Deep Sparse Autoencoder-based Feature Selection for SNPs Validation in Prostate Cancer Radiogenomics

Selezione Variabili tramite Deep Sparse Autoencoders per validazione di SNPs in Radiogenomica del Cancro alla Prostata

Michela Carlotta Massi[1,2], Francesca Ieva[1,2,3], Anna Maria Paganoni[1,2,3], Andrea Manzoni[1], Paolo Zunino[1], Nicola Rares Franco[1], Tiziana Rancati[4], and Catharine West[5,6]

Abstract Prostate cancer is the most diffused cancer affecting the male population. As therapies improve their effectiveness, surviving patients might be affected by complications induced by radiotherapy in the long run. To predict the onset of such rare late toxicities, because of the failure of phenotypic characteristics, the attention is shifting towards identifying specific genetic locations (Single Nucleotide Polimorphisms, or SNPs) associated with them. Because of the complexity of the problem, SNPs identified in a study are rarely validated on a different cohort of patients. In this case study we apply a novel approach for feature selection (namely a Deep Sparse Autoencoder-based Feature Selection method), to validate SNPs associated with radiotherapy-induced late toxicity causing urinary frequency variation (UFV).

Abstract *Il cancro alla prostata è il più diffuso tra la popolazione maschile. Nonostante il miglioramento nei trattamenti, i pazienti comunque essere affetti da complicazioni indotte dalla radioterapia nel lungo periodo. Per predire l'emergere di queste rare tossicità tardive, visto il fallimento nell'utilizzare caratteristiche fenotipiche dei pazienti, l'attenzione si sta spostando sull'identificare loci genetici (SNPs) a loro associate. Per la complessità del problema, le SNP individuate in uno studio sono raramente validate su una coorte differente di pazienti. In questo caso studio applichiamo un nuovo metodo di selezione delle variabili (un metodo basato su Deep Sparse Autoencoders), per validare le SNPs associate con la variazione tardiva della frequenza urinaria.*

Key words: Radiogenomics, Feature Selection, Autoencoders, Deep Learning, SNPs

[1] MOX Laboratory, Math Department, Politecnico di Milano, Milan, Italy
[2] CADS-Center for Analysis, Decisions and Society, Human Technopole, Milan, Italy
[3] CHRP-National Center for Healthcare Research and Pharmacoepidemiology, University of Milano-Bicocca, Milan, Italy
[4] Prostate Cancer Program, Fondazione IRCCS Istituto Nazionale dei Tumori, Milan, Italy
[5] Translational Radiobiology Group, Division of Cancer Sciences, University of Manchester
[6] Manchester Academic Health Science Centre, Christie Hospital, UK

1 Introduction

Prostate cancer is the most diffused cancer affecting the male population in Europe. According to the American Cancer Society, about 1 american man in 9 will be diagnosed with prostate cancer during his lifetime. Because of the recent advancements in treatments, survival rates are high, but patients might suffer from debilitating complications resulting from therapies in the long run (radio-therapy induced late toxicity) [1, 2].

Traditional methods based on patients' phenotypic characteristics and treatment details fail in stratifying the treated population and in predicting the onset of such negative, but still very rare, side-effects. For this reason, the attention is shifting towards investigating possible relations between the genotype and the adverse outcomes in the so called 'precision medicine' approach, driving the need for novel statistical methods to address this question.

This case study was conducted with the support of Fondazione IRCSS Istituto Nazionale dei Tumori, the Italian National Cancer Research Institute, that provided us with data regarding the REQUITE [3] cohort of prostate cancer patients, with the aim of validating some specific genetic markers (in the form of Single Nucleotide Polymorphisms, SNPs) that could be predictive for late toxicity. The identification and validation of predictive biomarkers is an objective of paramount importance in a setting such as Genome-Wide Association Studies (GWAS), as the complexity of the problem, the rarity of the *traits* (or negative outcome) and the numerosity of the genetic traits to evaluate makes it extremely complex and rare for different studies to recognize similar patterns in data.

In this short paper we present a novel approach to SNPs validation, exploiting a Deep Sparse Autoencoder-based (DSAE) feature selection method to identify relevant SNPs associated with radiotherapy-induced late Urinary Frequency Variation (UFV). The task at hand requires us to identify predictive features for an extremely small minority class in a setting characterized by complex non-linear interactions among genetic loci, small sample size, several confounding factors, noisy data and the need for results interpretability to drive real clinical research.

