf{D}$: categorical covariate dataset;
- $\tilde{\mathbf{D}}$: continuous covariate dataset;
- B: Enasamble iterations;
- δ: Threshold (quantile).

OUTPUT Feature set F.

```
$\mathbf{X} \leftarrow \{\underline{x_i} \in \mathbf{D} | Y_i = 0\}$
$\tilde{\mathbf{X}} \leftarrow \{\underline{\tilde{x}_i} \in \tilde{\mathbf{D}} | Y_i = 0\}$
$\mathbf{X'} \leftarrow \{\underline{x_i} \in \mathbf{D} | Y_i = 1\}$
$\tilde{\mathbf{X}}' \leftarrow \{\underline{\tilde{x}_i} \in \tilde{\mathbf{D}} | Y_i = 1\}$
O ← minority sample size or number of $\mathbf{X'}$ rows
Set $\mathbf{Q}$ // empty array
for $i = 1, ...., B$ do
    $\mathbf{X_{test}} \leftarrow$ Sample O observations from $\mathbf{X}$
    $\tilde{\mathbf{X}}_{\mathbf{test}} \leftarrow$ Sample the same observations in $\mathbf{X_{test}}$ from $\tilde{\mathbf{X}}$
    $\mathbf{X_{train}} \leftarrow \{\underline{x_i} \in \mathbf{X} | \underline{x_i} \notin \mathbf{X_{test}}\}$
    $\tilde{\mathbf{X}}_{\mathbf{train}} \leftarrow \{\underline{\tilde{x}_i} \in \tilde{\mathbf{X}} | \underline{\tilde{x}_i} \notin \tilde{\mathbf{X}}_{\mathbf{test}}\}$
    $\mathbf{X_{test}} \leftarrow concatenate(\mathbf{X_{test}}, \mathbf{X}')$
    $\tilde{\mathbf{X}}_{\mathbf{test}} \leftarrow concatenate(\tilde{\mathbf{X}}_{\mathbf{test}}, \tilde{\mathbf{X}}')$
    Define $Y_{test}$ containing class belongings of $\mathbf{X_{test}}$
    $\phi$, $\mathbf{W}$, $\mathbf{W'}$ ← train a deep sparse autoencoder using the normal dataset $\tilde{\mathbf{X}}_{\mathbf{train}}$ as
    input and $\mathbf{X_{train}}$ as output considering (2.3)
    $\mathbf{R}$ ← evaluate reconstruction error on $\tilde{\mathbf{X}}_{\mathbf{test}}$ and $\mathbf{X_{test}}$ as in (2.2)
    Label $\mathbf{R}$ with $Y_{test}$
    $\mathbf{Q} \leftarrow concatenate(\mathbf{Q}, \mathbf{R})$
end for
$\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$ ← $\{re_j \in \mathbf{Q} | Y_{test,j} = 0\}$, $\{re_j \in \mathbf{Q} | Y_{test,j} = 1\}$
$mean\_re_{maj}$ , $mean\_re_{min}$ ← Column mean of $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$
$\Delta \leftarrow mean\_re_{min}$ - $mean\_re_{maj}$
F ← $\{j | \Delta_j > \Delta_\delta, j \in \{1, ..., J\}\}$
```

The rationale behind this sampling procedure is the same used for outliers detection: since the DSAE is trained to reconstruct normal observations only, it will make higher RE when tested on outlier observations never experienced during training [39]. Moreover, the DSAE is trained to reconstruct the continuous input into a categorical output, so that the possible error due to imputed data is accounted for in the comparison. The input data are $\mathbf{\tilde{X}_{train}}$ where each of the J features is considered numerical and the output layer returns the probability distribution over the M categories in the categorical covariate. The network weights and reconstruction map are optimized to have the best possible representation and reconstruction of the J features exploiting the loss in (2.3).

$$\overline{\phi(\underline{\tilde{x}})} = argmin_{\phi} L(\underline{x}, \phi(\underline{\tilde{x}})) \tag{2.1}$$

$$L(x_j, \phi(\tilde{x}_j)) = -\sum_{k=0}^{M} (x_{jk} * log(\phi(\tilde{x}_{jk}))) \quad for \;\; j \in \{1, ..., J\} \tag{2.2}$$

$$L^S(\underline{x}, \phi(\underline{\tilde{x}})) = \sum_{j=1}^{J} L(x_j, \phi(\tilde{x}_j)) + \lambda |h(l)| \tag{2.3}$$

where $\underline{\tilde{x}} \in \mathbf{\tilde{X}_{train}}$ and $\underline{x} \in \mathbf{X_{train}}$. The loss is composed of two different parts, the first one being a cross-entropy that evaluates the difference in the probability distribution between the autoencoder outcome and the one-hot encoding of the corresponding categorical covariate. The second part instead is the L1 loss to induce sparsity in the latent representation of the covariates, improving the generalization ability of the model.

Once the network has been trained, the reconstruction error is evaluated on each sample of the test set as in (2.2). For each observation of the test set, the outcome is composed of J probability distributions over the M categories describing the likelihood of each feature belonging to each class.

$$\phi(\mathbf{\tilde{X}_{test}}) \in R^{(J,M)}$$

Since the autoencoder was trained exclusively on observation from the majority class, their representation is expected to be optimal, while the representation and consequently the reconstruction of minority class samples are expected to be less efficient:

$$\sum_{j=1}^{J} (L(x_j | Y = 1, \phi(\tilde{x}_j))) \geq \sum_{j=1}^{J} (L(x_j | Y = 0, \phi(\tilde{x}_j)))$$

where $\underline{x} \in \mathbf{X_{test}}$ and $\underline{\tilde{x}} \in \mathbf{\tilde{X}_{test}}$.

The most discriminant feature should present the highest reconstruction error difference in the two groups, or, more specifically, the RE distribution of the most discriminant feature should be statistically different in each class $re_j|Y = k \sim f_j^k$ with $k \in \{0, 1\}$ and $j \in \{1, ..., J\}$ and $f_0 \nsim f_1$.

The test set REs from each ensemble repetition are concatenated in Q and at the end of the ensemble procedure $\mathbf{Q} \in R^{(B*2*O,J)}$. Q is then split in $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$ based on the belonging of each observation in the test set.

For each feature j, the RE difference in the two groups is evaluated by taking the mean over all the observations in $\mathbf{Q_{maj}}$ and $\mathbf{Q_{min}}$.

$$l_j^b = \frac{\sum_{i=1}^{O} L(x_j^i|Y_i = 1, \phi(\tilde{x}_j^i)) - \sum_{i=1}^{O}(L(x_j^i|Y_i = 0, \phi(\tilde{x}_j^i))}{O}, \quad j \in J, b \in B \tag{2.4}$$

$$\Delta_j = \frac{\sum_{b=1}^{B} l_j^b}{B}, \quad j \in J \tag{2.5}$$

Averaging is performed also over ensemble repetition to achieve higher robustness in the selection method.

Finally, the features are ranked based on (2.5) in decreasing order. The highest-ranked features are those accurately reconstructed by the DSAE on the majority class (lower RE), and poorly reconstructed on the minority class (higher RE). To identify an exact set F, a threshold $\delta \in (0, 1)$ is defined. Δ_δ is the δ-th quantile evaluated on the distribution of $\{\Delta_j\}_{j=1}^{J}$. Then, all those features j whose average RE difference is above the user-defined quantile are selected:

$$F = \{j|\Delta_j > \Delta_\delta, j \in \{1, .., J\}\} \tag{2.6}$$

This is the same final selection procedure used in [38]. Note that there is an inverse relation between δ and the number of selected features: the higher the δ, the lower the number of selected features.

2.2. Distribution-based methodology for feature selection involving ensemble learning

In this section, a new selection methodology from the Reconstruction Error (RE) of the autoencoder is developed based on the possibility of simulating the distribution of the RE in the two groups thanks to the ensemble learning procedure. This method is introduced specifically in the context of feature selection performed with anomaly detection autoencoders and the rationality beneath the method is to select as discriminating features those presenting statistically different RE distributions in the two groups. Given the ensemble nature of the Algorithm 2.1, it is possible to consider the RE distributions evaluated on the test set as the best approximation available of its real value in the population. Indeed, the idea of ensemble learning is to build a prediction model by combining the strengths of a collection of simpler base models to reduce the variance of an estimated prediction function. This provides a way not only to estimate parameters and perform FS but also to approximate a sampling distribution, in this case, the RE in each group. Instead of simply summarizing the information coming from different ensemble iterations in a mean, it is possible to compute tests on the sampled distributions and generalize the results to the entire population. Feature Selection can then be performed by picking features presenting a statistically larger RE distribution in the minority class than in the majority class.

The feature selection from autoencoder reconstruction error is performed as reported in Algorithm 2.2 and explained in the following.

Algorithm 2.2 Ensamble-based feature Selection

INPUTS

- **Q**: matrix of the reconstruction error as computed in Algorithm 1.2 ($\mathbf{Q} \in R^{(B*2*O,J)}$).

OUTPUT Feature set F.

```
$\mathbf{Q_{maj}}$, $\mathbf{Q_{min}} \leftarrow \{re_j \in \mathbf{Q} | Y_{test,j} = 0\}$, $\{re_j \in \mathbf{Q} | Y_{test,j} = 1\}$
set F //empty set
for j in J do
    $p_j \leftarrow$ p-value of the Smirnov test,
    testing the difference in $\mathbf{Q_{maj}}$ and $\mathbf{Q_{min}}$ jth column
    if $p_j < \frac{0,05}{J}$ then
        Include j in F
    end if
end for
```

In particular, the matrix of reconstruction errors $\mathbf{Q} \in R^{(B*2*O,J)}$ computed as reported in Algorithm 1.2 splits in $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$ based on the belonging of each observation in the test set. Each of the $B * O$ observations in the two is considered as a sample extracted from each feature distribution $re_j | Y = k \sim f_j^k$ with $j \in \{1, ..., J\}$ and $k \in \{0, 1\}$. It is possible then to compare the samples and test if the feature distribution in different groups is statistically different. The analysis is performed via the Smirnov test. The Smirnov test is a non-parametric two-sample test, used to determine if two independent random samples appear to follow the same distribution. Let $\underline{x} : (x_1, x_2, ..., x_m)$ and $\underline{y} : (y_1, y_2, ..., y_n)$ be independent random samples of size m and n, respectively, from continuous or ordered categorical populations with CDFs of F and G, respectively. The null hypothesis of the test is the equality of distribution functions:

$$H_0 : F(t) = G(t), \ \ \forall t \in R$$

This null hypothesis can be tested against the two-sided alternative hypothesis:

$$H_1 : \exists t \in R : F(t) \neq G(t)$$

or it can be tested against a one-sided alternative hypothesis:

$$H_1 : F(t) \geq G(t) \ \ \forall t \in R \ \ and \ \ \exists t \in R : F(t) > G(t)$$

or

$$H_1 : F(t) \leq G(t) \ \ \forall t \in R \ \ and \ \ \exists t \in R : F(t) < G(t)$$

To compute the Smirnov test statistic, the empirical CDFs for the x and y samples are computed and the difference between the two distributions can be measured by the signed differences or by the absolute value of the difference depending on H_1 for every t. The maximum value of the selected difference between the two empirical cumulative distribution functions is then used as the test statistic [12]. For more information about the test refer to [12].

Once the test is performed, set F includes all the features whose test p-value is lower than the Bonferroni corrected threshold of 0.05.

The F set computed with Algorithm 2.2 can include an oversized number of features. This method can be also used as a pre-screening of the features. To restrict the selected set of features F, it is possible to compute for each $j \in F$ the RE overall mean difference in the two groups by taking the mean over all the observations in $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$ as in (2.5); rank them based on (2.5) in decreasing order and given a threshold $\delta \in (0, 1)$ and the Δ_δ, δ-th quantile evaluated on the distribution of $\{\Delta_j\}_{j \in F}$, select:

$$F_{combined} = \{j | \Delta_j > \Delta_\delta, j \in F\}$$

2.3. Generalization of anomaly detection autoencoders for feature selection to multivariate response

This section is devoted to the illustration of a multi-outcome feature selection. Multivariate FS is achieved by the proper definition of DSAEE groups, accounting for correlation in the endpoints. In particular, feature selection is performed starting from a unique train sample, extracted from the intersection of the majority classes of each endpoint, and an endpoint-depending test sample. This means that each learner is trained to represent a population that does not present any toxicity. Clinicians believe that interindividual differences in toxicity outcomes are only partly endpoint-specific [30], consequently starting from a control population extracted specifically from each endpoint majority class could lead to concealing of genetic risk pattern linked to general toxicity, due to their presence in both the majority and minority class. The selection is able to produce a set of SNPs describing each specific toxicity accounting for the dependency structure in the multivariate outcome and consequently increasing the statistical power of the model. The selected SNPs can then be combined to form an informative set of features correlated with general toxicity and able to distinguish generally radiosensitive patients.

A multi-outcome binary supervised learning setup is considered with an available set of N (input, targets) pairs

$$\mathbf{D} = \{(\underline{x}_1, \underline{y}_1), ..., (\underline{x}_N, \underline{y}_N)\}$$

where $\underline{y}_i = \{Y_{i1}, ..., Y_{iT}\}$ is the multi-endpoint target, each endpoint that takes values in $\{0, 1\}$ and $x_i \in \{1, ..., M\}^J$ with $i = \{1, ..., N\}$ is the input feature vector of true categorical data. Suppose that a fixed number of M categories is available for each feature. Finally, suppose that imbalance in the classes is present for each of the endpoints, with a minority class $Y_k = 1$ and a majority class $Y_k = 0$ with $k \in \{1, ..., T\}$. The rationale of the DSAEE follows in general the one presented in [38] with the intention of including correlation between the endpoints in the analysis to exploit information from other endpoints in the FS of each endpoint. A set of features is selected for each endpoint and a comprehensive outcome as detailed in the following and schematized in Algorithm 2.3.

In particular, $\mathbf{X}$ is defined as the dataset of the J categorical features. $\mathbf{Y}$ is defined as the target dataset, containing the T binary endpoint.

Algorithm 2.3 DSAEE for multi-outcome Feature Selection

INPUTS

- **D**: Dataset with J covariates and M outcomes;
- B: Enasamble iterations;
- $\underline{\alpha}$: Vector of thresholds (quantiles).

