sciendo

CAIM
ISSN 2038-0909

Commun. Appl. Ind. Math. 13 (1), 2022, 29–48 DOI: 10.2478/caim-2022-0003

Research

Double-stage discretization approaches for biomarker-based bladder cancer survival modeling

Mauro Nascimben[1*,2], Manolo Venturin[1], Lia Rimondini[2]

[1]Enginsoft SpA, Padova, Italy

[2]University of Eastern Piedmont Amedeo Avogadro, Dept. of Health Sciences, Novara, Italy

[*]Email address for correspondence: m.nascimben@enginsoft.com

Communicated by Giorgio Fotia

Received on 02 02, 2021. Accepted on 07 06, 2021.

Abstract

Bioinformatic techniques targeting gene expression data require specific analysis pipelines with the aim of studying properties, adaptation, and disease outcomes in a sample population. Present investigation compared together results of four numerical experiments modeling survival rates from bladder cancer genetic profiles. Research showed that a sequence of two discretization phases produced remarkable results compared to a classic approach employing one discretization of gene expression data. Analysis involving two discretization phases consisted of a primary discretizer followed by refinement or pre-binning input values before the main discretization scheme. Among all tests, the best model encloses a sequence of data transformation to compensate skewness, data discretization phase with class-attribute interdependence maximization algorithm, and final classification by voting feature intervals, a classifier that also provides discrete interval optimization.

Keywords: genetic expression, bladder cancer, discretization, survival rate modeling, data-driven biomarker research, machine learning

AMS subject classification: 92B15, 68T01, 92C50

1. Introduction

Alteration of genes regulating cell growth and differentiation causes cancer, a disease characterized by uncontrolled cell proliferation. One of the key principles of precision medicine applied to oncology is simultaneously profiling gene expression data (GED) from multiple sources to define a personalized model to contextualize patients' clinical outcome [1]. Identifying a subset of genes differentially expressed between conditions (i.e., healthy vs sick) provides the foundation of gene expression profiling. However, due to the amount of complex and heterogeneous bio-molecular data coming from the laboratories, it could be essential to reduce the genetic expression data-set to the most relevant genes underlying the typology of a specific disease. Gene regulatory networks establish complex relations between molecular regulators and other substances to control the expression levels of hundreds to thousands of genes. Inside these networks, identifying hub genes, which are highly correlated and interconnected with others, could be crucial because high connectivity implies a rapid transfer of information in the gene network [2]. Even small changes in hub genes could impact the major part of the network, thus being "markers" of aberrant behaviors in cellular swelling.

In the present work, we evaluated the predictive potential of bladder cancer hub genes identified by a medical team at Leipzig University (Germany) after an extensive literature review that summarized information collected from six public data-sets [3]. Identifying diagnostic and prognostic biomarkers for cancer screening is an essential step for the future development of a safe, non-risky, and preventive cancer treatment alternative to cystoscopy and bioptic histology [4]. In addition, the investigation of biomarkers based on GED offers valuable insights into a disease's course with the possible development of personalized treatment. In this study, GED was modeled using statistical learning methods to find a data-driven predictive procedure. Before building a statistical model, researchers on GED apply few steps as pre-processing, including data transformation [5] and data discretization [6]. We mainly evaluated the possibility of optimizing the discretization operation to achieve more accurate modeling of GED, comparing four experimental analysis pipelines by their probability of correctly ranking tumor outcome. Additionally, we explored the trade-off between accuracy and interpretability, the latter being a critical factor in medical decision making or health care policy because it could provide insights about biological processes related to cancer onset and progression, supporting the development of more effective therapies. For this reason, preference was given to simpler models able to explain the phenomenon following the principle of parsimony of explanations [7].

2. Materials and Methods

Data was organized by [8] and released under creative commons license. Authors of the data-set provided $N = 406$ anonymized clinical samples containing gene expression values of 14 hub genes related to bladder cancer. Detection of the hub genes was carried out with DAVID (Database for Annotation, Visualization and Integrated Discovery). The authors also added 11 seed genes from the most important modules outputted by the FUNRICH (Functional Enrichment analysis tool) software. In the present investigation, custom scripts in Python and R programming languages were used as data analytics tools.

