POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

Executive Summary of the Thesis

Multi-outcome feature selection via anomaly detection autoencoders

An application to radiogenomics in breast cancer patients

Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Author: Alessia Mapelli

Advisor: Prof. Francesca Ieva

Co-advisor: Michela Massi, Nicola Rares Franco

Academic year: 2021-2022

---

1. Introduction

Thanks to treatments such as radiotherapy, the survival in patients diagnosed with cancer is increasing. However, approximately 5% of patients receiving radiotherapy are particularly sensitive to irradiation and likely to develop long-term side effects [2]. The work in this thesis is developed within a large international study, namely, RADprecise [2] aiming at personalizing radiotherapy treatment for cancer patients by improving prediction models for radiosensitivity [2]. Radiosensitivity is a latent outcome, and it is only inferred through measurements of various types of Late Toxicities (LTs), side effects that arise when the radiation damages healthy tissue and can occur years after radiotherapy impairing quality of life. LTs are measured as binary responses or endpoints according to CTCAE (Common Terminology Criteria for Adverse Events) and multivariate inference is necessary for comprehensive treatment decisions. Traditionally, physicians base treatment decisions on model-based risk estimates called Normal Tissue Complication Probabilities (NTCP). Specifically, NTCPs model the risk of radioinduced complications in terms of radiation dose (D), and partial volume irradiated (v). 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. Genetic biomarkers are believed to be crucial in predicting LT development [1]. 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 LTs to the presence of associated genetic mutations in the patient DNA and can be exploited to incorporate genetic information into wider risk prediction models as NTCPs. 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 high-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. In the peculiar setting of genetic studies, proper FS is hindered by several endogenous and exogenous data complexities. Indeed, high-dimensional genetic data is most of the time available in very small samples. Moreover, the study of rare phenotypical traits (such as LT) determines mostly unbalanced settings, with very low case-control ratios that may violate asymptotic assumptions of statistical inference. Additionally, several raw genetic features are not directly measured: indeed, imputation methods estimate genotype probabilities at variants not genotyped to achieve completeness in genetic information. Ignoring the genotypic uncertainty and performing analysis with standard statistical tools generally affect inference, causing statistical significance to be lost for certain experimental configurations. 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 [1]. 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 potential 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 suboptimal when high-order interactions effect is more important than any single genetic variant. Most of the above-mentioned complexities have been recently addressed in [4, 5], where the authors develop a Deep Learning-based FS method for imbalanced data. Of note, the genetic features selected via their Deep Sparse AutoEncoder Ensemble (DSAEE) are subsequently included in an interaction-aware method for polygenic risk scoring (PRSi) [1]. 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 [5] 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 multi-endpoint framework. Techniques to properly handle imputation in the input data are also introduced in the method. 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 [1], the selection can be exploited to efficiently include genetic effects in clinical risk models.

2. Background: Anomaly detection autoencoders

An AutoEncoder (AE) is a neural network trained to copy its input to its output. AEs 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 $\underline{x}$ characterized by J features. The network can be seen as constituted by two parts: an encoder and a decoder. The encoder and decoder functions can be represented as: $\underline{h}_i = f(\mathbf{W}\underline{x}_i + \underline{b})$ in the encoder that maps each input vector $\underline{x}_i$ into an encoded version of itself (i.e. a latent representation), usually in a low dimensional space of size H; and $\overline{\underline{x}}_i = g(\mathbf{W}'\underline{h}_i + \underline{b}')$ in the decoder that maps back the latent representation vector to the original J-dimensional space. In general, we can define an AE as a map $\phi(\underline{x}_i) : R^J \rightarrow R^J$, such that $\phi(\underline{x}_i) = g(\mathbf{W}' f(\mathbf{W}\underline{x}_i + \underline{b}) + \underline{b}')$ and the weight are optimized so that the reconstruction $\overline{\underline{x}}_i = \phi(\underline{x}_i)$ is as close as possible, considering

some loss $L(\underline{x}, \overline{\underline{x}})$, to $\underline{x}_i$. 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, while L is typically a cross-entropy for categorical variables that measure the difference between input and reconstructed values probability distributions. Mimicking the identity function, the AE 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. AEs 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 [5].
Better representations can be achieved using constraints that force autoencoders not only to replicate the input but to learn effective representations of such input in the latent space. In a Deep Sparse AutoEncoder (DSAE), the L_1 penalization is applied on h(l), the function generating the latent representation, forcing the model to represent the input in the simplest way and incrementing generalization propriety of the model.