OUTPUT

- Feature set F_k, $k \in \{1, ...T\}$ one for each endpoint;
- Feature set F_{tot} linked to a comprehensive outcome.

```
$\mathbf{X} \leftarrow$ from $\mathbf{D}$ extract the covariate dataset with J features;
$\mathbf{Y} \leftarrow$ from $\mathbf{D}$ the targets dataset with T outcomes;
for $k = 1, ...., T$ do
    $\mathbf{X^k_{maj}} \leftarrow \{x_i \in \mathbf{X} | Y_{ik} = 0\}$
end for
$\mathbf{X_{maj}} \leftarrow \bigcap_{k=1}^{T} \mathbf{X^k_{maj}}$
for $k = 1, ...., M$ do
    $\mathbf{X'} \leftarrow \{x_i \in \mathbf{X} | Y_{ik} = 1\}$
    O $\leftarrow$ minority sample size or number of $\mathbf{X'}$ rows
    Set $\mathbf{Q}$ // empty array
    for $i = 1, ...., B$ do
        $\mathbf{X_{test}} \leftarrow$ Sample O observations from $\mathbf{X_{maj}}$
        $\mathbf{X_{train}} \leftarrow \{\underline{x}_i \in \mathbf{X_{maj}} | \underline{x}_i \notin \mathbf{X_{test}}\}$
        $\mathbf{X_{test}} \leftarrow concatenate(\mathbf{X_{test}}, \mathbf{X}')$
        Define $Y_{test}$ containing class belongings of $\mathbf{X_{test}}$ in endpoint $k$
        $\phi$, $\mathbf{W}$, $\mathbf{W'}$ $\leftarrow$ train a deep sparse autoencoder using the normal dataset $\mathbf{X_{train}}$
        considering (2.9)
        $\mathbf{R} \leftarrow$ evaluate reconstruction error on $\mathbf{X_{test}}$ as in (2.8)
        Label $\mathbf{R}$ with $Y_{test}$
        $\mathbf{Q} \leftarrow concatenate(\mathbf{Q}, \mathbf{R})$
    end for
    $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}} \leftarrow \{re_j \in \mathbf{Q} | Y_{test,j} = 0\}$, $\{re_j \in \mathbf{Q} | Y_{test,j} = 1\}$
    $mean\_re_{maj}$ , $mean\_re_{min} \leftarrow$ Column mean of $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$
    $\Delta \leftarrow mean\_re_{min}$ - $mean\_re_{maj}$
    $F_k \leftarrow \{j | \Delta_j > \Delta_{\alpha_k}, j \in \{j, ..., J\}\}$
end for
$F_{tot} \leftarrow \bigcup_{k=1}^{T} F_k$
```

First, the set of common controls needs to be defined. These are those patients that do not present any of the endpoints considered and it is computed as the intersection of the control samples of all the endpoints, in the following it is referred to as the majority class. For each endpoint, the specific set of selected features is computed as follows. First, the case sample (minority class) is defined, including all the patients that present the specific endpoint. Then, in each ensemble iteration b $\in \{1, ..., B\}$ a test set containing 2 * O data points, where O is the minority class numerosity, is constructed by concatenating all the minority class patients with a random sample of the same size from the common control sample. The remaining observations of the common majority class are included in the training set. This train-set structure allows us to train each DSAE learner in an unsupervised fashion only on the population not presenting any toxicity, and to test its performance when facing both groups' examples so that comparison of the RE in two populations is possible. The network weights and reconstruction map are optimized to have the best possible representation and reconstruction of the J features in the common majority class exploiting the loss in (2.9).

$$\overline{\phi(\underline{x})} = argmin_{\phi} L(\underline{x}, \phi(\underline{x})) \tag{2.7}$$

$$L(x_j, \phi(x_j)) = -\sum_{k=0}^{M} (x_{jk} * log(\phi(x_{jk}))) \quad for \;\; j \in \{1, ..., J\} \tag{2.8}$$

$$L^S(\underline{x}, \phi(\underline{x})) = \sum_{j=1}^{J} L(x_j, \phi(x_j)) + \lambda |h(l)| \tag{2.9}$$

where $\underline{x} \in \mathbf{X_{train}}$. The loss is composed of two different parts, the first one being a cross-entropy that evaluates the difference in the probability distribution between the autoencoder outcome and the one-hot encoding of the corresponding categorical covariate. The second part instead is the L1 loss to induce sparsity in the latent representation of the covariates, improving the generalization ability of the model.

Once the network has been trained, it is applied to the test set. For each observation of the test set, the outcome is composed of J probability distributions over the M categories describing the likelihood of each feature belonging to each class.

$$\phi(\mathbf{X_{test}}) \in R^{(J,M)}$$

The reconstruction error is evaluated on each sample of the test set as in (2.8).

The test set REs from each ensemble repetition are concatenated in **Q** and at the end of the ensemble procedure $\mathbf{Q} \in R^{(B*2*O,J)}$. $\mathbf{Q}$ is then split in $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}}$ based on the belonging of each observation in the test set. For each feature j, the RE difference in the two groups is evaluated by taking the mean over all the observations in $\mathbf{Q_{maj}}$ and $\mathbf{Q_{min}}$.

$$l_j^b = \frac{\sum_{i=1}^{O} L(x_j^i|\underline{x^i} \in \mathbf{Q_{min}}, \phi(x_j^i)) - \sum_{i=1}^{O}(L(x_j^i|\underline{x^i} \in \mathbf{Q_{maj}}, \phi(x_j^i))}{O}, \quad j \in J, b \in B \quad (2.10)$$

$$\Delta_j = \frac{\sum_{b=1}^{B} l_j^b}{B}, \quad j \in J \quad (2.11)$$

Averaging is performed also over ensemble repetition to achieve higher robustness in the selection method. Finally, the features are ranked based on (2.11) in decreasing order. The highest-ranked features are those accurately reconstructed by the DSAE on the majority class (lower RE), and poorly reconstructed on the minority class (higher RE). To identify an exact feature set for the endpoint considered, a threshold $\alpha_k \in (0,1)$ is defined. Δ_{α_k} is the α_k-th quantile evaluated on the distribution of $\{\Delta_j\}_{j=1}^J$. Then, all those features j whose average RE difference is above the user-defined quantile are selected:

$$F_k = \{j|\Delta_j > \Delta_{\alpha_k}, j \in \{1, .., J\}\}$$

Once the set of features is selected for each endpoint, it is possible to define a set F_{tot} as the union of all the features selected, that incorporates significant features for a comprehensive endpoint.

2.4. Multi-outcome feature selection via anomaly detection autoencoder in genomic applications

The last section presents a methodology that encapsulates the key components described in previous sections and implements a multi-outcome feature selection specifically developed in the context of clinical risk models and genomic data.

A multi-outcome binary supervised learning setup is considered with an available set of N (input, targets) pairs

$$\tilde{\mathbf{D}} = \{(\underline{\tilde{x}}_1, \underline{y}_1), ..., (\underline{\tilde{x}}_N, \underline{y}_N)\}$$

where $\underline{y}_i = \{Y_{i1}, ..., Y_{iT}\}$ is the multi-endpoint target, each endpoint takes values in $\{0,1\}$ and $\underline{\tilde{x}}_i \in R^J$ with $i = 1, ..., N$ is the input feature vector of imputed data or, in general, noisy data. Suppose that $\underline{x}_i$, true categorical feature vector, is known for each patient present in the training set and that a fixed number of M categories is available for each feature. Therefore a second dataset is available with N (input, target) pairs

$$\mathbf{D} = \{(\underline{x_1}, \underline{y}_1), ..., (\underline{x_N}, \underline{y}_N)\}$$

where $\underline{y}_i$ is the same multi-endpoint target and $\underline{x}_i \in \{1, ..., M\}^J$ with $i = 1, ..., N$ is the input feature vector of categorical data. If the true categorical feature vector is unknown, as in the denoising case, it is possible to simply round each imputation to the closest integer and consider it the categorical representation of the data. Finally, suppose that imbalance in the classes is present for each of the endpoints, with a minority class $Y_k = 1$ and a majority class $Y_k = 0$ with $k \in \{1, ..., T\}$. The method employs an ensemble of deep sparse autoencoders to perform multi-output feature selection. Each autoencoder is trained, as an anomaly detection autoencoder, to optimally represent a class of controls, and aim to distinguish them from the anomalies (cases). Anomaly detection autoencoders perform well, especially in unbalanced settings, where the normal sample presents several units (majority class) and the anomaly sample only a few (minority class). In a supervised learning setting, it is possible to observe how each feature is reconstructed in the majority and minority class samples. The difference in the distribution of the reconstruction error in the two groups is assessed by exploiting a non-parametric test, and discriminant features are selected as those presenting statistically different distributions between them. Multivariate FS is achieved by the proper definition of DSAE control and case groups, accounting for correlation in the endpoints. In particular, feature selection is performed starting from a unique control sample, extracted from the intersection of the majority classes of each target, and a target-dependent case sample. Each learner is trained to represent a population that does not present any of the targets considered. A set of features is selected for each endpoint and the union of all of them is considered significant in explaining a comprehensive outcome. The method is robust to error in the imputation of genomic data thanks to a procedure inspired by denoising autoencoders. Each learner is trained to reconstruct from the noisy input the categorical output, including imputation noise in the reconstruction process and imposing an unbiased analysis. The method is detailed in the following and schematized in Algorithm 2.4.

In particular, $\mathbf{X}$ is defined as the dataset of the J categorical features, and $\tilde{\mathbf{X}}$ as the dataset of the J continuous features. $\mathbf{Y}$ is defined as the target dataset, containing the T binary outcomes.

Algorithm 2.4 DSAEE for multi-outcome Feature Selection in the context of genomic data

```
INPUTS
  • $\mathbf{D}$ : categorical covariate dataset;
  • $\tilde{\mathbf{D}}$: continuous covariate dataset;
  • B: Enasamble iterations;
  • $\underline{\alpha}$: Vector of thresholds (quantiles).
OUTPUT
  • Feature set $F_k$, $k \in \{1, ...T\}$ one for each endpoint;
  • Feature set $F_{tot}$ linked to a comprehensive outcome.
$\mathbf{X} \leftarrow$ from $\mathbf{D}$ extract the categorical covariatre dataset with J features;
$\tilde{\mathbf{X}} \leftarrow$ from $\tilde{\mathbf{D}}$ extract the continuous covariatre dataset with J features;
$\mathbf{Y} \leftarrow$ from $\mathbf{D}$ the targets dataset with T outcomes;
for $k = 1, ...., T$ do
    $\mathbf{X^k_{maj}} \leftarrow \{\underline{x}_i \in \mathbf{X} | Y_{ik} = 0\}$;
    $\tilde{\mathbf{X}}^\mathbf{k}_\mathbf{maj} \leftarrow \{\underline{\tilde{x}_i} \in \tilde{\mathbf{X}} | Y_{ik} = 0\}$;
end
$\mathbf{X_{maj}} \leftarrow \bigcap_{k=1}^{T} \mathbf{X^k_{maj}}$
$\tilde{\mathbf{X}}_\mathbf{maj} \leftarrow \bigcap_{k=1}^{T} \tilde{\mathbf{X}}^\mathbf{k}_\mathbf{maj}$
```

First, the set of common controls needs to be defined. These are those samples that do not present any of the outcomes considered and it is computed as the intersection of the majority class samples from each endpoint. Two different datasets containing common controls are available $\tilde{\mathbf{X}}_\mathbf{maj}$ and $\mathbf{X_{maj}}$, respectively with continuous and categorical features and will be referred to as the majority class dataset in the following.

DSAEE for multi-outcome Feature Selection in the context of genomic data (Part II)

```
for $k = 1, ...., T$ do
    $\mathbf{X'} \leftarrow \{\underline{x}_i \in \mathbf{X} | Y_{ik} = 1\}$
    $\tilde{\mathbf{X}}' \leftarrow \{\underline{\tilde{x}_i} \in \tilde{\mathbf{X}} | Y_{ik} = 1\}$
    O ← minority sample size or number of $\mathbf{X'}$ rows
    Set $\mathbf{Q}$ //empty array
    for $i = 1, ...., B$ do
        $\mathbf{X_{test}} \leftarrow$ Sample O observations from $\mathbf{X_{maj}}$
        $\tilde{\mathbf{X}}_{\mathbf{test}} \leftarrow$ Sample the same observations in $\mathbf{X_{test}}$ from $\tilde{\mathbf{X}}_{\mathbf{maj}}$
        $\mathbf{X_{train}} \leftarrow \{\underline{x}_i \in \mathbf{X_{maj}} | \underline{x}_i \notin \mathbf{X_{test}}\}$
        $\tilde{\mathbf{X}}_{\mathbf{train}} \leftarrow \{\underline{\tilde{x}_i} \in \tilde{\mathbf{X}}_{\mathbf{maj}} | \underline{\tilde{x}_i} \notin \tilde{\mathbf{X}}_{\mathbf{test}}\}$
        $\mathbf{X_{test}} \leftarrow concatenate(\mathbf{X_{test}}, \mathbf{X}')$
        $\tilde{\mathbf{X}}_{\mathbf{test}} \leftarrow concatenate(\tilde{\mathbf{X}}_{\mathbf{test}}, \tilde{\mathbf{X}}')$
        Define $Y_{test}$ containing class belongings of $\mathbf{X_{test}}$ in endpoint $k$
        $\phi$, $\mathbf{W}$, $\mathbf{W'}$ ← train a deep sparse autoencoder using the continuous dataset $\tilde{\mathbf{X}}_{\mathbf{train}}$
        as input and the categorical one $\mathbf{X_{train}}$ as output considering (2.3)
        $\mathbf{R}$ ← evaluate reconstruction error on $\tilde{\mathbf{X}}_{\mathbf{test}}$ and $\mathbf{X_{test}}$ as in (2.2)
        Label $\mathbf{R}$ with $Y_{test}$
        $\mathbf{Q} \leftarrow concatenate(\mathbf{Q}, \mathbf{R})$

    end
    $\mathbf{Q_{maj}}$, $\mathbf{Q_{min}} \leftarrow \{re_j \in \mathbf{Q} | Y_{test,j} = 0\}$, $\{re_j \in \mathbf{Q} | Y_{test,j} = 1\}$
    set $F_k$ //empty set
    for $j$ $in$ $J$ do
        p$_j$← p-value of the Smirnov test,
        testing the difference in $\mathbf{Q_{maj}}$ and $\mathbf{Q_{min}}$ jth column
        if $p_j < \frac{0{,}05}{J}$ then
            Include j in $F_k$
        end
    end
end
$F_{tot} \leftarrow \bigcup_{k=1}^{T} F_k$
```

For each outcome, the specific set of selected features is computed as follows. The case sample (minority class) is defined, including all the patients that present the specific endpoint. Two different minority datasets are computed $\tilde{\mathbf{X}}'$ and $\mathbf{X'}$, respectively with continuous and categorical covariates. Then, in each ensemble iteration b ∈ $\{1, ..., B\}$

a test set containing 2 * O data points, where O is the minority class numerosity, is constructed by concatenating all the minority class patients with a random sample of the same size from the common control class patients. The remaining observations of the latest are included in the training set. Each DSAE learner is trained in an unsupervised fashion only on the population not presenting any outcome. The training includes a denoising procedure forcing the DSAE to reconstruct from the continuous input ($\tilde{\mathbf{X}}_{\mathbf{train}}$) its categorical representation ($\mathbf{X}_{\mathbf{train}}$) so that the possible error due to imputation in the data is accounted for in the comparison. The network weights and reconstruction map are optimized to have the best possible representation and reconstruction of the J features in the training set exploiting the loss in (2.3):

$$Loss(\underline{x}, \phi(\underline{\tilde{x}})) = \sum_{j=1}^{J} - \sum_{k=0}^{M} (x_{jk} * log(\phi(\tilde{x}_{jk}))) + \lambda |h(l)|$$

where $\underline{\tilde{x}} \in \tilde{\mathbf{X}}_{\mathbf{train}}$ and $\underline{x} \in \mathbf{X}_{\mathbf{train}}$. The loss function heeds the cross-entropy between the autoencoder outcome and the one-hot encoding of the corresponding categorical covariate and the L1 loss to improve the generalization ability of the model.