2.1. *Pre-processing*

The number of observations was reduced to $N = 405$ because one subject had multiple missing entries. In addition, a patient had one missing genetic expression value that was replaced by iterative imputation [9] using ten iterations and the mean value of the gene as initial guess. Each patient was labeled according to the disease exitus for binary classification purposes (survival rate 55%).

2.2. *Original data assessment*

Initially, Spearman rank-order correlations values were grouped by hierarchical clustering (Figure 1) to highlight relations between GED. The dendrogram shows two distinct assemblies of GED that correlate to each other, resembling the division in seed and hub genes. Nonetheless, these groups do not match perfectly the hub and seed sets, suggesting that genes among these two groups do not contribute equally to the informative content of the data-set. In the original paper [8], Dr. Zhang described an opposite behavior of CRYAB, TPM1, and CASQ2 genes compared to other hub genes. The negative correlation between these three genes and the other hub genes appears on the dendrogram, with their inclusion in the seed group. Moreover, gene expression data was pre-selected by medical doctors that authored the original data-set [8], and feature selection techniques based on variance or correlation may not consider all the intuitions underlying their research. To preserve all the knowledge present in the data-set and reduce the feature space removing redundant information, dimensionality reduction was usually preferred to feature selection because it creates new synthetic features by combining the original ones. Nevertheless, feature selection was still compared to dimensionality reduction when exploring innovative approaches as in Sections 3.2 and 3.3.

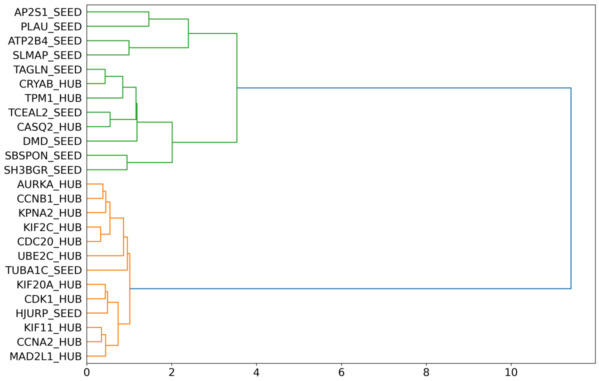
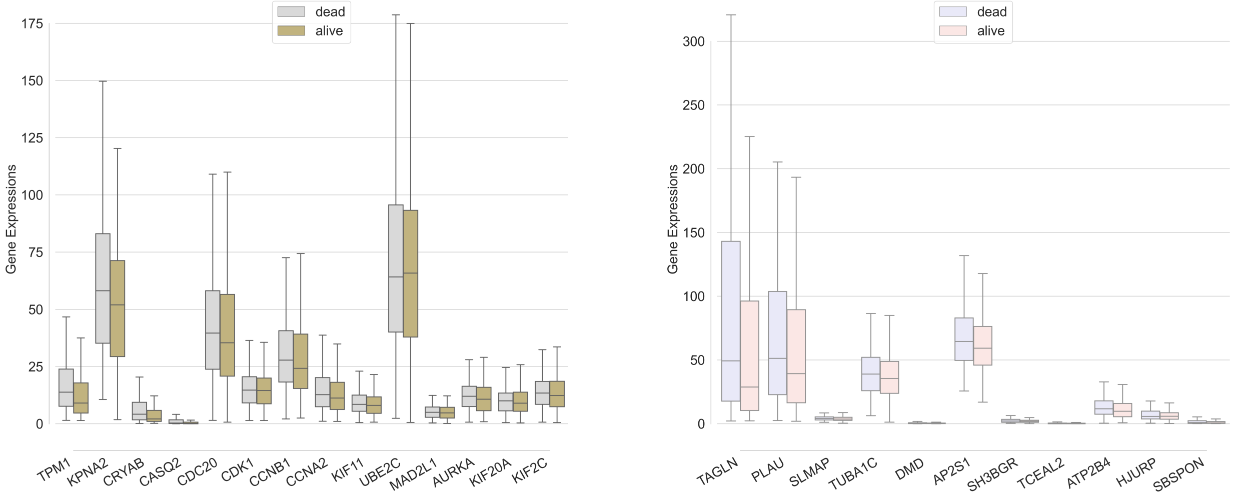