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

The parameter λ is usually optimized through grid search.
AEs 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. 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 in training. Data points with high loss are considered to be anomalies. 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. A "Denoising" version of autoencoders can be exploited to reconstruct noise-free corrupted versions of their inputs.

3. Methodology

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, LTs in radiogenomics; 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 sample 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 it is possible to simply round each imputation to the closest integer. 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 DSAE to perform multivariate FS. Each AE is trained, as an anomaly detection autoencoder, to optimally represent a class of controls and distinguish them from the anomalies (cases). Multivariate FS is achieved by the proper definition of these groups, accounting for correlation in the endpoints. The method is detailed in the following and schematized in Figure 1.
The FS 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 the population that does not present any toxicity. Genetic variants influencing radiosensitivity are only partly toxicity-specific. Consequently starting from a control population extracted from each endpoint majority class could lead to

the concealing of genetic risk patterns linked to general toxicity, due to their presence in both the majority and minority classes.

$$\mathbf{X_{maj}} = \{\underline{x}_i | Y_{ik} = 0 \ \forall \ k\}$$

For each outcome, k, the specific set of selected features is computed as follows.

(i) The case sample (minority class) is defined, including all the patients that present endpoint k *(B)*.

$$\mathbf{X_{min}} = \{\underline{x}_i | Y_{ik} = 1\}$$

(ii) 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 class patients. The remaining observations of $\mathbf{X_{maj}}$ are included in the training set *(C)*.

(iii) Each DSAE learner is trained on the training set. The training includes a denoising procedure forcing the DSAE to reconstruct from the continuous noisy input its categorical representation so that the possible error due to imputation is accounted for in the comparison *(D)*. 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 following loss:

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

$$+\lambda |h(l)|$$

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

(iv) Once the network has been trained, it is applied to the test set *(E)*. The evaluation of its performance when facing both majority and minority class examples enables the comparison of the RE in two populations to detect features able to distinguish between them.