Once the network has been trained, it is applied to the test set. The evaluation of its performance when facing both majority and minority class examples enables the comparison of the RE in two populations. For each observation of the test set, the outcome is composed of J probability distributions over the M categories describing the likelihood of each feature belonging to each class.

$$\phi(\tilde{\mathbf{X}}_{\mathbf{test}}) \in R^{(J,M)}$$

The reconstruction error is evaluated on each sample of the test set as in (2.2):

$$RE(x_j, \phi(\tilde{x}_j)) = - \sum_{k=0}^{M} (x_{jk} * log(\phi(\tilde{x}_{jk}))) \quad for \;\; j \in \{1, ..., J\}$$

The test set REs from each ensemble repetition are concatenated in Q and labeled according to the class belongings of $\mathbf{X}_{\mathbf{test}}$ in the specific endpoint. At the end of the ensemble procedure $\mathbf{Q} \in R^{(B*2*O,J)}$. $\mathbf{Q}$ is then split in $\mathbf{Q}_{\mathbf{maj}}$, $\mathbf{Q}_{\mathbf{min}}$, dividing majority and minority class observations in the test set. Each of the $B * O$ observations in the two is considered as a sample extracted from each distribution $x_j | x_j \in \mathbf{Q}_{\mathbf{min}} \sim f_j^{min}(x)$ and $x_j | x_j \in \mathbf{Q}_{\mathbf{maj}} \sim f_j^{maj}(x)$. It is possible then to compare the samples and test if the feature distributions in different groups are statistically different. The analysis is performed via the Smirnov test, a non-parametric two-sample test, used to determine if two independent random samples appear to follow the same distribution. Once the test is performed, the set F_k includes all the features whose test p-value is lower than the Bonferroni corrected threshold of 0.05.

If the F_k includes an oversized number of features. A second selection method can be applied to F_k to restrict it. It is possible to compute for each $j \in F_k$ the RE difference in the two groups by taking the mean over all the observations in $\mathbf{Q_{maj}}$ and $\mathbf{Q_{min}}$ as in (2.5):

$$\Delta_j = \frac{\sum_{b=1}^{B} \left(\sum_{i=1}^{O} L(x_j^i | \underline{x^i} \in \mathbf{Q_{min}}, \phi(x_j^i)) - \sum_{i=1}^{O} (L(x_j^i | \underline{x^i} \in \mathbf{Q_{maj}}, \phi(x_j^i)) \right) /O,}{B}, \quad j \in J$$

rank them based on Δ_j in decreasing order and given a threshold $\alpha_k \in (0,1)$ and the Δ_{α_k}, δ-th quantile evaluated on the distribution of $\{\Delta_j\}_{j \in F_k}$, select:

$$F = \{j | \Delta_j > \Delta_{\alpha_k}, j \in F_k\}$$

Once the set of features is selected for each endpoint, it is possible to define a set F_{tot} as the union of all the features selected, that incorporates significant features for a comprehensive endpoint.

Evaluation metrics employed in the analysis

The methodologies presented in the previous section are separately tested on simulated datasets and the comprehensive algorithm is applied to REQUITE case study. To highlight their proprieties, performance metrics are employed.

The methods' performance in classification tasks is evaluated based on the classical metrics indexes of binary classification. In this work, we evaluated those in Table 2.1. Note that since data present unbalanced classes few performance indexes evaluating the prediction of the minority class are introduced. The metrics' definitions are stated hypothesizing a binary classification with subjects positive and negative to the outcome. The method's performance in FS tasks is evaluated by metrics indices described in Table 2.2 considering the metric introduced in [39] and an additional one evaluating the ability of the method in the selecting features specific to the single outcome considered in presence of highly correlated outputs. Finally, to evaluate autoencoders' performance metrics presented in Table 2.3 are exploited. In the definitions of Table 2.3 $\underline{x}$ represent the true value and $\overline{\underline{x}}$ the predicted value.

Metrics used to evaluate binary classification

Metrics	Description	Formula
AUC	Area under the curve of the receiver operating characteristic curve (ROC). A ROC curve is a graph showing the performance of a classification model at all classification thresholds. A ROC curve plots True Positive Rate(TPR) (y) vs. False Positive Rate (FPR)(x) at different classification thresholds. AUC is usually approximated using a Riemann sum.	True Positive Rate(TPR): $TRP = \frac{TP}{TP+FN}$ False Positive Rate (FPR): $FPR = \frac{FP}{FP+TN}$
Precision	Precision describes the ratio between true positive and total number of sample classified as positive. The precision describes the ability of the classifier not to label as positive a sample that is negative.	$P = \frac{TP}{TP+FP}$
AP	Avarage precision: AP summarizes a precision(P)-recall(R) curve as the weighted mean of precisions achieved at each threshold, with the increase in recall from the previous threshold used as the weight. It indicates whether the classifier is able to correctly identify all the positive samples without accidentally marking too many negative samples as positive	$AP = \sum_n (R_n - R_n - 1) * P$ where $R = \frac{TP}{TP+FN}$ and $P = \frac{TP}{TP+FP}$
F1	Measure of the test accuracy evaluated from the precision and recall of the classification	$F1 = \frac{2*P*R}{P+R}$
Sensitivity	Percentage of true positive among those that are positive (True Positive Rate)	$TRP = \frac{TP}{TP+FN}$
Specificity	Percenatage of true negative among those that are negative (True Negative Rate)	$TNR = \frac{TN}{TN+FP}$
NPV	Negative Predictive Values: percentage of true negative among those who are classified as negative (True Negative Rate)	$NPV = \frac{TN}{TN+FN}$

Table 2.1: **Definition and description of the metrics used to evaluate binary classification method**. The table contains in the columns the metrics name used in the thesis text, their descriptions, and the formulas used to compute them given the confusion matrix of the classification model.

Metrics used to evaluate FS

Metrics	Description	Formula
Tests: Wasserstein distance Wilcoxon paired signed-rank test t-test	To quantify and evaluate the separation of the class-specific RE distributions parametric and non-parametric methods are adopted to avoid strict assumptions on the distributions of RE. The Wasserstein distance quantifies the difference in the shape of the two empirical distributions, the Wilcoxon paired signed-rank test evaluates whether the two related samples come from the same distribution (two-sided test), or there is a stochastic order between the two distributions (one-sided test) and the t-test evaluate the mean difference in the groups under the normality assumption.	
FSP	FS Performance evaluates the capability of the algorithm to produce a meaningful ranking of features and selecting the most informative features to discriminate the two classes. It is a useful-to-selected ratio where both informative and redundant features are considered useful.	$FSP = \|IRF\|/\|F\|$ where $\|IRF\|$: total number of informative and redundant features selected and $\|F\|$: total number of features selected
FSC	FS Correspondence quantifies the ability of the methodology in selecting every and only informative feature avoiding redundant ones.	$FSC = \|IF\|/\|TIF\|$ where $\|IF\|$: total number of informative features selected and $\|TIF\|$: total number of true informative features

Table 2.2: **Definition and description of the metrics used to evaluate feature selection**. The table contains in the columns the metric names used in the thesis text, their description, and the formulas used to compute them.

Metrics used to evaluate Autoencoder perfromances

Metrics	Desciption	Formula
MSE	Mean Square Error (MSE) measures the square of Euclidean distance between the estimated values and the actual value, i.e. the average squared difference between them.	$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - \overline{x}_i)^2$
Accuracy / Binary accuracy	The frequency with which the estimated value matches the true value	$acc = \frac{\sum_i (x_i == \overline{x}_i)}{N}$
Binary crossentropy	The cross-entropy calculates the difference between two probability distributions p and q. In binary classification, the true probability p_i is the true label, and the given distribution q_i is the predicted value of the current model	cross $\leftarrow -E_p(logq)$ where E is the expected value w.r.t. p bin_cross $\leftarrow -\sum_i \sum_j p_j log(q_j)$= =$-\sum_i x_i log(\overline{x}_i)$+ (1- x$_i$)$log(1-\overline{x}_i)$
AUC	See: Table 2.4	See: Table 2.4

Table 2.3: **Definition and description of the metrics used to evaluate autoencoder performance.** The table contains in the columns the metric names used in the thesis text, their descriptions, and the formulas used to compute them.

3 | Applications

3.1. Simulation setting

In this section, the distinctive aspects introduced in the proposed method are validated through a simulation study. To do that, simulated data needs to reproduce peculiar characteristics of genomic data, namely categorical features (i.e. the variants) and the presence of complex interactions determining the endpoints' onset. Moreover, to test the improved power of the proposed FS algorithm for multivariate targets, we simulated a generative model determining correlated endpoints. The algorithm exploited to generate the simulated data begins with the construction of the multivariate target endpoint. Specifically, the multivariate output is constructed as a matrix of binary features with a user-defined intra-features correlation structure. Then, the genotype data is generated as a set of binary covariates representing the variants (with a defined variant's frequency). At this point, to simulate the complex genotype-multivariate phenotype relationship, the algorithm defines, for each dimension of the multivariate endpoint a set of interacting features (hereby called "pattern") with a specified co-occurrence frequency with its corresponding dimension. More details about the generative process can be found in Appendix B and in [36].

3.1.1. Denoising proprieties of the developed method

This section aims at validating the utility of the denoising aspect introduced in the novel DSAEE algorithm. That is, to verify the capability of the denoising DSAEE to improve imputation error handling.

This simulated experiment includes a dataset with 1000 observations and 100 variants with a 10% relative frequency. The output is univariate with a minority class composed of 100 samples. The length of the associated pattern contains 20 variants and the co-occurrence frequency of the pattern and endpoint is 70%. The noisy dataset is generated starting from the original categorical dataset adding random noise as in Algorithm 3.1.

Algorithm 3.1 Noise addition procedure

INPUTS

- **D** : categorical covariate dataset.

OUTPUT : $\tilde{\mathbf{D}}$: continuous covariate dataset.

```
def AddNoise(x) {
ok ← sample from ∼ Bi(0.7)
if ok then
    $\epsilon$ ← sample from ∼ Exp(5)
    if x is 0 then
        $x \leftarrow min(1, x + \epsilon)$
    else
        x ← $max(0, x - \epsilon)$
    end if
end if
return x
}
$\mathbf{X}$ ← from $\mathbf{D}$ extract the categorical covariate dataset with J features
$\mathbf{Y}$ ← from $\mathbf{D}$ the targets dataset with T outcomes;
$\tilde{\mathbf{X}}$ ← apply elementwise AddNoise to $\mathbf{X}$
$\tilde{\mathbf{D}} \leftarrow merge(\tilde{\mathbf{X}}, \mathbf{Y})$
```

To test the denoising capability of the Algorithm 2.1, referred to in the following as denoising DSAEE, is compared to the DSAEE Algorithm 1.2, referred to in the following as DSAEE. The denoising DSAEE algorithm introduces exclusively the denoising procedure with respect to the DSAEE algorithm. Changes and improvements from one algorithm representation to the other are then the consequence of better imputation error handling. In both cases, the same simple autoencoder architecture is implemented and its in-training convergence is analyzed. Each AE is composed by an encoder with one 90-nodes hidden layer, followed by a bottleneck layer of 50 nodes and a symmetrical decoder and the ensemble is composed of 10 learners. More details about the architecture are present in Table B.1 in Appendix B.

To quantify and evaluate the reconstruction ability of the autoencoder, metrics presented in Table 2.3 are adopted and their evolution is studied during the training process. In Figure 3.1 in-train loss of both methods is plotted to be compared.

In the DSAEE methodology, features are considered to be continuous variables. To evaluate the loss the mean squared error is adopted. In the denoising DSAEE features are, instead, reconstructed as categorical variables, and the loss is evaluated via binary-cross entropy. During both trainings, a validation split is performed to describe how the autoencoder behaves when tested on unseen data.

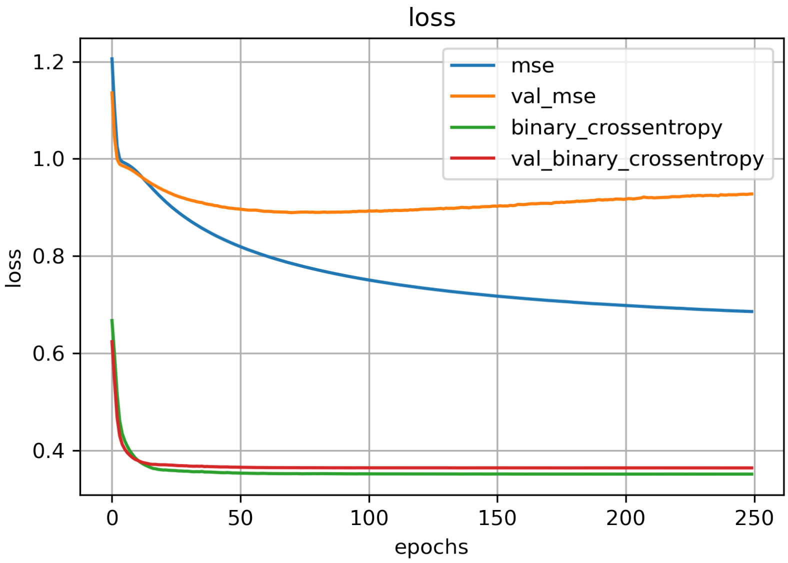

Figure 3.1: **In-train loss of the compared methodologies.** In the DSAEE the loss is evaluated via MSE since noisy data are treated like continuous variables, while in denoising DSAEE the loss is evaluated via cross-entropy since the encoder output distribution is compared to the real categorical value. The training was performed excluding a validation set to mimic the performance of both algorithms on unseen data. The loss on the validation sets is also presented

The accuracy is reported in Figure 3.2.

Further insight into the representation capability of the DSAEE denoising can be extracted from the mean AUC. The denoising DSAEE reconstructs binary data evaluating the probability of the variation presence. The AUC can be computed for each feature reconstruction and averaged over all. The in-train AUC is shown in Figure 3.3.

Several observations emerge from these plots. First, the denoising DSAEE keeps both the validation and training set loss very low and very close with respect to DSAEE, showing a better reconstruction performance on seen and unseen data and better generalization ability. Moreover, both loss and accuracy plots reveal a faster and smoother convergence in the denoising DSAEE. Reducing the computational effort in training has great advantages on the total computational time enabling, in ensemble learning algorithms, a higher number of repetitions to be performed and higher performance.