For this reason, the work presented in this study exploits a feature selection method tailored to identify relevant features to discriminate the minority class from the majority class, and improve minority class classification accuracy. The adopted methodology for this case study was developed in a previous work in [4], where a detailed description of the algorithm can be found. For this reason, in Section 2 we will provide only a brief description of the main concepts, while the rest of the paper will be devoted to the case study.

The benefits of this Deep Learning (DL) model for our objective are several: it is a non-linear and stratified model, allowing to learn complex and hyerarchical relationships in data; additionally, the model deals well with large numbers of features, and has the capability of autonomously ignore noise.

2 Methods

Deep Sparse AutoEncoders (DSAE). An AutoEncoder (AE) is a neural network whose output provides a reconstruction of the input (Hinton and Salakhutdinov, 2006). The network can be seen as constituted by two parts: an encoder and a decoder.
The encoder function $\mathbf{h}_i = f(\mathbf{W}\mathbf{x}_i + b)$, encodes each input vector $\mathbf{x}_i$ into an encoded version of itself of size H. Here $f : \mathrm{R}^\mathrm{J} \rightarrow \mathrm{I\!R}^\mathrm{H}$ is usually non-linear, $\mathbf{W}_{H\times J}$ is called *weight matrix* and $\mathbf{b}$ is an H-dimensional *bias* vector.
The decoder maps back the encoded vector to the J-dimensional space in most cases using a squashing non-linear function $\hat{\mathbf{x}}_i = g(\mathbf{W}'\mathbf{h_i} + \mathbf{b}')$, $g : \mathrm{I\!R}^\mathrm{H} \rightarrow \mathrm{I\!R}^\mathrm{J}$ with parameters $\mathbf{W}'$ and $\mathbf{b}'$. The model is trained through gradient descent of the loss function $L(\mathbf{x}, \hat{\mathbf{x}})$; where L is typically the Mean Squared Reconstruction Error (RE), i.e. the mean squared Euclidean distance between the input values and the reconstructed values for each observation.
To force the model to learn more useful representaitons of the input data, one approach is to force sparsity in the central hidden layer. A sparse representation can be obtained adding a penalty term that penalizes the L_1 norm of the vector $\mathbf{h}_i^{(l)}$ of activation of the hidden nodes (where (l) indicates the layer the hidden nodes belong to, and it should be considered the most internal layer in case of a Deep AE), for each observation i, controlled by the parameter λ, i.e.:

$$L_i = L(\mathbf{x}_i, \hat{\mathbf{x}}_i) + \lambda|\mathbf{h}_i^{(l)}|. \tag{1}$$

The parameter λ can be optimized through grid search or can be arbitrarily chosen in the design phase of the model.

DSAE for Minority Class Feature Selection. The main idea behind the choice of a DSAE as a mean to perform feature selection is that the model trained to reconstruct *normal* observations only (majority class, or *healthy* patients) would make higher errors by trying to reconstruct anomalous patterns in *outliers* (minority class, or *unhealthy* patients showing late toxicity). Indeed, it is on the analysis of the average RE performed by the model on each feature for each class that we identify those that could discriminate better between the two classes. A detailed description of the proposed methodology can be found in our previous work [4]. In Figure 1 we propose a schema of the algorithm after the trained DSAE is tested on both healthy and unhealthy patients. Note that as a result the algorithm provides a subset of features which dimension depends on a parameter ($\delta \in [0,1]$) set by the user: the closer the δ value is to 1, the smaller the subset.

3 Urinary Frequency: Case Study Setting

From the original dataset, we selected a cohort of 1,296 patient, among which 55 (4.2%) belonged to the class of *cases* (y=1, i.e. the patients reported radiotherapy-induced late UVF), while 1241 (95.8%) belonged to the *controls*'s class. Each pa-