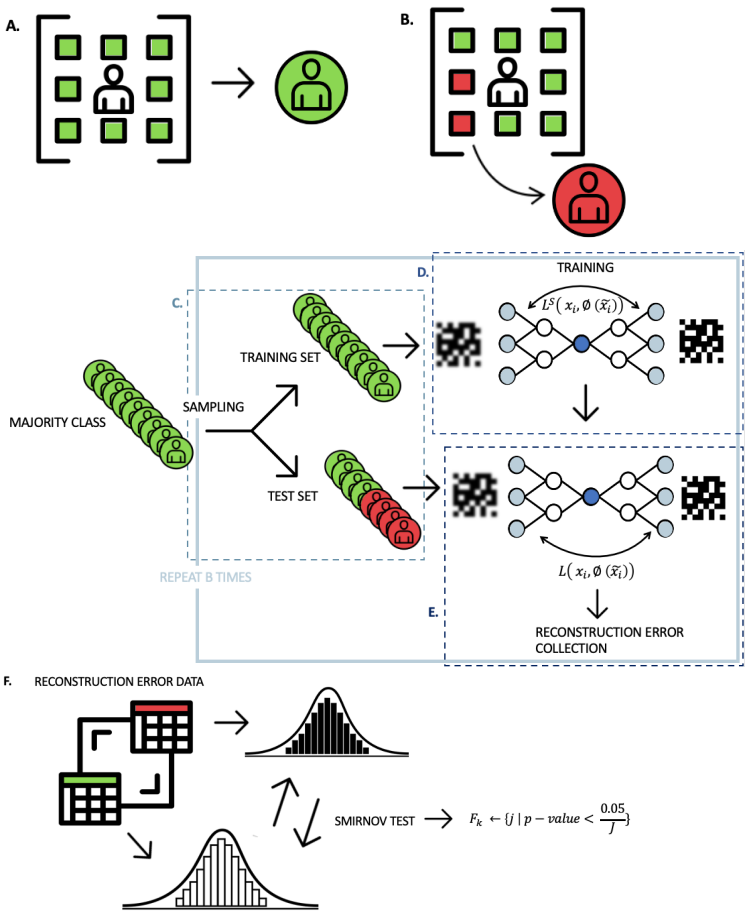

Figure 1: **A methodology illustration.** (**A**) Definition of the control sample as those patients that do not present any toxicity (green). (**B**) Definition of the endpoint-dependent case sample as those patients presenting the specific endpoint (red). (**C**) In each ensemble iteration, a random sample of the same size as the case sample is extracted from the common control class patients and included in the test set while the remaining observations of the latter are included in the training set. (**D**) The DSAEE is trained on the training set. (**E**) Each AE in the ensemble is applied to the test set sample and the reconstruction error is stored in a unique dataset. (**F**) Data are grouped based on their belonging to the case and controls and the distributions of the reconstruction error in the two groups are compared via the Smirnov test for each covariate. Those with a p-value smaller than the Bonferroni correct threshold are selected.

The reconstruction error is evaluated on each sample of the test set as:

$$RE(x_j, \phi(\tilde{x}_j)) = -\sum_{k=0}^{M} (x_{jk} * log(\phi(\tilde{x}_{jk})))$$

$$for \;\; j \in \{1, ..., J\}$$

The test set REs from each ensemble repetition are concatenated in a unique dataset $\mathbf{Q}$ and labeled according to the belonging to the majority or minority class of the samples. At the end of the ensemble procedure, REs are divided into two datasets $\mathbf{Q_{maj}}$ and $\mathbf{Q_{min}}$ including respectively the majority and minority class observations.

(v) Each of the B*O observations in $\mathbf{Q_{min}}$ and $\mathbf{Q_{maj}}$ is considered as a sample extracted from each feature distribution $re_j|$ minority sample $\sim f_j^{min}$ and $re_j|$ majority sample $\sim f_j^{maj}$ with $j \in \{1, ..., J\}$. It is possible then to compare the samples for each feature and test if the distributions in different groups are statistically different *(F)*. 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. This selection method will be referred to in the following as distribution-based selection.

(vi) If the F_k includes an oversized number of features. A second selection method can be applied to F_k to restrict it. We can estimate the vector of average RE per feature per class: l_{min} and l_{maj}, both belonging to R^J, where each element is computed as:

$$l_{min,j} = \frac{1}{B * O} \sum_{t=0}^{B*O} Q_{min,t}$$

$$l_{maj,j} = \frac{1}{B * O} \sum_{t=0}^{B*O} Q_{maj,t}$$

It is possible to compute for each $j \in F_k$ Δ_j, the difference between the average RE on the minority class and the majority class.

$$\Delta_j = l_{min,j} - l_{maj,j}$$

The features can then be ranked in decreasing order according to Δ_j. To identify an exact set F_k^{comb} we need to define a threshold $\delta \in (0, 1)$, such that Δ_δ is the δ-th quantile evaluated on the distribution of $\{\Delta_j\}_{j \in F_k}$. We, therefore, select all those features j whose average RE difference is above the predefined threshold:

$$F_k^{comb} = \{j \in F_k | \Delta_j \geq \Delta_\delta\}$$

This second selection methodology will be referred to in the following as combined selection.

Once the set of features is selected for each endpoint, it is possible to compute the union of all the features selected and consider it as the set of significant features for a comprehensive endpoint.

The multivariate FS developed 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. The method is robust to error in the imputation of genomic data thanks to a procedure inspired by denoising autoencoders and to the imbalance in the classes thanks to a procedure inspired by anomaly detection autoencoders. Finally, the FS is high-order interaction aware thanks to autoencoders' intrinsic hierarchical structure, and the selection method allows for the interpretability of the selection.

4. Simulation study

4.1. Simulation setting

In this section, we validate the distinctive aspects introduced in the proposed method 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.