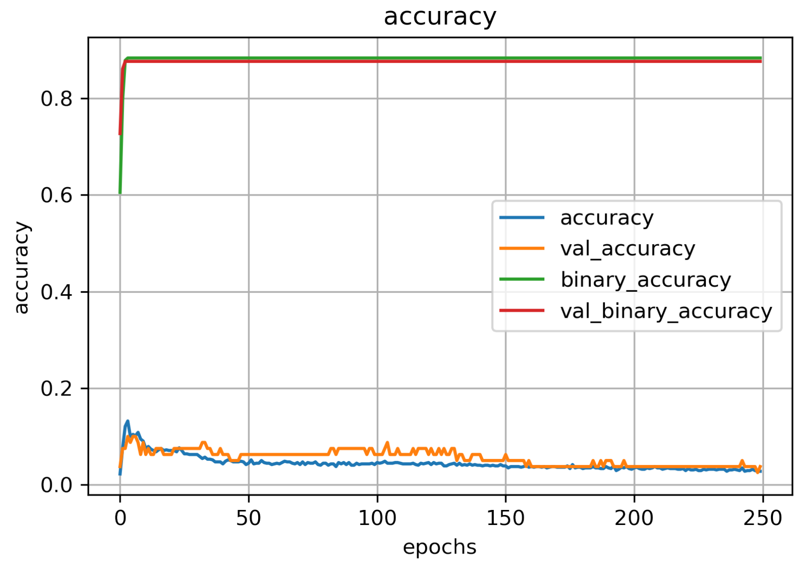

Figure 3.2: **In-train accuracy of the compared methodologies.** The training was performed excluding a validation set to mimic the performance of the autoencoder on unseen data

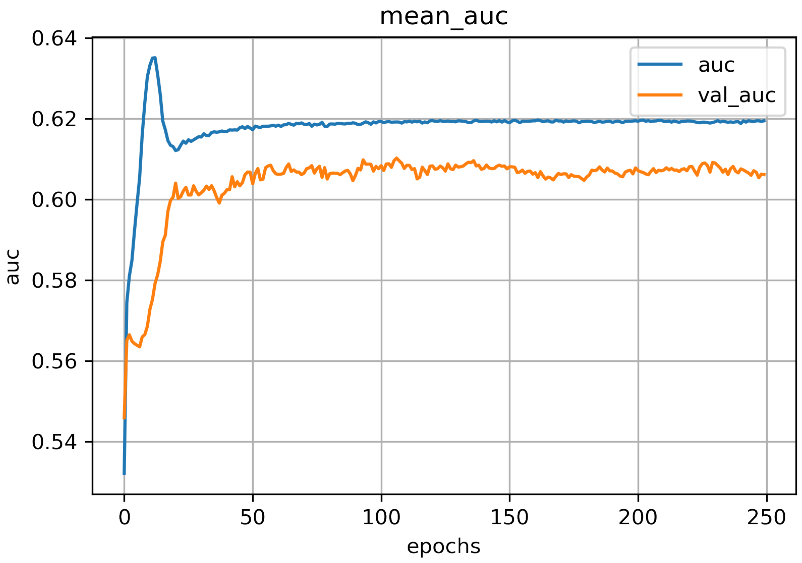

Figure 3.3: **In-train AUC** of DSAEE denoising Algorithm. The training was performed excluding a validation set to mimic the performance of the autoencoder on unseen data

Finally, the AUC performance of the denoising algorithm stabilized at approximately 62% in both the training and the validation set. The performance in AUC, although, probably weakened by the high noise-to-signal ratio present in the data, shows the ability of the autoencoder to isolate the signal, and compute a latent representation that enables a good reconstruction of noisy data.

In conclusion, the simulated analysis reveals the denoising DSAEE ability to reconstruct from noisy data their true categorical values. In this stage, accuracy in data representation is fundamental to better distinguish the classes of the binary outcome, and consequently to perform a better feature selection when applied within the multivariate methodology.

3.1.2. Distribution-based methodology for feature selection to improve selection precision and correspondence

The aim of this section is to verify the improvement in selection performance including the distributional approximation of the reconstruction error in the discovery of significant covariates.

This simulated experiment includes a dataset with 1000 observations and 100 variants with a 10% relative frequency. The output is univariate with a minority class composed of 100 samples. The length of the associated pattern contains 20 variants and the co-occurrence frequency of the pattern and endpoint is 70%. To test the improvement in performance, the selection method presented in Section 2.2, referred to in the following as distributional FS, is compared to the one presented in Algorithm 1.2, referred to in the following as ranking FS. In addition, the combined selection where a pre-screening of the feature is performed via distributional FS and the final selection is computed on the pre-screening exploiting the ranking FS is analyzed. A simple architecture is implemented and unique training is performed before selecting independently the relevant features with the methods. Each AE is composed of an encoder with one 90-nodes hidden layer, followed by a bottleneck layer of 50 nodes and a symmetrical decoder, and the ensemble is composed of 10 learners. More details about the architecture are present in Table B.2 in Appendix B. A single DSAEE is trained according to Algorithm 1.2. Once matrix $\mathbf{Q}$ is computed, each selection methodology is applied to the same $\mathbf{Q}$. Selection performances are evaluated via metrics present in Table 2.2, namely FSP, which evaluates the percentage of useful features selected over the total number of selected features, and FSC, which evaluates the ability of the method in selecting every and only informative feature avoiding redundant ones. The δ parameter of the ranking and combined selection is optimized in each method first on FSP, as the fundamental goal is to avoid introducing non-informative features in the selection, and secondly on FSC to get the best possible set of features. The results are presented in Table 3.1.

Results show how the distributional FS method selects a large number of features that

Result from the simulation study of FS

Distributional FS		Combined FS		Ranking FS	
FSP	FSC	FSP	FSC	FSP	FSC
0,171	1	1	**0,584**	1	**0,417**

Table 3.1: **Result metrics from the different FS methods**. In the table metrics evaluation from Table 2.2 are reported for each of the methods compared. In the first columns are reported the result using Algorithm 2.2, in the second results from the combined methodology and in the third the results using Algorithm 1.2

cover all the covariates important for the endpoint, introducing, on the other hand, a large number of irrelevant features in the selection. A possible reason for this behavior is the clean separation of the class in the simulation setting. In applications, usually, groups are overlapping and the set of features statistically able to distinguish between them is restricted. The combined method resolves this issue by selecting the most relevant feature between those that present a different distribution in the two groups. All the selected features are significant, as in the ranking FS, but the FSC is 17% higher than the ranking selection method.
Through these observations, the improvement in the novel selection method can be validated. Other results are presented in Section 3.1.3. In that context, simulation analyses are performed on a multivariate outcome, FS is performed with each methodology and can be compared to robustify the findings of this section. Moreover, FSC in multivariate analysis is also a measure of how the methods can select specific features of one endpoint avoiding those related to correlated outcomes.

In conclusion, the simulated analysis reveals the distributional method improves the FSC of feature selection and, more importantly, that the proposed combined methodology can reach ranking method performance in FSP and improve performance in FSC.

3.1.3. Inclusion of intra-endpoints correlation structure in the FS improve the method performance

This section aims to verify the capability of the multi-outcome methodology described in Section 2.3 to improve variable selection in terms of the metrics defined in Table 2.2. In the simulation, more performance evaluations of the distribution FS methodology are included.

This simulated experiment includes a dataset with 1000 observations and 100 variants

Simulation dataset multivariate case

	Numerosity	Cases	Correlation	Pattern-endpoint co-occurance
Total	1000	50		
y_1	1000	27	a (0.8)	0.7
y_2	1000	12	a (0.8)	0.8
y_3	1000	7	a (0.8)	0.9
y_4	1000	25	b (1)	0.9
y_5	1000	20	b (1)	0.6

Table 3.2: **Multi-outcome simulation dataset summary.** 5 outcomes are considered in the simulation. Their numerosities in the minority classes with their correlation structure and theit co-occurrance with the associated patterns are reported in the table

with a 10% relative frequency. The output is multivariate with five outcomes and a comprehensive minority class composed of 50 samples. The correlation structure within the multivariate output is composed of two sets of correlated dimensions: the first 3 with a correlation of 0.8 and the last two with a correlation of 0.7. Each outcome is associated with a pattern of variants. The length of the associated patterns ranges from 10 to 15 variants. The co-occurrence frequency of pattern and endpoint ranges from 60% to 90%. Details about the dataset are presented in Table 3.2 To test the improvement in the multi-outcome selection the multivariate algorithm, described in Algorithm 2.3, is tested against an algorithm performing univariate selection for each outcome as described in Algorithm 1.2. The multi-outcome algorithm introduces the definition of a unique control set, i.e. the null class across all target dimensions. Conversely, the univariate algorithm is trained on the control group of each endpoint separately. Changes and improvements from one algorithm selection to the other are then to be attributed only to the introduction of outcomes correlation in the FS. The final selection is performed exploiting the three different selection methods studied in Section 3.1.2.

To compare the two algorithms in terms of feature selection only, the same AE architecture was exploited for both. Specifically, each AE is composed of an encoder with one 90-nodes hidden layer, followed by a bottleneck layer of 50 nodes and a symmetrical decoder and the ensemble is composed of 10 learners. The parameter regulating the sparsity term of the loss (i.e. *lambda*) is instead optimized by grid search during the training of each DSAEE. More details about the architecture are present in Table B.3 in Appendix B.

An additional endpoint is defined.The "at least one" (alo) endpoint is defined for each

Multivariate method performance on the simulated dataset

	Distributional FS		Combined FS		Ranking FS	
	FSP	FSC	FSP	FSC	FSP	FSC
y_1	0,562	0,857	1	0,571	1	0,286
y_2	0,615	0,125	1	0,125	0,8	0,375
y_3	0,611	0,75	0,857	0,5	0,8	0,25
y_4	0,357	1	1	0,875	1	0,25
y_5	0,375	1	1	0,857	1	0,57
alo			0,688	0,355		
alo_union			1	0,55		

Table 3.3: **Multi-outcome method performance on simulation dataset.** For each of the considered outcomes, metrics evaluation of the selection via Distributional FS, Ranking FS, and Combined methodology is presented. The last two rows present the result of the best working algorithm on selection for a comprehensive endpoint by the direct definition of at least one endpoint, as the $max_i\ y_i$, and the union of the selected features for each outcome.

sample as the maximum among all the other outcomes.

$$y_{alo} = max_{i=0}^{5} y_i$$

If an observation presents any of the considered endpoints then $y_{alo} = 1$ otherwise $y_{alo} = 0$. The endpoint is introduced to represent the presence of overall toxicity. Aiming at explaining y_{alo} exploiting a univariate algorithm the DSAEE is trained on the null class across all target dimensions, as for the other endpoints, and tested on cases presenting at least one endpoint. Finally, in both the multivariate and the univariate methods, the union of all the selected covariates for each endpoint is evaluated in predicting a comprehensive endpoint. The multivariate selection for overall toxicity can then be compared to the univariate one. To quantify and evaluate the selection ability of the methods, metrics presented in Table 2.2 are adopted to evaluate the feature selection for each outcome and for the comprehensive endpoint. In Table 3.3 metrics evaluation is reported in the case of multivariate selection, while in Table 3.4 metrics evaluation is reported in the case of the comparison univariate methodology.

The first result emerging from the analysis concerns the comparison between the various selection method considered. Both within the multivariate and univariate selection the combined selection is the best-performing one. As observed in the previous section, the distribution FS presents good performances in FSC but defines a set of covariates too large

Univariate method performance on the simulated dataset

	Distributional FS		Combined FS		Ranking FS	
	FSP	FSC	FSP	FSC	FSP	FSC
y_1	0,394	0,714	0,7	0,429	0,7	0,286
y_2	0,444	0	0,666	0	0,5	0,143
y_3	0,4545	0,571	1	0,286	0,6	0,429
y_4	0,25	0,875	1	0,5	1	0,625
y_5	0,21	0,5	0,5	0	0,3	0
alo_union			0,737	0,452		

Table 3.4: **Univariate method performance on simulation dataset.** For each of the considered outcomes, metrics evaluation of the selection via Distributional FS, Ranking FS, and Combined methodology is presented. The last row presents the result of the best working algorithm on selection for a comprehensive endpoint by the union of the selected features for each outcome.

and therefore includes insignificant covariates among those selected; the combined model instead, achieves the same FSP as the ranking method with an improvement in FSC. These results consolidate the conclusions from the previous simulation study. The comparison between multivariate and univariate selection methodology is performed focusing on the best-performing selection, namely the combined FS.

From the comparison, it is possible to observe that the multivariate model shows an improvement in FSP in almost every endpoint, and when the FSP is the same the FSC metrics increase, implying a selection focused on the discovery of every and only significant feature that better avoid those linked to correlated endpoints. One of the objectives in developing a multi-outcome feature selection is to be able to define a set of features informative about general radiosensitivity. It is possible to observe that the union of selected variables for individual endpoints identifies a set of covariates more informative for y_{alo} with respect to the selection performed univariately on the y_{alo} endpoint and the union set of variants chosen in the univariate case. These analyses validate the hypothesis that the multivariate method improves the selection of variables for individual outcomes, avoiding the pitfalls of a univariate selection when a strong correlation exists between endpoints. Indeed, the independent univariate selection might identify as predictive for a certain target, features that are actually determinants for a correlated target. This, while in principle still granting an acceptable predictive power of the selected features, may affect the interpretation of the underlying generative mechanism. In the context of genetic studies, this would translate into false discoveries of the biological interactions determining the phenotype of interest.

3.2. Case study: REQUITE data

As mentioned in the introduction, the selection and discovery of genomic variants predictive of late toxicities can inform downstream models such as PRSs and NTCPs. Therefore, in this section, the case study application of the proposed algorithm on the REQUITE breast cancer Cohort is presented. The features selected via the multivariate method described in Section 2.4 is finalized to introduce genetic information in the NTCP modes together with clinical covariates. Based on the selected SNPs, PRSs are defined for each patient, and an attempt is made to introduce genetic information through the scores in the personalized risk models.

The dataset employed in this analysis is presented in Section 1.2.2 To recap two different datasets are available, each containing N = 599 triplets (genome, boost, endpoints)

$$\mathbf{D} = \{(\underline{x}_1, z_1, \underline{y}_1), \dots, (\underline{x}_N, z_N, \underline{y}_N)\}$$

and

$$\tilde{\mathbf{D}} = \{(\underline{\tilde{x}}_1, z_1, \underline{y}_1), \dots, (\underline{\tilde{x}}_N, z_N, \underline{y}_N)\}$$

where for every patient i belonging to $\{1, \dots, 599\}$, $\underline{\tilde{x}_i}$ belonging to R^{122} is the vector containing values in [0,2] of the 122 SNPs considered; $\underline{x_i}$ belonging to $\{0, 1, 2\}^{122}$ is the vector containing rounded values of the 122 SNPs considered; $\underline{y_i}$ belongs to $\{0, 1\}^6$ and it's the vector describing the presence or absence for each of the 6 endpoints evaluated, while z_i belonging to $\{0, 1\}$ describes the delivery or not of an additional radiotherapy dose. In the analysis conducted, two groups of patients are considered based on whether or not they received an additional dose of radiation therapy. Those who did are referred to in the following as "boost" group and the others are "no boost" group. The pool of genetic features to select from included 122 variants previously identified in the literature as correlated to radio-induced late toxicities in breast cancer patients. The features selected via the proposed multivariate method are meant to be exploited to construct a PRS for breast cancer late toxicities. Therefore the algorithm is applied to a multivariate target combining the six (highly correlated) late toxicity endpoints. This resulted in six dimension-specific sets of selected SNP exploited independently to build six different PRSs (one for each late toxicity endpoint). The PRSs are computed following the PRSi algorithm presented in [22] that exploits FIM (Frequent Itemset Mining) routines to create a list of possible significant interactions and builds the score by weighting the contribution of each interaction term accordingly to the weights obtained when fitting a logistic regression model with the considered endpoint as the outcome, more details can be found in Section 1.3.5. The computed scores are ultimately added to an NTCPs model involving influential clinical covariates selected in an affiliated work, within the RADprecise project, aiming at predicting long-term side effects from clinical information. In the analyses, the prediction ability of the PRSi logistic regression is evaluated to estimate the selection ability of the multivariate FS introduced in the thesis. Moreover, the prediction improvement in the addition of the genomic score to clinical risk models is assessed by comparing the performance of NTCPs including genetic information to those based only on clinical information.