Figure 1. Hierarchical clustering on the Spearman rank-order correlations

In Figure 2, probability distributions of the raw data from the hub and seed groups are shown divided by classes of clinical outcomes. In most genes, raw data distribution can be described by a shape heavily skewed towards the right tail. During oncogenesis, the accumulation of mutations in several proto-oncogenes, and malignant tumor events associated with cancer progression, could determine a skewed GED distribution, as also seen in previous works [10,11]. Skewed data may negatively affect training of machine learning algorithms like gaussian naïve bayes or neural networks [12], or impair the interpretation of feature importance. Moreover, the data-set is slightly imbalanced towards one class (survivors), introducing another source of disturbance that could potentially prejudice classifier performance. Strategies involving data transformation were integrated into the proposed analysis pipelines together with the weighting of class instances during classification to overcome data and labels skewness. In Figure 3, skewness and kurtosis values were divided by their standard error and plotted together with dashed gray lines depicting how many standard errors the sample excess kurtosis or skewness deviates from zero, assuming a normal univariate distribution [13]. This procedure resembles a two-tailed test that approximates the 0.05 significance level with a threshold of ≈ 1.96. Methods for skewness and kurtosis threshold estimation were summed up in Supplementary Materials (SM) Section 2. Scatterplot values are localized above the horizontal threshold, suggesting a positive skewness in all GED. Few genes, those on the left of the vertical gray line, do not show positive kurtosis.



(a) Hub genes

(b) Seed genes

Figure 2. The initial probability distribution of all genes in the two groups

2.3. *Data transformation*

Each gene expression was log-transformed to correct the right skewness and achieve a more symmetrical distribution [14]. We also tested the possibility of using non-linear rank transformations during this study so the data could be mapped to a uniform or normal distribution instead of applying logarithmic conversion (Section 3.4). Uniform distribution reduces the impact of potential outliers and spreads out the most frequent values, re-distributing feature informative content in an attempt to balance heterogeneous ranges of values as seen in the original GED (Figure 2). This procedure involves the estimation of the cumulative distribution for each genetic expression. After this step, the quantile function maps the cumulative distribution to the uniform or normal one. An advantage of uniform distributions is that they can represent both discrete and continuous data. In the present study, transformations were applied to

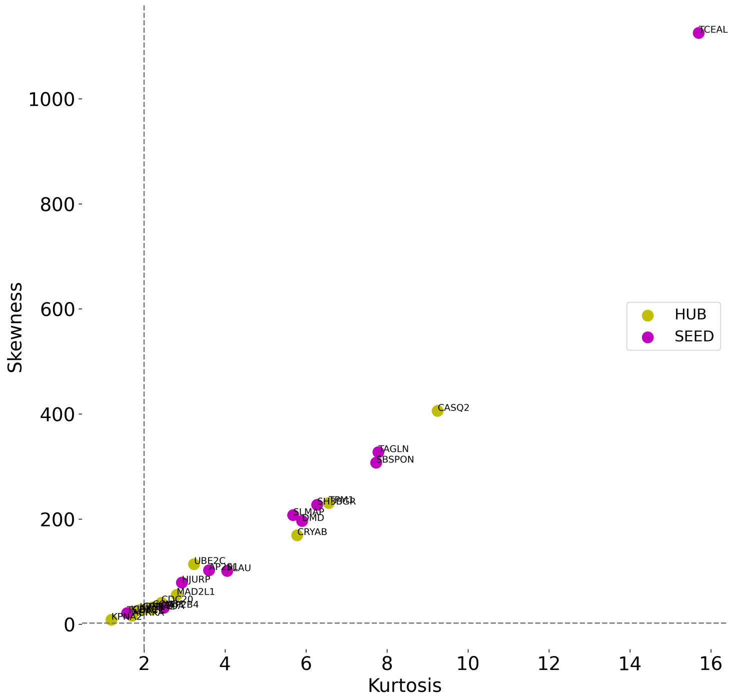


Figure 3. Visual representation of skewness and kurtosis of each gene

each gene data across subjects. A common practice in medicine (including gene expression data) involves normalizing data with z-scores [15] to facilitate comparisons between individuals. However, there could be some concerns on the application of z-scores to markedly skewed data before any transformation, and for this reason, classic approaches handle skewness log-transforming the data before applying z-scores. In Table 1, each gene information is collected after logarithmic and z-score transformations of the original values; normality was tested using D'Agostino and Pearson's departure from normality test [16] while outliers were detected in terms of distance from the median absolute deviation [17] using consistency constant of 1.4826 and Harrell-Davis quantile estimation [18].