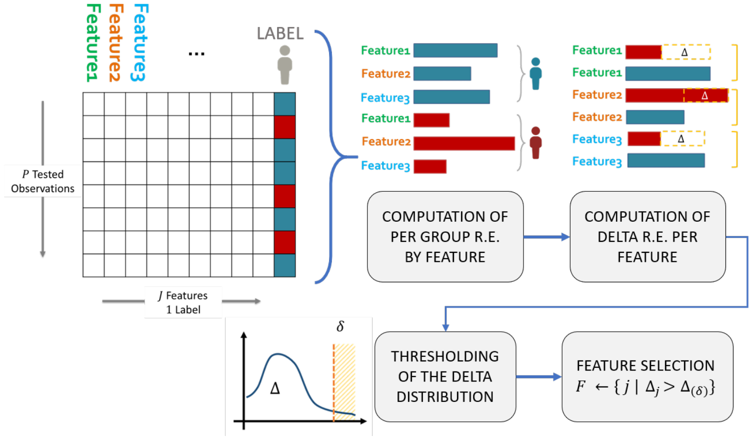

Fig. 1 Schema of the proposed algorithm: after training the DSAE on healthy patients only, the model is supplied with a test set composed of healthy and unhealthy patients. The passages in this schema depict all the steps from the collection of the RE to the feature selection based on δ. More details can be found in [4].

tient was characterized by 43 genetic traits (SNPs), among which 9 were identified in previous studies as predictive biomarkers for this endpoint. In Table 1 we list the biomarkers to validate. The unbalancing of the classes and the complexity of the problem (there potentially exist complex non-linear relations among biomarkers affecting the outcome [6]) makes this field of application an interesting fit for the peculiarities and potentials of our proposed model [4]. We performed the training and testing of the DSAE 50 times, extracting from the 1,296 the training set (1,186 observations, i.e. all *controls* except the 55 included in the test set) and test set (110 observations, half *cases* and half *controls*). The algorithm was implemented in Python. The DSAE had one input layer with 43 nodes, and three encoding hidden layers (with 40, 30 and 20 nodes respectively), followed by three decoding layers (30, 40, 43 nodes respectively). The training of each DSAE was performed for 400 hundred epochs, with a batch size of 10 observations, and the whole procedure of sampling, training and testing took on average (over the 50 repetitions) 3.22 minutes to complete. Note that the training time of the whole algorithm highly depends on the number of repetitions of sampling and training, and could be highly reduced in case less repetitions are needed to capture the most relevant variations between the two classes, or the number of minority class observations is sufficiently large to require a smaller number of sampling procedures on the healthy

SNP	Reference
rs17599026	Kerns et al. (2016) [5]
rs342442	Kerns et al. (2016) [5]
rs8098701	Kerns et al. (2016) [5]
rs7366282	Kerns et al. (2016) [5]
rs10209697	Kerns et al. (2016) [5]
rs4997823	Kerns et al. (2016) [5]
rs7356945	Kerns et al. (2016) [5]
rs6003982	Kerns et al. (2016) [5]
rs10101158	Kerns et al. (2016) [5]

Table 1 SNPs previously identified in literature as associated with late UVF

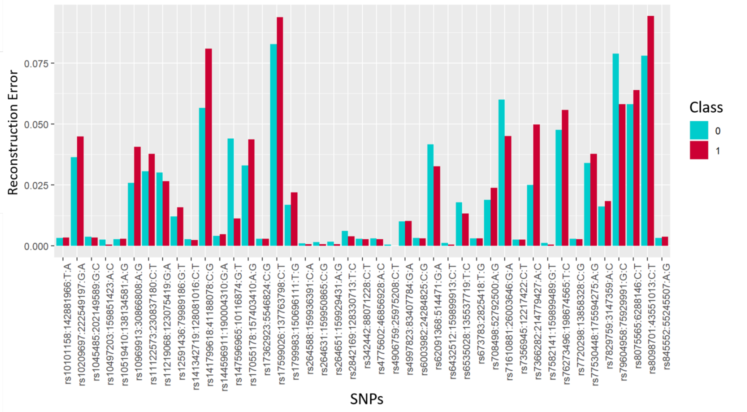

Fig. 2 Reconstruction Error by SNP and by Group (cases in red and controls in blue)

population to guarantee a robust comparison.
Once the AE was trained to reconstruct the training set of the healthy population, the test set (composed of healthy and unhealthy observations) was supplied to the model, collecting the Reconstruction Error (RE).
As in the process depicted in Figure 1, we obtained a matrix where each patient (row) that belonged to the test set at least once was associated with an outcome and a set of REs for each feature (SNP). This allowed us to group patients w.r.t. the enpoint, and estimate the average RE for each SNP for *cases* and *controls*.