Analyses

The multivariate feature selection is performed following Algorithm 2.4. The common control sample definition is refined. Only patients in the boost group are eligible to be part of the control sample. This restriction, in agreement with the clinical counterpart in the RADprecise process, is necessary to avoid selecting patients in whom toxicities had not presented because of the scarce radiation dose. If those patients are included, there is a high chance of introducing risk patterns that require higher doses to express within the controls masking their effect in the case sample.

Given an ensemble of 10 DSAE, the autoencoders architecture is optimized in this phase based on the in-train metrics described in Table 2.3, while the λ parameter of the L1 loss is optimized to get the best separation between the reconstruction errors of the two classes. The separation is assessed via non-parametric tests reported in same Table. A grid search is performed with $\lambda \in \{1.0, 0.5, 0.25, 0.1, 0.05, 0.001, 0.0001, 0.00001, 0.000001, 0.0\}$. Details about autoencoders' architecture for each endpoint can be found in Table C.1 and C.2 in Appendix C. An example of the in-train accuracy and loss for the y^6 endpoint is reported in Figure 3.5 and Figure 3.4.

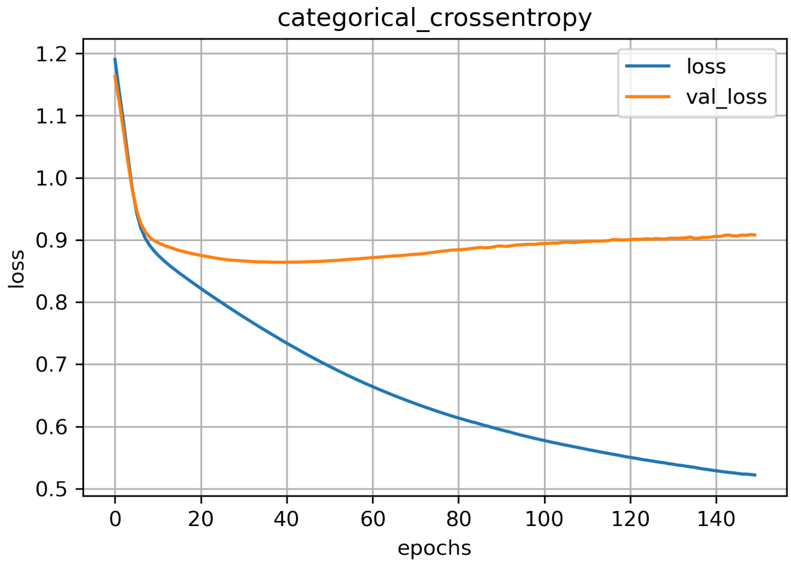

Figure 3.4: **In-train loss of the DSAEE** for y^6.In the DSAEE the loss is evaluated via cross-entropy since the autoencoder output distribution is compared to the rounded categorical value. The training was performed excluding a validation set to mimic the performance on unseen data. The loss on the validation sets is also presented

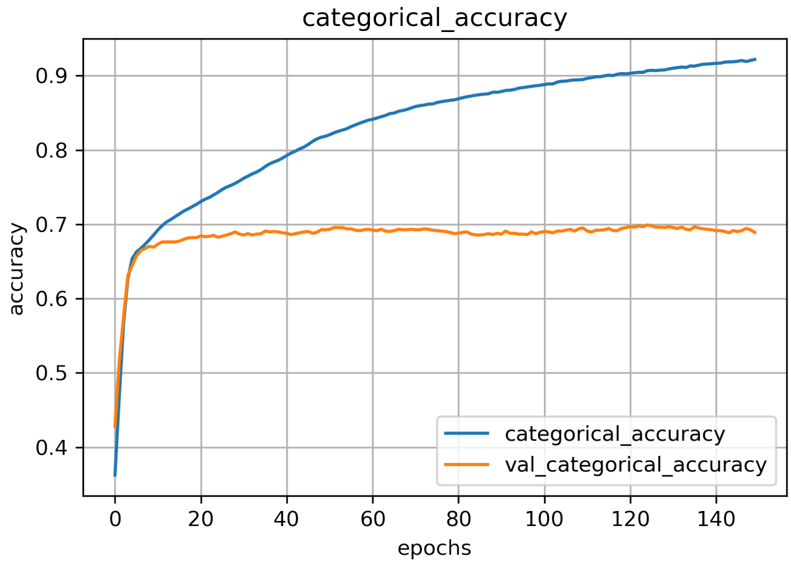

Figure 3.5: **In-train accuracy of the DSAEE** for y^6. The training was performed excluding a validation set to mimic the performance of the autoencoder on unseen data and the accuracy , measured during training, is plotted for both the training and the validation set

The final variable selection is performed by the distributional selection methods producing a final set of informative features for each endpoint.

The integrity of the dataset is then divided into a train and a test set. The train set is exploited to train and optimize the parameters of the risk models fitted with PRSi algorithm while the test set is exploited to evaluate the performance of the fitted model in the end. The user-defines parameters of the PRSi, namely the threshold γ for high-frequency appearing patterns in the FIM routine, the K selected patterns and the threshold ES, for the accuracy at which the DSAE training is stopped, are optimized via cross-validation. The threshold γ ranges in
$\{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$, K in $\{3, 5, 10, 15, 20, 25, 30, 35, 40\}$ and ES in $\{0.8, 0.85, 0.90, 1\}$. Details of the chosen parameter for each endpoint are presented in Table C.3 in Appendix C.

The PRSi prediction models are computed both for the entire training set and for the two groups separately. Those are evaluated via metrics defined in Table 2.1. Starting from the set of selected features a Lasso and a Ridge regression are computed to assess the importance of interaction inclusion in the prediction. The AUC of these models is reported with PRSi prediction model performance. Results are presented in Table 3.5 for the entire train set, in Table 3.6 for the "no boost " group and in Table 3.7 for the "boost" group. All the reported results are computed in cross-validation. An example of the patterns selected by the PRSi algorithm for y^6 is reported in Figure 3.6.

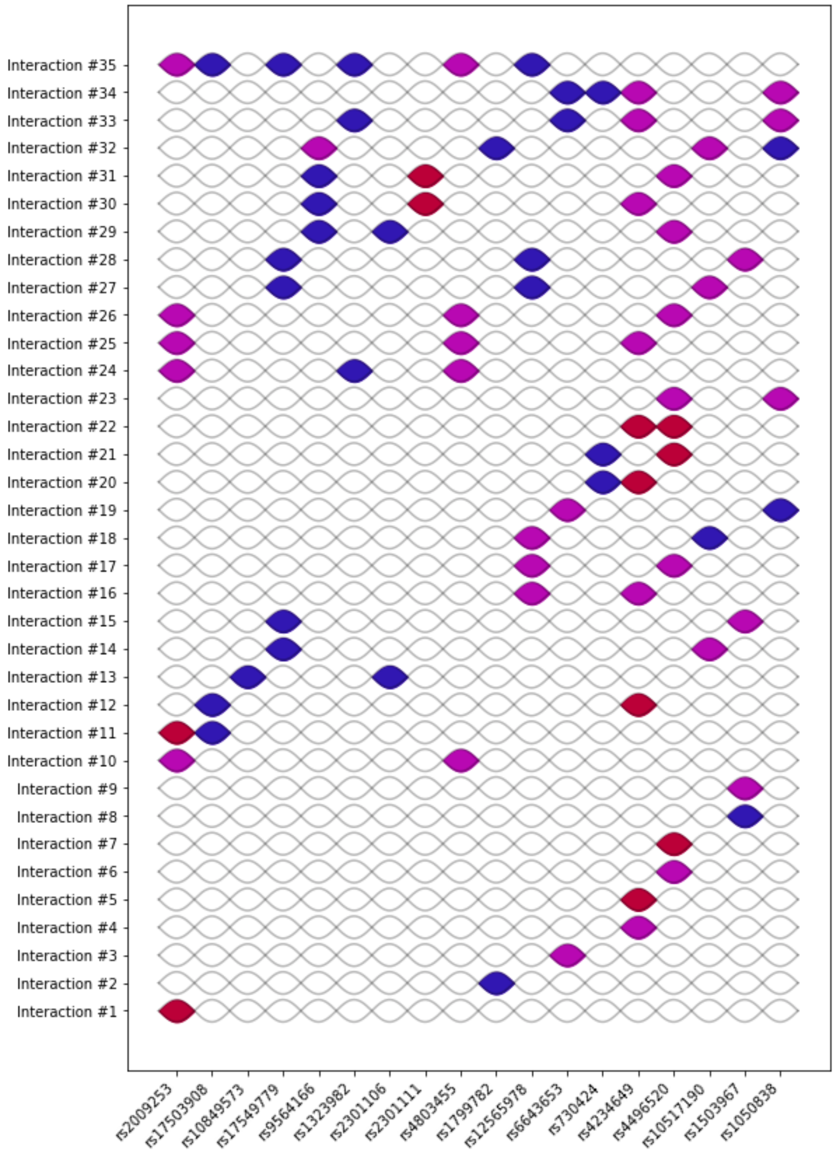

Figure 3.6: **Illustration of patterns selected** for y^6. Interactions selected by the PRSi algorithm to predict y^6 are represented in order. High-order interactions with four and six SNPs are selected for the description of the toxicity uprising.

To assess the performance of the feature selection and the ability of PRSi algorithm to define the most informative high-order interactions for endpoint prediction, its performance is compared to state-of-the-art models. A Random Forest and an XGboost algorithm are exploited to create prediction models for each endpoint from the initial set of literature-identified SNP. These models usually work very well in classification problems where high-order interactions between covariates are relevant but they could lose predictive ability when the classes are unbalanced, as in our case. The performance of the state-of-the-art models in cross-validation, in terms of AUC, can be found in Table 3.8

PRS model performance computed via hiPRS

Endpoint	AUC	AP	f1	sens	spe	NPV	Lasso_AUC	Ridge_AUC
y^1	**0.6**	0.55	0.48	0.42	0.81	**0.616**	0.59	0.57
y^2	**0.53**	0.23	0.3	0.58	0.58	**0.873**	0.49	0.5
y^3	**0.64**	0.18	0.29	0.78	0.55	**0.969**	0.54	0.54
y^4	**0.64**	0.34	0.34	0.64	0.61	**0.915**	0.68	0.68
y^5	**0.6**	0.23	0.28	0.64	0.56	**0.928**	0.58	0.56
y^6	**0.76**	0.36	0.5	0.77	0.89	**0.98**	**0.41**	**0.44**

Table 3.5: **PRS model performance computed via hiPRS on the entire training set.** For each endpoint a PRS model is computed and evaluated via metrics presented in Table 2.2. In addition, two linear models are fitted considering the same set of SNPs selected. The AUC of these models is reported to assess the importance of high-order interactions. All the reported results are computed in cross-validation.

PRS model performance computed via hiPRS on no boost group

Endpoint	AUC	AP	f1	sens	spe	NPV	Lasso_AUC	Ridge_AUC
y^1	**0,64**	0,582	0,7	0,82	0,6	**0.80**	0,58	0,57
y^2	**0,61**	0,4	0,43	0,32	0,56	**0,814**	0,15	0,18
y^3	**0,68**	0,45	0,51	0,73	0,7	**0.917**	0,61	0,63
y^4	**0,91**	0,67	0,77	1	0,91	**1**	0,97	0,73
y^5	**0,93**	0,6	0,7	1	0,87	**1**	0,41	0,33
y^6	_	_	_	_	_	_	_	_

Table 3.6: **PRS model performance computed via hiPRS on the set of patients that didn't receive the boost dose.** For each endpoint a PRS model is computed and evaluated via metrics presented in Table 2.2. In addition, two linear models are fitted considering the same set of SNPs selected. The AUC of these models is reported to assess the importance of high-order interactions. All the reported results are computed in cross-validation.

PRS model performance computed via hiPRS on boost group

Endpoint	AUC	AP	f1	sens	spe	NPV	Lasso_AUC	Ridge_AUC
y^1	**0.6**	0.55	0.48	0.42	0.81	**0.616**	0.59	0.57
y^2	**0.53**	0.23	0.3	0.58	0.58	**0.873**	0.49	0.5
y^3	**0.64**	0.18	0.29	0.78	0.55	**0.969**	0.54	0.54
y^4	**0.64**	0.34	0.34	0.64	0.61	**0.915**	0.68	0.68
y^5	**0.6**	0.23	0.28	0.64	0.56	**0.928**	0.58	0.56
y^6	**0.76**	0.36	0.5	0.77	0.89	**0.98**	**0.41**	**0.44**

Table 3.7: **PRS model performance computed via hiPRS on the set of patients that received the additional dose.** For each endpoint a PRS model is computed and evaluated via metrics presented in Table 2.2. In addition, two linear models are fitted considering the same set of SNPs selected. The AUC of these models is reported to assess the importance of high-order interactions. All the reported results are computed in cross-validation.

Once PRS are defined, it is possible to associate each patient, of both training and test set, its score. This score can be used directly to predict the endpoint or in combination with other clinical covariates. Three risk models can then be defined: (i) a model based solely on clinical covariates correlated with the selected endpoint, (ii) one solely on the PRS score, and (iii) the last combining the clinical and genomic information. The models are fitted on the training set and validated on the test set, measuring their performance throw the metrics described in Table 2.1. Results of model (i) are reported for the training set in Table 3.9 and for the test set in Table 3.10; results of model (ii) are reported for the training set in Table 3.11 and for the test set in Table 3.12; and results of model (iii) are reported for the training set in Table 3.13 and for the test set in Table 3.14. The same three risk models are also fitted focusing exclusively on the prediction of the "boost group". The results can be found in Appendix C. Here are reported only model (i) and (iii) performance on the test set, resp. in Table 3.15 and in Table 3.16, to perform a comparison.

Results

The feature selection method developed performance can be assessed through the performance of the PRSi prediction models. The selection seems to work quite well on the full dataset. The multivariate FS is performed on the entire cohort since genetic variants from which toxicities arise should be radiation-dose independent. However, their expression and interactions could be radiation-dose related [16] opening the possibility to create

Prediction performance of the state-of-the-art model

Endpoint	RF_AUC	XGB_AUC
y^1	0,506	0,503
y^2	0,5	0,472
y^3	0,5	0,426
y^4	0,5	0,496
y^5	0,5	0,485
y^6	0,5	0,473

Table 3.8: **Prediction performance of the Random Forest and XGboos classifier applied with the aim of predicting the endpoint** from the initial literature-identified SNPs. The performance is assessed via AUC computation. All the reported results are computed in cross-validation.