Transformed values from all patients are shown in Figure 4 arranged over a circular heat-map [19] together with annotations: three inner circles represented gender (dark brown for females and yellow for males), age (subdivided in three groups: (0-35y) "young" in blue, (36-65y) "adults" in green and (66-99y) "elderly" in red) and human sub-populations (white sub-population in gray, black and afro-americans in pink and asians in blue). An important insight observed from this graph is the prevalence of male patients with bladder tumors after 65 years old. Descriptive information on metadata is available in the Supplementary Materials Section 1. Categorical data from all patients was compared to survival chance: chi-square test of independence is significant for age and human sub-populations, meaning that these two variables are related to disease outcome (SM Table 35). However, insights provided by this statistical test are limited to the sample under exam and cannot be generalized.

2.4. *Data discretization*

Gene expression measures the activity of a gene as reflected by the number of its RNA copies present at a specific moment in a cell. Laboratory measurements of GED are represented in a continuous domain of values, but there are reasons why it could be useful to infer gene expression data in the discrete domain. In a data-driven analysis, discretization reduces the amount of information and simplifies the learning process, for instance, to hasten training of gradient boosting decision trees [20]. Heterogeneous data-sets

Table 1. Sum-up table of log-transformed and z-scored GED

Gene	Type	Range	0.25 Perc.	0.75 Perc.	Median	Normal	Skewness	Kurtosis	Outliers
TPM1	HUB	5.503	-0.796	0.674	-0.02	Yes	0.226	-0.086	2
KPNA2	HUB	6.988	-0.558	0.696	0.215	No	-0.976	1.64	2
CRYAB	HUB	5.349	-0.717	0.679	-0.067	Yes	0.201	-0.366	0
CASQ2	HUB	3.936	-0.807	0.696	-0.387	No	1.008	-0.048	0
CDC20	HUB	6.972	-0.502	0.65	0.162	No	-0.951	1.704	6
CDK1	HUB	5.773	-0.595	0.676	0.168	No	-0.747	0.623	2
CCNB1	HUB	5.95	-0.569	0.695	0.117	No	-0.546	0.343	3
CCNA2	HUB	5.669	-0.616	0.736	0.126	No	-0.567	0.098	1
KIF11	HUB	6.248	-0.567	0.72	0.164	No	-0.751	0.728	3
UBE2C	HUB	8.652	-0.462	0.672	0.192	No	-1.354	3.913	7
MAD2L1	HUB	7.53	-0.589	0.7	0.191	No	-0.794	1.578	2
AURKA	HUB	5.867	-0.574	0.675	0.167	No	-0.72	0.548	2
KIF20A	HUB	6.288	-0.569	0.638	0.145	No	-0.794	1.168	5
KIF2C	HUB	6.345	-0.482	0.681	0.188	No	-1.104	1.79	8
TAGLN	SEED	4.86	-0.757	0.789	-0.067	No	0.141	-0.657	0
PLAU	SEED	5.077	-0.707	0.702	0.062	No	-0.237	-0.459	0
SLMAP	SEED	8.098	-0.62	0.637	0.066	No	0.034	1.496	8
TUBA1C	SEED	8.343	-0.581	0.674	0.112	No	-1.082	3.467	4
DMD	SEED	4.108	-0.997	0.766	-0.369	No	0.689	-0.524	0
AP2S1	SEED	7.005	-0.614	0.597	-0.033	No	0.278	0.568	4
SH3BGR	SEED	7.748	-0.666	0.587	0.059	No	-0.006	1.175	6
TCEAL2	SEED	5.115	-0.714	0.568	-0.714	No	1.475	1.871	0
ATP2B4	SEED	5.984	-0.558	0.68	0.095	No	-0.533	0.444	3
HJURP	SEED	7.489	-0.549	0.703	0.108	No	-0.805	1.54	6
SBSPON	SEED	4.881	-0.833	0.627	-0.162	No	0.497	-0.093	0