4.2. Denoising proprieties of the developed method

We first aim at validating the utility of the denoising aspect introduced in the novel DSAEE algorithm. That is, we want 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 exponential noise. To test the denoising capability of the method a denoising DSAEE, reconstructing the continuous input to its categorical representation, is compared to a DSAEE algorithm reconstructing a continuous output. 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 AE architecture is implemented and its in-training convergence is analyzed. 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. To quantify and evaluate the reconstruction ability of the autoencoder, the loss, MSE in the DSAEE and binary-cross-entropy in the denoising DSAEE, and accuracy are considered and their evolution is studied during the training process. In Figure 2 in-train loss of both methods is plotted to be compared and in Figure 3 the accuracy is reported. 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 overall. The in-train AUC is shown in Figure 4.

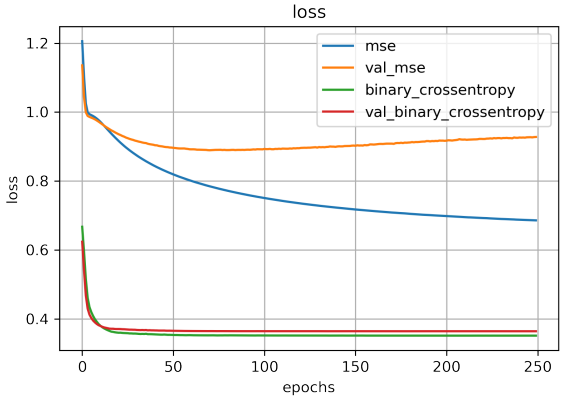

Figure 2: **In-train loss of the compared methodologies.** In the DSAEE the loss is evaluated via MSE since noisy data are treated like continuous variables, while in denoising DSAEE the loss is evaluated via cross-entropy since the encoder output distribution is compared to the real categorical value. 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.

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. Finally, the AUC performance of the denoising algorithm stabilized at approximately 62% in both the training and validation sets. 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.

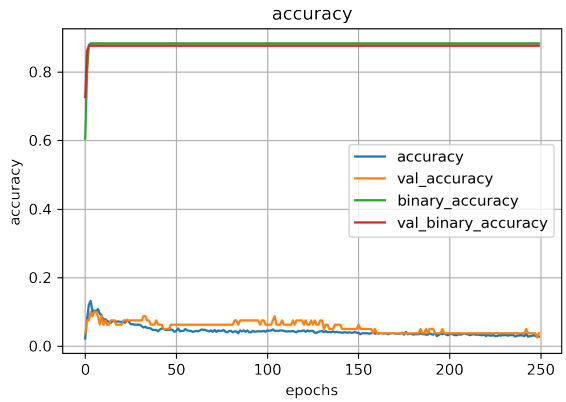

Figure 3: **In-train accuracy of the compared methodologies.** The training was performed excluding a validation set to mimic the performance of the autoencoder on unseen data. The accuracy of the validation sets is also presented.

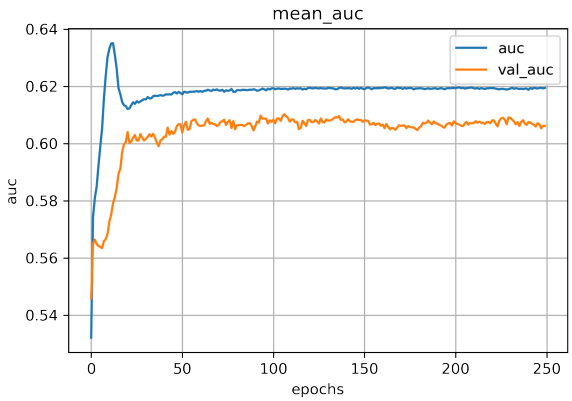

Figure 4: **In-train AUC** of the denoising DSAEE. The training was performed excluding a validation set to mimic the performance of the autoencoder on unseen data. The mean AUC on the validation sets is also presented.

In conclusion, the simulated analysis reveals the denoising DSAEE's 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.