NTCP model performance including clinical covariates on training set

Endpoint	AUC	AUC (95% CI)	prec	f1	sens	spe	NVP
y^1	0.743	(0.697, 0.790)	0.618	0.690	0.781	0.572	0.747
y^2	0.751	(0.676, 0.826)	0.284	0.4	0.674	0.757	0.942
y^3	0.731	(0.648, 0.814)	0.139	0.235	0.75	0.617	0.967
y^4	0.757	(0.673, 0.841)	0.224	0.328	0.611	0.805	0.957
y^5	0.785	(0.688, 0.883)	0.159	0.264	0.777	0.729	0.980
y^6	0.871	(0.790, 0.952)	0.126	0.217	0.785	0.814	0.991

Table 3.9: **NTCP model performance including clinical covariates on the training set.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

NTCP model performance including clinical covariates on test set

Endpoint	AUC	AUC (95% CI)	prec	f1	sens	spe	NVP
y^1	0.724	(0.626, 0.821)	0.637	0.698	0.770	0.618	0.755
y^2	0.766	(0.638, 0.893)	0.318	0.424	0.636	0.802	0.938
y^3	0.552	(0.293, 0.811)	0.117	0.19	0.5	0.708	0.948
y^4	0.729	(0.521, 0.938)	0.384	0.434	0.5	0.919	0.948
y^5	0.653	(0.394, 0.911)	0.166	0.258	0.571	0.807	0.965
y^6	0.863	(0.718, 1)	0.12	0.214	1	0.784	1

Table 3.10: **NTCP model performance including clinical covariates on the test set.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

Prediction model performance including exclusively PRS score on train set

Endpoint	AUC	AUC (95% CI)	prec	f1	NVP	sens	spe
y^1	0.565	(0.519, 0.612)	0.55	0.478	0.578	0.422	0.697
y^2	0.835	(0.782, 0.888)	0.422	0.517	0.945	0.666	0.864
y^3	0.692	(0.605, 0.779)	0.127	0.220	0.972	0.824	0.525
y^4	0.728	(0.654, 0.803)	0.147	0.25	0.973	0.842	0.540
y^5	0.772	(0.688, 0.856)	0.162	0.256	0.967	0.607	0.786
y^6	0.974	(0.959, 0.989)	0.283	0.441	1	1	0.91

Table 3.11: **Prediction model performance including exclusively PRS score on the train set.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

Prediction model performance including exclusively PRS score on test set

Endpoint	AUC	AUC (95% CI)	prec	f1	sens	spe	NVP
y^1	0.654	(0.568, 0.740)	0.25	0.217	0.192	0.49	0.408
y^2	0.498	(0.332, 0.664)	0.0968	0.156	0.4	0.417	0.816
y^3	0.513	(0.333, 0.694)	0.089	0.156	0.625	0.505	0.945
y^4	0.499	(0.307, 0.690)	0.042	0.059	0.1	0.772	0.896
y^5	0.727	(0.592, 0.863)	0	0	0	0.606	0.9
y^6	0.447	(0.147, 0.748)	0.042	0.078	0.5	0.579	0.968

Table 3.12: **Prediction model performance including exclusively PRS score on the test set.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

NTCP model performance including clinical covariates and PRS score on training set

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV	prs p-value
y^1	0.753	(0.707, 0.799)	0.643	0.693	0.751	0.630	0.740	0.0181
y^2	0.894	(0.844, 0.943)	0.54	0.635	0.767	0.907	0.965	4.81e-10
y^3	0.804	(0.733, 0.875)	0.19	0.312	0.875	0.693	0.985	0.0002
y^4	0.821	(0.750, 0.892)	0.217	0.339	0.777	0.741	0.973	4.76e-05
y^5	0.895	(0.843, 0.947)	0.214	0.345	0.888	0.785	0.99	5.85e-06
y^6	0.989	(0.9769, 1)	0.359	0.528	1	0.939	1	5.26e-05

Table 3.13: **NTCP model performance including clinical covariates and PRS score on the training set.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2. The last column contains the p-value of the coefficient test, namely $H_0 : \beta_{PRS} = 0$ vs $H_0 : \beta_{PRS} \neq 0$.

NTCP model performance including clinical covariates and PRS score on test set

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV
y^1	0.670	(0.566, 0.774)	0.623	0.653	0.688	0.636	0.7
y^2	0.709	(0.526, 0.893)	0.28	0.389	0.636	0.763	0.935
y^3	0.522	(0.245, 0.798)	0.106	0.18	0.625	0.592	0.953
y^4	0.681	(0.451, 0.910)	0.219	0.333	0.7	0.747	0.961
y^5	0.464	(0.235, 0.694)	0.085	0.148	0.571	0.587	0.953
y^6	0.611	(0.264, 0.957)	0.074	0.133	0.666	0.755	0.987

Table 3.14: **NTCP model performance including clinical covariates and PRS score on the test set.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

NTCP model performance including clinical covariates on the test set of boost group

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV
y^1	0.712	(0.598, 0.825)	0.659	0.698	0.743	0.634	0.722
y^2	0.748	(0.611, 0.885)	0.258	0.4	0.888	0.634	0.975
y^3	0.506	(0.204, 0.808)	0.111	0.19	0.666	0.609	0.961
y^4	0.703	(0.485, 0.921)	0.333	0.417	0.555	0.870	0.943
y^5	0.549	(0.267, 0.830)	0.043	0.08	0.5	0.195	0.842
y^6	0.862	(0.713, 1)	0.15	0.261	1	0.779	1

Table 3.15: **NTCP model performance including clinical covariates on the test set of boost group.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

NTCP model performance including clinical covariates and PRS score on the test set of boost group

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV
y^1	0.689	(0.572, 0.805)	0.646	0.713	0.795	0.585	0.75
y^2	0.674	(0.495, 0.853)	0.263	0.357	0.555	0.777	0.924
y^3	0.593	(0.347, 0.838)	0.125	0.2	0.5	0.744	0.953
y^4	0.638	(0.429, 0.846)	0.375	0.352	0.333	0.935	0.923
y^5	0.559	(0.265, 0.853)	0.216	0.043	0.08	0.5	0.842
y^6	0.771	(0.674, 0.867)	0.13	0.23	1	0.74	1

Table 3.16: **NTCP model performance including clinical covariates and PRS score on the test set of boost group.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

different risk models for patient groups receiving different radiation doses. Indeed, PRSi predictive models improve their performance when different selections are performed between the two groups of patients. The introduction of performance measures focusing on the prediction of the minority class exposes the difficulty of the logistic regression model in classifying correctly patients presenting toxicities. This can be expected since imbalanced class distribution can greatly reduce the predictive power of a binary logistic regression and entail poor predictive performance, especially for the minority class [44]. The ability of the model in detecting the minority class could be improved using imbalance learning methods as cost-sensitive models, that typically introduce different costs for the majority and minority class observations. Notably, the NPV, the percentage of true negatives among those classified as negative, is quite high in all models. In the radiogenic context, a characteristic fundamental for the model to be practicable is the patient's ability to avoid misclassification of patients as non-radiosensitive.

The difference between AUC performance in the classical linear models and PRSi assesses the importance of interaction in the genomic context. An example can be found in endpoint y^6 where the ability of the developed algorithm to consider high-order interaction in feature selection is fundamental in performing a good prediction. This can be assessed also from Figure [**?**] where the patterns selected by the PRSi algorithm present interactions composed of three or more SNPs.

State-of-the-art models are introduced as comparison methods and to have a first measure of how informative the genomic covariates are for each endpoint. The low results obtained can be caused by a low signal-to-noise ratio and by the imbalanced and imputed nature of the data.

Clinical models perform quite well on both the train and test set but, as PRSi models, aren't able to accurately distinguish the minority set. Sensitivity, specificity, and NPV are though pretty good. PRS scores are able to predict well the outcome on the train set but lose performance on the test set. The pronounced difference is probably due to overfitting. Only 320 samples are available in the training set and optimization of the parameters is performed in cross-validation reducing further the set of starting information. A larger dataset is needed to perform a more robust analysis. Moreover, the addition of the PRS score in the clinical NTCP model is significant if the p-value of the coefficient test is considered, but does not improve the performance w.r.t. clinical models. The missing improvement could be due to the fact that the score is simply added to an already large set of clinical covariates. The models are not reduced and the attempt is simply to add to the contribution of the correct clinical covariates a genetic contribution. A better way to integrate clinical and genomic data should be studied. The presence of many covariates lead also to the overfitting that can be observed in the gap between the training and test performance of the complete model.

4 | Conclusions and Discussion

The innovation of this work is the development of a methodology able to detect the most important genomic features in a multi-outcome setting. The proposed method builds upon the original work in [39], where an ensemble of anomaly detection AEs (i.e, the DSAEE algorithm) is exploited to select predictive features to discriminate between classes. In this work, the DSAEE is extended to allow FS for multivariate binary outcomes and enriched with a denoising technique that robustifies the analysis to imputation error in genomic data. Similarly to its predecessor, the method developed is designed to overcome the challenges imposed by the peculiar setting of genomic studies. In particular, it is meant to tackle, class imbalance derived from the study of rare traits, and the need to account for predictive high-order interactions among features, due to the complex biological mechanisms determining phenotypic traits. The developed methodology can be applied to clinical case studies with the aim of identifying SNPs associated with late-toxicity endpoints and including genetic information in personalized risk models, such as NTCPs.

Based on simulation studies, we can say that the developed model succeeds in improvement in the representation of noisy data thanks to the denoising technique. Moreover, the inclusion of the intra-endpoint correlation structure in the FS improves the accuracy in the selection of highly influential features that provide intrinsic information and discriminant properties for class separability. The accurate definition of influential feature for the specific toxicity can be fruitful for an interpretation of biologically relevant variants. Indeed, the model, in addition to selection, can be exploited for the discovery of influential genetic variants or validation of variants previously identified in the literature as correlated to radio-induced toxicities. The developed method, thanks to a well-performing FS, can improve the definition of genetic predisposition to general toxicities and can be employed by physicians to take more informed individual decisions in cancer treatment. The importance of the model lies in its clinical applicability.

The method can be generalized to all contexts where it is necessary to perform a multivariate FS with unbalanced classes and similar data characteristics.

Some of the limitations of the developed model are the need for ground truth definition of noisy input data and the difficulty to scale in input features due to the high computational cost.

The full methodology is applied to the REQUITE Breast Cancer Cohort. Case study results show the ability of the developed method in selecting informative features evaluating the impact of high-order interaction. The selection is finalized to introduce genetic information in the NTCP modes together with clinical covariates. Based on the selected SNPs, PRSs are defined for each patient, and an attempt is made to introduce genetic information through the scores in the personalized risk models. However, the significance of the genetic information in describing the presence of toxicity is notably reduced when summarized in a single score and the addition of the scores to the clinical-based risk models does not improve the ability of the models to distinguish radiosensitive patients.

The limitation in the available data, the presence of high-dimensional clinical covariates, and the employment of logistic regression impact the performance of the NTCP model containing both clinical and genetic covariates. The strong overfitting in the train set hield a model not up to standards. Moreover, better methodologies to integrate clinical and genetic variants need to be implemented. A possibility is to conduct feature selection of SNPs independently from clinical data, capturing those variants able to introduce information uncorrelated with clinical covariates. This could be achieved by applying the developed methodology to groups of control and cases defined as misclassified radiosensitive patients and correctly classified patients.

This work paves the way for several further developments. Variational autoencoders have already been proposed for anomaly detection [3, 23]. Anomaly detection in this case is performed considering the reconstruction probability, a probabilistic measure that takes into account the variability of the distribution of covariates. It has a theoretical background making it a more principled and objective anomaly score than the reconstruction error considered here [3]. However, variation autoencoders require a large training dataset that hinders its applicability in this thesis work. Improvement can be done also on the denoising characteristic of the autoencoder. Denoising terms can be introduced directly in AEs loss avoiding the need to define or approximate the true value of the considered features [24]. In this work, we chose to use a denoising method that with good performance was more interpretable and controllable. Finally, it would be interesting to further develop the multivariate setting developing a full multivariate methodology for multiple endpoint PRS definition.

Bibliography

[1] D. Abshire and M. K. Lang. The Evolution of Radiation Therapy in Treating Cancer. *Seminars in Oncology Nursing*, 34(2):151–157, May 2018. doi: 10.1016/j.soncn.2018.03.006.

[2] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. 2015.

[3] J. An and S. Cho. Variational autoencoder based anomaly detection using reconstruction probability. *Special Lecture on IE*, 2(1):1–18, 2015.

[4] C. N. Andreassen, B. S. Rosenstein, S. L. Kerns, et al. Individual patient data meta-analysis shows a significant association between the ATM rs1801516 SNP and toxicity after radiotherapy in 5456 breast and prostate cancer patients. *Radiotherapy and Oncology*, 121(3):431–439, Dec. 2016. doi: 10.1016/j.radonc.2016.06.017.

[5] Y. S. Aulchenko, M. V. Struchalin, and C. M. van Duijn. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*, 11(1):134, Dec. 2010. doi: 10.1186/1471-2105-11-134.

[6] M. Avanzo, J. Stancanello, M. Trovò, et al. Complication probability model for subcutaneous fibrosis based on published data of partial and whole breast irradiation. *Physica medica: PM: an international journal devoted to the applications of physics to medicine and biology: official journal of the Italian Association of Biomedical Physics (AIFB)*, 28(4):296–306, Oct. 2012. doi: 10.1016/j.ejmp.2011.11.002.

[7] D. Bank, N. Koenigstein, and R. Giryes. Autoencoders, Apr. 2021. arXiv:2003.05991 [cs, stat].

[8] F. Bao, Y. Deng, and Q. Dai. Acid: Association correction for imbalanced data in gwas. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15 (1):316–322, 2018. doi: 10.1109/TCBB.2016.2608819.

[9] G. C. Barnett, C. E. Coles, R. M. Elliott, et al. Independent validation of genes and polymorphisms reported to be associated with radiation toxicity: a

prospective analysis study. *The Lancet Oncology*, 13(1):65–77, Jan. 2012. doi: 10.1016/S1470-2045(11)70302-3.

[10] G. C. Barnett, R. M. Elliott, J. Alsner, et al. Individual patient data meta-analysis shows no association between the SNP rs1800469 in TGFB and late radiotherapy toxicity. *Radiotherapy and Oncology*, 105(3):289–295, Dec. 2012. doi: 10.1016/j.radonc.2012.10.017.

[11] G. C. Barnett, D. Thompson, L. Fachal, et al. A genome wide association study (GWAS) providing evidence of an association between common genetic variants and late radiotherapy toxicity. *Radiotherapy and Oncology*, 111(2):178–185, May 2014. doi: 10.1016/j.radonc.2014.02.012.

[12] V. W. Berger and Y. Zhou. Kolmogorov–Smirnov Test: Overview. In N. Balakrishnan, T. Colton, B. Everitt, et al., editors, *Wiley StatsRef: Statistics Reference Online*. Wiley, 1 edition, Sept. 2014. ISBN 978-1-118-44511-2. doi: 10.1002/9781118445112.stat06558.

[13] R. M. Buckley, A. C. Harris, G.-D. Wang, et al. Best practices for analyzing imputed genotypes from low-pass sequencing in dogs. *Mammalian Genome*, 33(1):213–229, Mar. 2022. doi: 10.1007/s00335-021-09914-z.

[14] C. Burman, G. Kutcher, B. Emami, and M. Goitein. Fitting of normal tissue tolerance data to an analytic function. *International Journal of Radiation Oncology*Biology*Physics*, 21(1):123–135, May 1991. doi: 10.1016/0360-3016(91)90172-Z.

[15] M. Ceccarelli, editor. *Bioinformatica: sfide e prospettive*. Number 321 in Collana RCOST / Tecnologie del software. FrancoAngeli, Milano, 2006. ISBN 978-88-464-8278-5.

[16] J. A. Cesaretti, R. G. Stock, D. P. Atencio, et al. A Genetically Determined Dose–Volume Histogram Predicts for Rectal Bleeding among Patients Treated With Prostate Brachytherapy. *International Journal of Radiation Oncology*Biology*Physics*, 68(5):1410–1416, Aug. 2007. doi: 10.1016/j.ijrobp.2007.02.052.