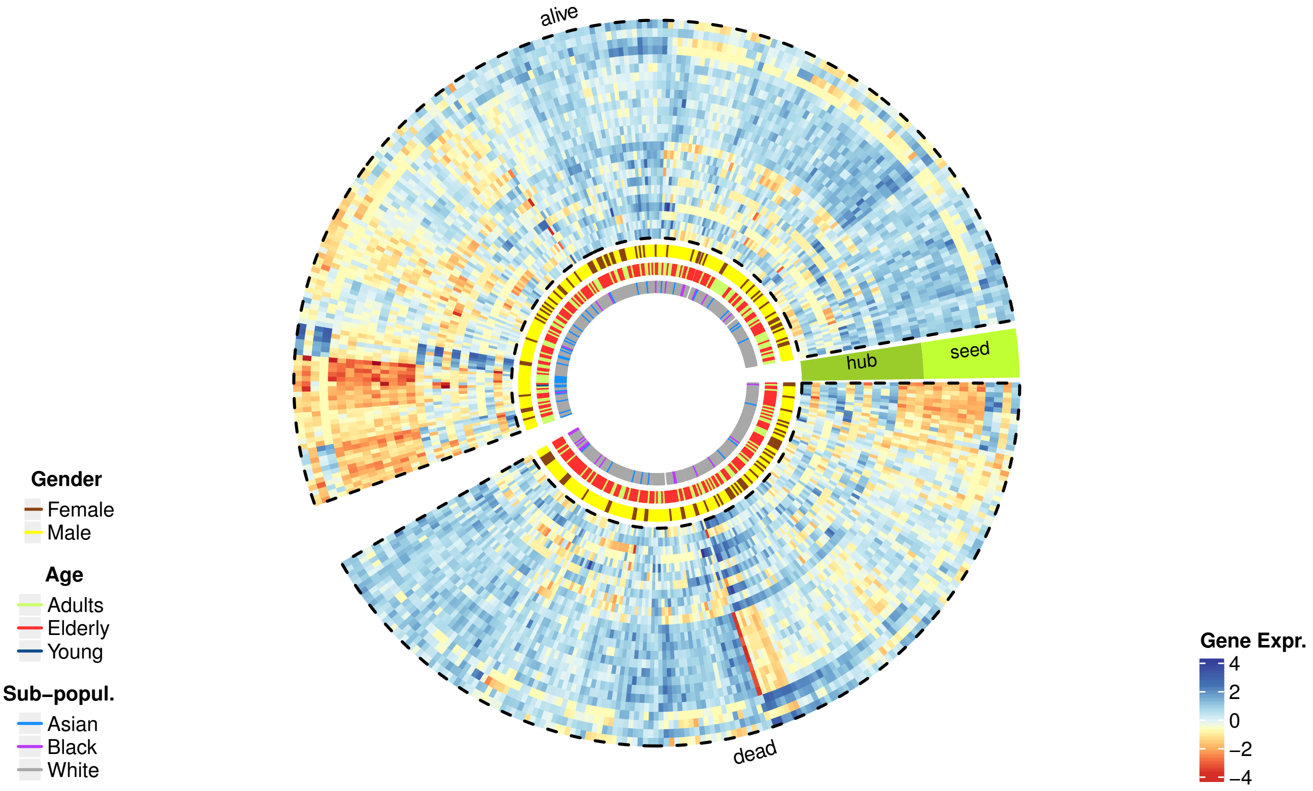


Figure 4. Visualization of the gene expression variables including annotations for all patients