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

This section aims at verifying the capability of the multi-outcome methodology to improve feature selection. The method's performance in FS tasks is evaluated by considering:

- $FSPrecision = |IRF|/|F|$
  where $|IRF|$ is the total number of informative and redundant features selected and $|F|$ is the total number of features selected
- $FSCorrespondence = |IF|/|TIF|$
  where$|IF|$ is total number of informative features selected and $|TIF|$ is total number of true informative features

This simulated experiment includes a dataset with 1000 observations and 100 variants 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%. To test the improvement in the multi-outcome selection the multivariate algorithm is tested against the algorithm described in [5] which performs univariate selection for each outcome separately. 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. 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 com- posed of 10 learners. The parameter regulating the sparsity term of the loss (i.e.λ) is instead optimized by grid search during the training of each DSAEE. Features are selected via both the distribution-based and the combined selection methods. The δ parameter of the combined selection is optimized in each method first on FSP, as the fundamental goal is

Multivariate method performance on the simulated dataset

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

Table 1: **Multi-outcome method performance on simulation dataset.** For each of the considered outcomes, metrics evaluation of the selection via Distributional 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.

Univariate method performance on the simulated dataset

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

Table 2: **Univariate method performance on simulation dataset.** For each of the considered outcomes, metrics evaluation of the selection via Distributional 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.

to avoid introducing non-informative features in the selection, and secondly on FSC to get the best possible set of features. An additional endpoint is defined. The "at least one" (alo) endpoint is defined for each sample as

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

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. 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. Tables 1 and 2 report the results for the multivariate and the univariate FS algorithms respectively.
The first result emerging from the analysis concerns the comparison between the two selection methods considered. In the simulation setting the combined FS works better since the distributional-based FS defines a set of covariates too large and therefore includes insignificant covariates. A possible reason for this behavior is the clean separation of the classes in the simulation setting. In applications, usually, groups are overlapping and the set of features statistically able to distinguish between them is restricted. 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 metric increase, implying a selection focused on the discovery of every and only significant features that better avoid those linked to correlated endpoints. Through the results presented in this section, it is possible to verify another propriety of the multivariate selection. 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 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 determinant 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.

5. Case study application in radiogenomics

As the original DSAEE algorithm, the multivariate DSAEE presented in this work was tailored to tackle the complexities of real-life genomic research. 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, we briefly present the case study application (detailed in [3]) of the proposed algorithm on the RADPrecise Breast Cancer Cohort, with the aim of identifying variants associated with LT. The considered sample is a subset of 599 patients with a documented follow-up visit three years after the initial cancer treatment. Six late toxicity endpoints are considered: five have an incidence below 10%, while one occurs for approximately 47% of the subjects. The pool of genetic features to select from included 122 variants previously identified in the literature as correlated to radio-induced LTs 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) LT endpoints. This resulted in six dimension-specific sets of selected SNP exploited (i) independently to build six different PRSs (one for each LT endpoint) and (ii) in combination (i.e. their union), to define a unique PRS to predict the overall risk of any LT. The PRSs are computed following the PRSi algorithm presented in [1] 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. This being an unsupervised setting it is hard to comment on the precision of the results without the required clinical expertise. However, the metrics of the classification models exploited in the definition of the PRSs reveal good predictive models (i.e AUC $\geq 60\%$), with some limitations in the discrimination of the minority class. Good results are probably induced by an effective initial selection.

6. Discussion and Conclusions

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 [5], 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. 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, features' imputation (i.e. noise), 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 is applied to the RADPrecise Breast Cancer Cohort, with the aim of identifying variants associated with LT and constructing a PRS for breast cancer LTs. 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 multivariate model succeeds in the accurate selection of highly influential features that provide intrinsic information and discriminant properties for class separability. The accurate definition of influential features 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 LTs. 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 a ground truth definition of noisy input data and the difficulty to scale in input features due to the high computational cost. Further development can be introduced in the model. Variational autoencoders have already been proposed for anomaly detection. 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. 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. 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.

7. 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.

References

[1] 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.

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

[3] A. Mapelli. Multi-outcome feature selection via anomaly detection autoencoders: An application to radiogenomics in breast cancer patients. Politecnico di Milano, Master's Thesis, 2022.

[4] 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.

[5] 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.