[17] A. Cobat, L. Abel, A. Alcaïs, and E. Schurr. A General Efficient and Flexible Approach for Genome-Wide Association Analyses of Imputed Genotypes in Family-Based Designs: Efficient and Flexible Approach for Genome-Wide Association Analyses. *Genetic Epidemiology*, 38(6):560–571, Sept. 2014. doi: 10.1002/gepi.21842.

[18] I. C. Covert and S.-I. Lee. DEEP UNSUPERVISED FEATURE SELECTION. page 20.

[19] X. Dai, G. Fu, S. Zhao, and Y. Zeng. Statistical Learning Methods Applicable to Genome-Wide Association Studies on Unbalanced Case-Control Disease Data. *Genes*, 12(5):736, May 2021. doi: 10.3390/genes12050736.

[20] S. De Langhe, T. Mulliez, L. Veldeman, et al. Factors modifying the risk for developing acute skin toxicity after whole-breast intensity modulated radiotherapy. *BMC Cancer*, 14(1):711, Dec. 2014. doi: 10.1186/1471-2407-14-711.

[21] S. Feng and M. F. Duarte. Graph autoencoder-based unsupervised feature selection with broad and local data structure preservation. *CoRR*, abs/1801.02251, 2018.

[22] N. R. Franco, M. C. Massi, F. Ieva, et al. Development of a method for generating SNP interaction-aware polygenic risk scores for radiotherapy toxicity. *Radiotherapy and Oncology*, 159:241–248, June 2021. doi: 10.1016/j.radonc.2021.03.024.

[23] S. Givnan, C. Chalmers, P. Fergus, et al. Anomaly Detection Using Autoencoder Reconstruction upon Industrial Motors. *Sensors*, 22(9):3166, Apr. 2022. doi: 10.3390/s22093166.

[24] S. Gupta and A. Gupta. Dealing with Noise Problem in Machine Learning Datasets: A Systematic Review. *Procedia Computer Science*, 161:466–474, 2019. doi: 10.1016/j.procs.2019.11.146.

[25] D. M. Hawkins. *Identification of Outliers*. Springer Netherlands, Dordrecht, 1980. ISBN 978-94-015-3994-4. OCLC: 851385856.

[26] B. Howie, C. Fuchsberger, M. Stephens, et al. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8):955–959, July 2012. doi: 10.1038/ng.2354.

[27] N. C. Institute. Common Terminology Criteria for Adverse Events (CTCAE). (`https://ctep.cancer.gov/protocoldevelopment/electronic_applications/ctc.htm`). [Online; accessed 12-October-2022].

[28] T. Ishihara and K. Yamamoto. A Test for Multiple Binary Endpoints with Continuous Latent Distribution in Clinical Trials. *Journal of Statistical Theory and Applications*, 20(4):463–480, Dec. 2021. doi: 10.1007/s44199-021-00003-3.

[29] J. Kang, R. Schwartz, J. Flickinger, and S. Beriwal. Machine learning approaches for predicting radiation therapy outcomes: A clinician's perspective. *International Journal of Radiation Oncology*Biology*Physics*, 93(5):1127–1135, 2015. doi: https://doi.org/10.1016/j.ijrobp.2015.07.2286.

[30] S. L. Kerns, C. M. L West, C. N. Andreassen, et al. Radiogenomics: the search for genetic predictors of radiotherapy response. *Future Oncology*, 10(15):2391–2406, Dec. 2014. doi: 10.2217/fon.14.173.

[31] S. L. Kerns, S. Kundu, J. H. Oh, et al. The Prediction of Radiotherapy Toxicity Using Single Nucleotide PolymorphismBased Models: A Step Toward Prevention. *Seminars in Radiation Oncology*, 25(4):281–291, Oct. 2015. doi: 10.1016/j.semradonc.2015.05.006.

[32] D. Krebsforshungszentrum. RADprecise. (`https://www.dkfz.de/en/epidemiologieâĂŞkrebserkrankungen/units/genepi/ge_pr13_RADprecise.htm/`), n.d. [Online; accessed 9-February-2022].

[33] S. A. Lambert, G. Abraham, and M. Inouye. Towards clinical utility of polygenic risk scores. *Human Molecular Genetics*, 28(R2):R133–R142, Nov. 2019. doi: 10.1093/hmg/ddz187.

[34] E. Lee, C. Takita, J. L. Wright, et al. Genome-wide enriched pathway analysis of acute post-radiotherapy pain in breast cancer patients: a prospective cohort study. *Human Genomics*, 13(1):28, Dec. 2019. doi: 10.1186/s40246-019-0212-8.

[35] L. B. Marks, E. D. Yorke, A. Jackson, et al. Use of Normal Tissue Complication Probability Models in the Clinic. *International Journal of Radiation Oncology*Biology*Physics*, 76(3):S10–S19, Mar. 2010. doi: 10.1016/j.ijrobp.2009.07.1754.

[36] M. C. Massi. *Patient Representations from Complex Biological Systems for Precision Medicine*. PhD thesis, Politecnico di Milano, 2022.

[37] M. C. Massi, F. Gasperoni, F. Ieva, et al. A Deep Learning Approach Validates Genetic Risk Factors for Late Toxicity After Prostate Cancer Radiotherapy in a REQUITE Multi-National Cohort. *Frontiers in Oncology*, 10:541281, Oct. 2020. doi: 10.3389/fonc.2020.541281.

[38] M. C. Massi, N. R. Franco, A. Manzoni, et al. Learning High-Order Interactions for Polygenic Risk Prediction. preprint, Genomics, Apr. 2022.

[39] M. C. Massi, F. Gasperoni, F. Ieva, and A. M. Paganoni. Feature selection for imbalanced data with deep sparse autoencoders ensemble. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):376–395, June 2022. doi: 10.1002/sam.11567.

[40] C. Mbah, H. Thierens, O. Thas, et al. Pitfalls in Prediction Modeling for Normal Tissue Toxicity in Radiation Therapy: An Illustration With the Individual Radiation

Sensitivity and Mammary Carcinoma Risk Factor Investigation Cohorts. *International Journal of Radiation Oncology, Biology, Physics*, 95(5):1466–1476, Aug. 2016. doi: 10.1016/j.ijrobp.2016.03.034.

[41] C. Mbah, K. De Ruyck, S. De Schrijver, et al. A new approach for modeling patient overall radiosensitivity and predicting multiple toxicity endpoints for breast cancer patients. *Acta Oncologica*, 57(5):604–612, May 2018. doi: 10.1080/0284186X.2017.1417633.

[42] N. H. G. R. I. (NHGRI). Polygenic Risk Scores. (`https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores.htm/`), n.d. [Online; accessed 10-October-2022].

[43] S. J. Pocock, N. L. Geller, and A. A. Tsiatis. The Analysis of Multiple Endpoints in Clinical Trials. *Biometrics*, 43(3):487, Sept. 1987. doi: 10.2307/2531989.

[44] S. Priselac. Outlier-robust Logistic Regression for Imbalanced Data. page 57.

[45] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[46] R. Ristl, S. Urach, G. Rosenkranz, and M. Posch. Methods for the analysis of multiple endpoints in small populations: A review. *Journal of Biopharmaceutical Statistics*, 29(1):1–29, 2019. doi: 10.1080/10543406.2018.1489402.

[47] D. B. Seal, V. Das, S. Goswami, and R. K. De. Estimating gene expression from DNA methylation and copy number variation: A deep learning regression model for multi-omics integration. *Genomics*, 112(4):2833–2841, July 2020. doi: 10.1016/j.ygeno.2020.03.021.

[48] P. Seibold, S. Behrens, P. Schmezer, et al. XRCC1 Polymorphism Associated With Late Toxicity After Radiation Therapy in Breast Cancer Patients. *International Journal of Radiation Oncology*Biology*Physics*, 92(5):1084–1092, Aug. 2015. doi: 10.1016/j.ijrobp.2015.04.011.

[49] P. Seibold, A. Webb, M. E. Aguado-Barrera, et al. REQUITE: A prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer. *Radiotherapy and Oncology*, 138:59–67, Sept. 2019. doi: 10.1016/j.radonc.2019.04.034.

[50] C. J. Talbot, G. A. Tanteles, G. C. Barnett, et al. A replicated association between polymorphisms near TNF and risk for adverse reactions to radiotherapy. *British Journal of Cancer*, 107(4):748–753, Aug. 2012. doi: 10.1038/bjc.2012.290.

[51] C. J. Talbot, G. A. Tanteles, G. C. Barnett, et al. A replicated association between polymorphisms near TNF and risk for adverse reactions to radiotherapy. *British Journal of Cancer*, 107(4):748–753, Aug. 2012. doi: 10.1038/bjc.2012.290.

[52] S. Terrazzino, P. La Mattina, G. Gambaro, et al. Common Variants of GSTP1, GSTA1, and TGF1 are Associated With the Risk of Radiation-Induced Fibrosis in Breast Cancer Patients. *International Journal of Radiation Oncology*Biology*Physics*, 83(2):504–511, June 2012. doi: 10.1016/j.ijrobp.2011.06.2012.

[53] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697.

[54] C. West, D. Azria, J. Chang-Claude, et al. The REQUITE Project: Validating Predictive Models and Biomarkers of Radiotherapy Toxicity to Reduce Side-effects and Improve Quality of Life in Cancer Survivors. *Clinical Oncology*, 26(12):739–742, Dec. 2014. doi: 10.1016/j.clon.2014.09.008.

[55] L. Yu, Z. Zhang, X. Xie, et al. Unsupervised Feature Selection Using RBF Autoencoder. In H. Lu, H. Tang, and Z. Wang, editors, *Advances in Neural Networks – ISNN 2019*, volume 11554, pages 48–57. Springer International Publishing, Cham, 2019. ISBN 978-3-030-22795-1 978-3-030-22796-8. doi: 10.1007/978-3-030-22796-8_6. Series Title: Lecture Notes in Computer Science.

A | Appendix A

Appendix A contains details about the REQUITE database. Available information is grouped into three main topics: toxicities, patients' history, and treatment data. The covariates are both continuous and categorical. For continuous variables, the minimum, maximum, 1st quantile, 3rd quantile, mean and median are reported. Categorical variables are summarized by the number of samples belonging to each category. Additional covariates are available as reported in [49].

Patient toxicities data summary

Atrophy	0 :5611	1 :2493	2 : 725	3 : 67	NA's: 361
Nipple retraction	0 :7824	1 : 893	2 : 92	NA's: 448	
Oedema	0 :7207	1 :1581	2 : 149	3 : 6	NA's: 314
Skin ulceration	0 :8723	1 : 153	2 : 47	3 : 14	NA's: 320
Telangiectasia tumour bed	0 :8711	1 : 213	2 : 15	NA's: 318	
Telangiectasia outside tumour bed	0 :8685	1 : 230	2 : 18	NA's: 324	
Skin induration tumour bed	0 :5287	1 :3099	2 : 489	3 : 43	NA's: 339
Skin induration outside tumour bed	0 :7905	1 : 894	2 : 107	3 : 16	NA's: 335
Erythema	0 :6908	1 :1577	2 : 466	3 : 29	NA's: 277
Arm lymphodema	0 :8703	1 : 245	2 : 32	3 : 2	NA's: 275
Skin hyperpigmentation	0 :7425	1 :1460	2 : 46	NA's: 326	
Pneumonitis	0 :8889	1 : 27	2 : 14	3 : 1	NA's: 326
Pain	0 :6108	1 :3090	NA's: 59	NA	
Pain severity	1 :2525	2 : 444	3 : 118	NA's:6170	
Swollen arm	0 :8603	1 : 574	NA's: 80	NA	

Table A.1: **Summury of late toxicities in the sample**

Patients' history data summary

Subject Id	Length:2034	Class :character			
Visit	Length:2034	Class :character			
Date of visit	Length:2034	Class :date			
Height (cm)	Min. :140.0	1st Qu.:158.0	Median :163.0	Mean :162.8	3rd Qu.:168.0
	Max.: 187.0	NA's:10			
Weight at cancer diagnosis (kg)	Min. : 36.00	1st Qu.: 60.00	Median : 68.00	Mean : 70.11	3rd Qu.: 78.00
	Max.: 187.00	NA's: 13			
Age at radiotherapy start (yrs)	Min. :23.00	1st Qu.:50.00	Median :58.00	Mean :58.25	3rd Qu.:66.00
	Max.: 90.00				
Bra cup size	Min. : 1.000	1st Qu.: 3.000	Median : 4.000	Mean : 3.902	3rd Qu.: 5.000
	Max.: 11.000	NA's: 68			
Band size	Min. : 1.000	1st Qu.: 5.000	Median : 6.000	Mean : 6.033	3rd Qu.: 7.000
	Max.: 10.000	NA's: 90			
Smoker	0 :1142	1 : 505	2 : 83	3 : 282	NA's: 22
Alcohol intake	0 :872	1 : 66	2 : 59	3 :980	NA's: 57
Menopausal status	1 : 491	2 :1364	3 : 150	NA's: 29	
Hormone replacement therapy	0 :1005	1 : 328	NA's: 701		
Diabetes	0:1908	1: 126			
History of heart disease	0 :1892	1 : 141	NA's: 1		
Ra	0:1976	1: 58			
Systemic lupus erythematosus	0:2030	1: 4			
Other collagen vascular disease	0:2020	1: 14			
Hypertension	0:1467	1: 567			
Depression	0:1795	1: 239			
Ace inhibitor	0 :1890	1 : 143	NA's: 1		
Amiodarone	0:2027	1: 7			
Analgesics	0:1834	1: 200			
Antidepressant	0:1796	1: 238			
Breast cancer family history	0 :1625	1 : 405	NA's: 4		
Radiotherapy toxicity family history	0 :1535	1 : 64	NA's: 435		
Ethnicity	1 :1913	7 : 23	5 : 16	3 : 15	17 : 11
	(Other): 50	NA's: 6			
Household income	2 : 312	3 : 240	4 : 164	1 : 107	5 : 92
	(Other): 85	NA's: 1034			
Household members	Min. :1.000	1st Qu.:2.000	Median :2.000	Mean :2.261	3rd Qu.:3.000
	Max.: 8.000	NA's: 424			
Previous malignancies	0 : 474	1 : 16	NA's:1544		