4 Results

In Figure 2 we report the results of the procedure just described. The bars in the barplot represent the reconstruction error for each group (cases in red, controls in blue). On the x-axis one can read all the 43 SNPS, each one with two associated bars. To validate the SNPs in a robust way, we selected different values for the δ threshold (δ equal to 0.75, 0.8, 0.85 and 0.9). In Table 2 are reported the 9 SNPs previously identified in literature (already mentioned in Table 1) as predictive for the onset of late UFV after radiotherapy. As shown in Table 2, for δ=0.75 the model identifies as relevant (thus validating) 4 out of 9 SNPs coming from literature. Interestingly, the four identified SNPs present the highest odds ratio w.r.t. the outcome, according to the study that first mentioned them. Note that the study in [5] was performed on a different cohort of patients. The fact that the identified SNPs are those most evidently related to the outcome on different data is both a proof of our methodology to identify the most discriminative features, and of the generalizability of its results. Unfortunately, we do not have access to the data from the mentioned study to cross-validate our method on that cohort. Another notable aspect of our results, is that the four identified SNPs remain relevant almost for all δ values, except for one that is excluded after the last threshold (0.9).

ODDS RATIO	**THRESHOLD**			
	0.75	0.8	0.85	0.9
3,2	**rs7366282**	**rs7366282**	**rs7366282**	**rs7366282**
3,12	**rs17599026**	**rs17599026**	**rs17599026**	**rs17599026**
2,66	**rs10209697**	**rs10209697**	**rs10209697**	rs10209697
2,41	**rs8098701**	**rs8098701**	**rs8098701**	**rs8098701**
1,8	rs10101158	rs10101158	rs10101158	rs10101158
1,74	rs7356945	rs7356945	rs7356945	rs7356945
0,51	rs342442	rs342442	rs342442	rs342442
0,51	rs6003982	rs6003982	rs6003982	rs6003982
0,49	rs4997823	rs4997823	rs4997823	rs4997823
TOTAL SNPS	43	43	43	43
TOTAL SELECTED	11	9	7	5
TOTAL IDENTIFIED	4	4	4	3
PERCENTAGE IDENT/SEL	36.36%	44.44%	57.14%	60.00%
PERCENTAGE SEL/TOT	25.58%	20.93%	16.28%	11.63%

Table 2 Results of the SNPs validation for UFV. SNPs validated by our methodology are in bold, for different threshold values. The first column reports the ORs associated with these SNPs in the study in [5].

5 Conclusion

In this paper we presented a novel approach to SNPs validation through the use of a DSAE-based feature selection method to select relevant minority class features. We applied the methodology to a case study that required us to validate SNPs previously identified in literature as predictive for the onset of radiotherapy-induced late UFV. Despite the complex unsupervised setting does not allow us to compare our results with a *ground truth*, the robustness of the identified SNPs and the height of the Odds Ratio associated to them on another cohort of patients support our results.

Using a DL approach in a GWAS seems therefore to be a viable strategy to tackle the peculiar complexities of this setting, and opens the venue for relevant future research.

References

1. M. J. Zelefsky, A. Pinitpatcharalert, *et al.*, “Early tolerance and tumor control outcomes with high-dose ultrahypofractionated radiation therapy for prostate cancer,” *European urology oncology*, 2019.
2. T. Rancati and C. Fiorino, “Predicting toxicity in external radiotherapy: A critical summary,” in *Modelling Radiotherapy Side Effects*, pp. 337–363, CRC Press, 2019.
3. P. Seibold, A. Webb, *et al.*, “Requite: A prospective multicentre cohort study of patients undergoing radiotherapy for breast, lung or prostate cancer,” *Radiotherapy and Oncology*, vol. 138, pp. 59–67, 2019.
4. M. C. Massi, F. Ieva, F. Gasperoni, and A. M. Paganoni, “Minority class feature selection through semi-supervised deep sparse autoencoders,” *MOX Report 38/2019*, 2019.
5. S. L. Kerns, L. Dorling, L. Fachal, *et al.*, “Meta-analysis of genome wide association studies identifies genetic markers of late toxicity following radiotherapy for prostate cancer,” *EBioMedicine*, vol. 10, pp. 150–163, 2016.
6. B. Liu, Y. Wei, Y. Zhang, and Q. Yang, “Deep neural networks for high dimension, low sample size data.,” in *IJCAI*, pp. 2287–2293, 2017.