Table A.2: **Summury of patients' data**

Treatment data summary

Surgery	1 :2022	NA's: 12			
Surgery type	1 :1107	2 : 907	NA's: 20		
Axillary surgery	0 : 163	1 :1859	NA's: 12		
Post operative haematoma	0 :1721	1 : 31	2 : 229	NA's: 53	
Post operativeoedema	0 :1811	1 : 145	NA's: 78		
Post operative infection	0 :1900	1 : 91	NA's: 43		
Infection antibiotics	0 : 4	1 : 76	2 : 6	NA's:1948	
Tumour quadrant	2 :774 (Other):321	3 :239 NA's:173	1 :228	8 :151	6 :148
Tumour locality	1 :1762	2 : 247	3 : 10	NA's: 15	
Tumour histological grade	1 : 407	2 :1036	3 : 509	4 : 1	NA's: 81
Tumour histological type	1 :1303 NA's: 18	2 : 192	3 : 270	4 : 38	5 : 213
Pathological tumour size (mm)	Min. : 0.00 Max.:128.00	1st Qu.: 9.00 NA's:73	Median : 14.00	Mean : 15.38	3rd Qu.: 20.00
Chemo neo adjuvant	0 :1833	1 : 190	NA's: 11		
Chemo neoadjuvant anthracycline	0 : 24	1 : 170	NA's:1840		
Sys treatment	0 : 406	1 :1617	NA's: 11		
Sys tamoxifen	0 :796	1 :815	NA's:423		
Sys aromatase	0 :806	1 :801	NA's:427		
Radio breast dose	Min. :28.50 Max.:56.00	1st Qu.:40.05	Median :50.00	Mean :45.30	3rd Qu.:50.00
Radio breast fractions	Min. : 5.00 Max.:31.00	1st Qu.:15.00	Median :25.00	Mean :20.46	3rd Qu.:25.00
Radio breast ct volume (cm3)	Min. : 38.0 Max.:5450.0	1st Qu.: 442.0 NA's:17	Median : 711.0	Mean : 808.7	3rd Qu.:1049.0
Radio imrt	0 :1028	1 :1003	NA's: 3		
Radio type imrt	1 : 807	2 : 196	NA's:1031		
Radio axillary levels	0 :1774 5 : 63	1 : 13 NA's: 21	2 : 14	3 : 87	4 : 62
Radio supraclavicular fossa	0 :1758	1 : 261	NA's: 15		
Radio boost	0: 653	1:1381			

Table A.3: **Summury of treatment's data**

B | Appendix B

B.1. Simulation algorithm

The algorithm employed in the simulation studies is developed in [36]. One of the main complexities of running simulation studies is the definition of the generating mechanism for the simulated data. Indeed, to showcase the value of the proposed algorithms, the generated data need to mimic real variomics data and phenotypes of complex traits, whilst presenting a peculiar structure. In particular,

- to represent genotype information, a potentially very large set J of binary variables needs to be generated, with a probability P(j = 1) representing a true distribution of variants in the population, namely a minor allele frequency (MAF);
- the target binary phenotype needs to be either univariate or multivariate according to the specific experimental setting. For immediacy of notation, $\mathbf{Y} \in R^{NxT}$ in bold is the matrix of the multivariate target, its subvectors of size N being Y_k, with $k \in \{1, ...T\}$, while the univariate will be identified by the vector $Y \in R^N$. In the case of the multivariate target, the dimensions Y_k need to present some internal correlation (and/or association) structure;
- genotype data has to include high-order interaction associated with the target phenotype. These are defined as patterns that in practice results in sequences of co-occurring genetic variants. In the case of homogeneous trait, the association of the patterns to the target Y has to be defined differentially w.r.t. the negative class (i.e. the controls population) only. In the case of heterogeneous trait, these patterns need to be associated to $\mathbf{Y}$ w.r.t. the negative class, while also imposing an association between each Y_k and at least one specific pattern not forcefully associated with the others. The simulated dataset needs to present an association for each pattern specifically to one Y_k. The association among $\mathbf{Y}$ dimensions might lead patterns to be associated with multiple Y_ks: this is a desirable result as we wish our algorithm to be able to identify and discriminate between all forms of association, prioritizing for each sub-target the variants' patterns directly associated with them.

In the case of the univariate outcome, the algorithm allows the user to define a number of desired patterns to be included in the generated data, the maximal length l_{max} of each of these patterns, and the association strength, or coverage, of each pattern with the positive class $\alpha \in (0, 1)$. In brief, for a final dataset of size N, with n observations belonging to the positive class, and J variants, the algorithm works as follows: First, it generates the matrix $\mathbf{X}$ of N rows and J columns (binary genetic features). Each entry of this matrix $\mathbf{X}$ is filled with 1 with a probability specified by the user (which should resemble the MAF), 0 otherwise. Then, it generates the binary target variable $Y \in R^N$, including exactly n positive class observations (i.e. Y = 1) and Nn negative class observations (Y = 0). Then, the algorithm generates the desired number of interaction patterns with a probabilistic process. The final patterns will be binary vectors of length equal to the respective l_{max}. Then, for each pattern and each associated l_{max}, the algorithm will randomly pick αn positive class observations and impose the pattern on the picked observations over the first l_{max} randomly populated columns of $\mathbf{X}$ not previously exploited in this process. The aforementioned imposition of the patterns works as the application of the logic condition OR when logically combining two binary vectors.

In the multivariate simulation, the process is similar except for the generation to a multivariate $\mathbf{Y}$. The algorithm allows defining a structure for the $\mathbf{Y}$ in terms of interrelationships among Y_k subgroup vectors. The level of association among the dimensions y_p can be specified for each of the subgroups by imposing a value of relatedness $r \in (0, 1)$, that will result in a joint probability $P(Y_{k=i} = 1, Y_{k=j} = 1) \geq r_c \forall i, j$ in the same structure c. The algorithm imposes such conditions by generating the C structures separately. First, it splits the total number of observations to be generated in C parts: in a dataset that will comprise n positive class observations (or cases), the algorithm computes $n_Y = \frac{n}{C}$. Then, for each structure, the algorithm generates a first binary vector c_1 of size n_Y with $P(c_1 = 1) = 1$. Then, it generates the other binary vector of the same subgroup c_i of the same size, with a probability $P(c_i = 1) = r_c$. The vectors are concatenated column-wise obtaining the structure matrix $\mathbf{M}^{(C)}$ with n_Y rows and a number of columns equal to the number of conditions in each structure. The process is repeated until all C structures have been covered, resulting in C structure matrices $\mathbf{M}^{(C)}$. Then, all the structure matrices are combined placing them on the diagonal of the final target matrix $\mathbf{Y}$, which has n rows and T columns. The remaining entries of the target matrix are filled with 1s with a predefined low probability $p_{filling}$, 0 otherwise to introduce further stochasticity. Once the multivariate target has been defined, the algorithm imposes patterns in a way that is very similar to the univariate case.

Input layer dimension	100
Output layer dimension	100
Number of hidden layer	3
Neurons in the first hidden layer	90
Neurons in the second hidden layer	50
Neurons in the third hidden layer	90
Activation in the first hidden layer	hyperbolic tangent
Activation in the second hidden layer	rectified linear unit
Activation in the third hidden layer	hyperbolic tangent
Activation in the output layer	sigmoid
Optimizer	adam
Loss	MSE in DSAEE and Categorical cross-entropy for denoising DSAEE
Epochs	250
Batch Size	50
λ	0,01
Early Stopping	Not used

Table B.1: **Details in Autoencoder definition in denoising simulation study**. The Table report detail the autoencoder architecture and training process in the denoising simulation case. In the simulation, two autoencoders were trained with the same parameters.

B.2. Details on autoencoder architecture in the simulation settings

This section contains the detail of autoencoder architecture in the simulated setting. Details for each of the simulations are presented in a separate table. Tables report detail about the architecture, such as the dimension of the input layer and the number of neurons in each layer, and details about the training, such as the number of epochs of the value of the parameter related to the L1 loss.

Input layer dimension	100
Output layer dimension	100
Number of hidden layer	3
Neurons in the first hidden layer	90
Neurons in the second hidden layer	50
Neurons in the third hidden layer	90
Activation in the first hidden layer	hyperbolic tangent
Activation in the second hidden layer	rectified linear unit
Activation in the third hidden layer	hyperbolic tangent
Activation in the output layer	sigmoid
Optimizer	adam
Loss	Categorical cross-entropy
Epochs	250
Batch Size	50
λ	0,001
Early Stopping	Not used

Table B.2: **Details in Autoencoder definition in the FS simulation study**. The Table report detail about the autoencoder architecture and training process in the FS simulation case.

Input layer dimension	100
Output layer dimension	100
Number of hidden layer	3
Neurons in the first hidden layer	90
Neurons in the second hidden layer	50
Neurons in the third hidden layer	90
Activation in the first hidden layer	hyperbolic tangent
Activation in the second hidden layer	rectified linear unit
Activation in the third hidden layer	hyperbolic tangent
Activation in the output layer	sigmoid
Optimizer	adam
Loss	Categorical cross-entropy
Epochs	250
Batch Size	50
λ	Outcome dependent, choose between $\{0.005, 0.001, 0.0001\}$
Early Stopping	Not used

Table B.3: **Details in Autoencoder definition in the multi- outcome FS simulation study**. The Table report detail on the autoencoder architecture and training process in the FS simulation case.

C | Appendix C

Appendix C contains details about the case study application. Details about the DSAE architectures are reported in Table C.1 and C.2 while details about the chosen hyperparameters in the hiPRS algorithm for each endpoint can be found in Table C.3. Finally, additional results on the boost group are available in the Tables reported below.

DSAEE details in REQUITE case study

Endpoint	**Input layer dim**	**Output layer dim**	**DSAE hidden layers**	**Activations**	**Optimizer**
y^1	122	122x3	[90,50,90]	[htan, reLu, htan]	Adam
y^2	122	122x3	[100,50,100]	[htan, reLu, htan]	Adam
y^3	122	122x3	[100,50,100]	[htan, reLu, htan]	Adam
y^4	122	122x3	[100,50,100]	[htan, reLu, htan]	Adam
y^5	122	122x3	[100,60,100]	[htan, reLu, htan]	Adam
y^6	122	122x3	[90,50,90]	[htan, reLu, htan]	Adam

Table C.1: **Details in Autoencoder definition in the REQUITE case study**. The Table report detail of the autoencoder architecture and training process.

DSAEE details in REQUITE case study (part II)

Endpoint	loss	Epochs	Batch Size	λ	H_0 test rejection
y^1	Categorical cross-entropy	150	50	0.000001	no
y^2	Categorical cross-entropy	150	50	0.1	yes
y^3	Categorical cross-entropy	150	50	0.05	no
y^4	Categorical cross-entropy	150	50	0.001	yes
y^5	Categorical cross-entropy	150	50	0.0001	yes
y^6	Categorical cross-entropy	150	50	0.000001	yes

Table C.2: **Details in Autoencoder definition in the REQUITE case study**. The Table report detail of the autoencoder architecture and training process. The last column refers to the rejection of the null hypothesis that the RE has the same distribution in the groups by one of the tests described in Table 2.2.

hiPRS hyperparameters definition in REQUITE case study

Endpoint	SNPs considered	K	γ	ES
y^1	5	3	0,1	0,85
y^2	14	30	0,15	1
y^3	4	5	0,1	0,85
y^4	8	20	0,35	0,8
y^5	9	10	0,01	0,8
y^6	18	35	0,3	0,9

Table C.3: **hiPRS hyperparameters definition in REQUITE case study.** For each endpoint the hyperparameters are optimized in cross-validation. γ is the threshold for high-frequency appearing patterns in the FIM routine, K is the number of selected interactions, and ES is the threshold of accuracy at which the training is stopped. The threshold γ ranges in $\{0.01, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$, K in $\{3, 5, 10, 15, 20, 25, 30, 35, 40\}$ and ES in $\{0.8, 0.85, 0.90, 1\}$.

NTCP model performance including clinical covariates on the training set of boost group

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV
y^1	0.741	(0.688, 0.793)	0.713	0.632	0.567	0.778	0.65
y^2	0.762	(0.685, 0.84)	0.217	0.344	0.833	0.568	0.959
y^3	0.753	(0.648, 0.857)	0.168	0.257	0.541	0.790	0.956
y^4	0.775	(0.694, 0.855)	0.227	0.338	0.666	0.752	0.954
y^5	0.813	(0.727, 0.898)	0.192	0.31	0.8	0.738	0.979
y^6	0.855	(0.765, 0.945)	0.141	0.239	0.785	0.79	0.988

Table C.4: **NTCP model performance including clinical covariates on the training set of boost group.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

Prediction model performance including exclusively PRS score on the train set of boost group

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV
y^1	0.611	0.553-0.669	0.58	0.544	0.512	0.644	0.578
y^2	0.881	0.845-0.918	0.370	0.533	0.956	0.752	0.991
y^3	0.850	0.784-0.915	0.158	0.27	0.92	0.622	0.99
y^4	0.816	0.740-0.892	0.298	0.390	0.566	0.87	0.953
y^5	0.698	0.594-0.802	0.224	0.293	0.423	0.881	0.949
y^6	0.924	0.875-0.972	0.17	0.289	0.933	0.796	0.996

Table C.5: **Prediction model performance including exclusively PRS score on the train set of boost group.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

Prediction model performance including exclusively PRS score on the test set of boost group

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	NPV
y^1	0.5724	0.4517-0.693	0.395	0.395	0.395	0.422	0.422
y^2	0.5143	0.3039-0.7246	0.10	0.175	0.5833	0.197	0.75
y^3	0.6982	0.5247-0.8716	0.103	0.187	1	0.3658	1
y^4	0.5014	0.3121-0.6907	0.119	0.21	0.888	0.253	0.952
y^5	0.6433	0.4499-0.8367	0.04918	0.0895	0.5	0.292	0.888
y^6	0.5268	0.2277-0.8259	0.055	0.105	1	0.19	1

Table C.6: **Prediction model performance including exclusively PRS score on the test set of boost group.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2.

NTCP model performance including clinical covariates and PRS score on the training set of boost group

Endpoint	AUC	AUC (CI)	AP	f1	sens	spe	prs p-value	NPV
y^1	0.7648	0.7143-0.8153	0.725	0.6798	0.6397	0.766	0.03391	0.688
y^2	0.9184	0.8819-0.955	0.36	0.526	0.972	0.752	9.22e-08	0.995
y^3	0.9248	0.8773-0.9724	0.302	0.46	0.958	0.826	6.51e-06	0.996
y^4	0.8959	0.8401-0.9516	0.444	0.547	0.714	0.916	2.75e-05	0.971
y^5	0.8631	0.8019-0.9244	0.2079	0.333	0.84	0.75	9.5e-05	0.983
y^6	0.9467	0.8899-1	0.5238	0.62857	0.7857	0.9687	2.87e-05	0.990

Table C.7: **NTCP model performance including clinical covariates and PRS score on the training set of boost group.** For each endpoint a prediction model is computed by logistic regression and evaluated via metrics presented in Table 2.2. The last column contains the p-value of the coefficient test, namely $H_0 : \beta_{PRS} = 0$ vs $H_0 : \beta_{PRS} \neq 0$.

List of Figures

List of Tables

Acknowledgements

This research was made possible thanks to the ERA PerMed Cofund program, grant agreement No. ERAPERMED2018-44, RADprecise— Personalized radiotherapy: incorporating cellular response to irradiation in personalized treatment planning to minimize radiation toxicity. I thank them for giving me the opportunity to participate in this international project that made me grow both as a researcher and as a person. I would like to thank Professor Francesca Ieva who has been my supervisor during this work. I thank her for proposing such an interesting project, for following me and showing me the right path while at the same time leaving me with the freedom to experiment and put my ideas into practice. I also thank her for the time and opportunity she gave me including the possibility to be a visitor at Human Technopole which made my thesis work more productive and gave me the opportunity to get to know an amazing research environment. I thank Dott. Michela Massi who has been my co-supervisor. She was extremely present and caring toward me. I thank her for her helpfulness and welcoming attitude. Michela helped in developing and implementing my ideas by guiding me in the right direction and encouraging me to always search for something new to develop. I thank Dott. Nicola Rares Franco, who supported me in the development of this thesis by always providing encouraging feedback and infusing a lot of positivity into the work. Finally, I thank all my family, my friends, and the people who have been close to me during these years, who helped me grow and discover what I'm passionate about, who gave me lightness in the most complicated moments, and who always pushed me on to do my best.