could be more easily compared using binned representations. At the same time, another advantage is the suppression of the noise present in raw data. On the other hand, selecting an algorithm able to fit the values from the continuous to the discrete domain should be carefully evaluated to reduce the loss of information. Authors compared the performance of different classifiers applied to both continuous and discretized genetic expression data. The author's procedure included a preliminary reduction of the feature set to decrease their correlation using the minimum redundancy maximum relevance (MRMR) algorithm for binned data and F-test or t-test feature ranking on continuous values. Among classifiers applied on continuous and discretized data, their research showed better performance on the latter. Discretization is also a procedure that could facilitate the identification of sub-groups of genes involved in cancer genesis as found by [21]. The authors used a discretization scheme to define three intervals, finding a harmonious relationship between binned data and transcriptomic profiles in renal cell carcinoma. In general, discretization algorithms could be divided into *splitting* and *merging* algorithms [22], also called top-down and bottom-up (SM Section 3).

2.5. *Primary discretization algorithms*

Discretization of GED measurements creates a non-overlapping partition of the vast spectrum of continuous values coming from gene expressions. The fundamental division could be intended in three levels: "activation", "inhibition", or "no modifications" as assumed by some forms of the MRMR algorithm. However, there is no limit to the discretization levels achievable by an algorithm, and it depends on the kind of inference planned and also by the trade-off between computational complexity and information loss that each discretization entails [23]. Discretizers were evaluated in two modes

1. as a stand-alone primary discretizer
2. inserted in a sequence of two discretization phases: a pre-binning followed by a primary discretizer or a primary discretizer followed by refinement or optimization of the levels

Primary discretization algorithms are listed below, whereas extended details were enclosed in Supplementary Materials Section 3.1.

- CAIM (Class-Attribute Interdependence Maximization) [24]
- CACC (Class-Attribute Contingency Coefficient) value, as conceptualized by [25]
- Ameva, using the formulation proposed in [26]
- MDLP (Minimum Description Length Principle) [27]
- ChiMerge [28]
- Modified Chi2 (Mod Chi2) as presented in [29]
- Extended Chi2 (Ext Chi2) [30]
- CART decision tree-based discretization [31]
- XGB discretizer [32]

2.6. *Classifiers*

Baseline estimators included naïve bayes (NB), random forest (RF), logistic regression (LR), and support vector machines (SVM) similarly to [33] except for the usage of random forest instead of linear discriminant analysis. They were adjusted to account for class imbalance using cost-sensitive learning, weighting class instances during model training, or adjusting prior probabilities of the classes (for naïve bayes). A fifth classifier (i.e., Dummy [34,35]) acted as chance level reference and behaved independently from the information contained in the training data. During Section 3.2, the voting feature intervals classifier (VFI) [36] was applied on the discretized data-set; it has already been applied to biological data in [37]. Within Section 3.3 and Section 3.4, additional classifiers were added in the analysis and described in the relative section. Classifiers' efficiency was measured by ROC AUC applying 10-fold stratified cross-validation (CV) to preserve the representation of target classes in the training set. In addition, cross-validation offers the possibility of capturing the inter- and intra-subject variability typical of biomedical data showing the standard deviation of CV rounds. When performance scores were reported in tables, decimals were rounded to reduce row length.

2.7. *Statistical tests*

Non-parametric statistical tests were preferred for their capacity to handle unknown statistical distributions and for being less sensitive to outliers. Depending on the situation encountered the Mann-Whitney U, Wilcoxon signed-rank, or the Kruskal-Wallis tests were applied to determine differences between experimental results. In the case of Kruskal-Wallis test, post-hoc interpretation was with Dunn's test and Bonferroni correction. Tables with statistical summaries were included in Supplementary Materials (Section 9) with values rounded to reduce text width.

3. Results

Data investigation involved four different numerical experiments, each one with a dedicated pipeline for survival rate modeling, as summarized in Figure 5. A standard approach was embodied in Section 3.1, while three alternatives involving specific sequences of discretization stages were proposed throughout other numerical experiments: a primary discretizer preceding a refinement scheme (Section 3.2 and Section 3.3) or a pre-binning followed by the primary discretization algorithm (Section 3.4). In addition, a class separability measure was introduced (mathematical formulation in SM Section 4) to determine how severable a set of classes are in their multi-dimensional feature space: it provides an alternative score to support our analysis assessing PCA and NCA outcomes.

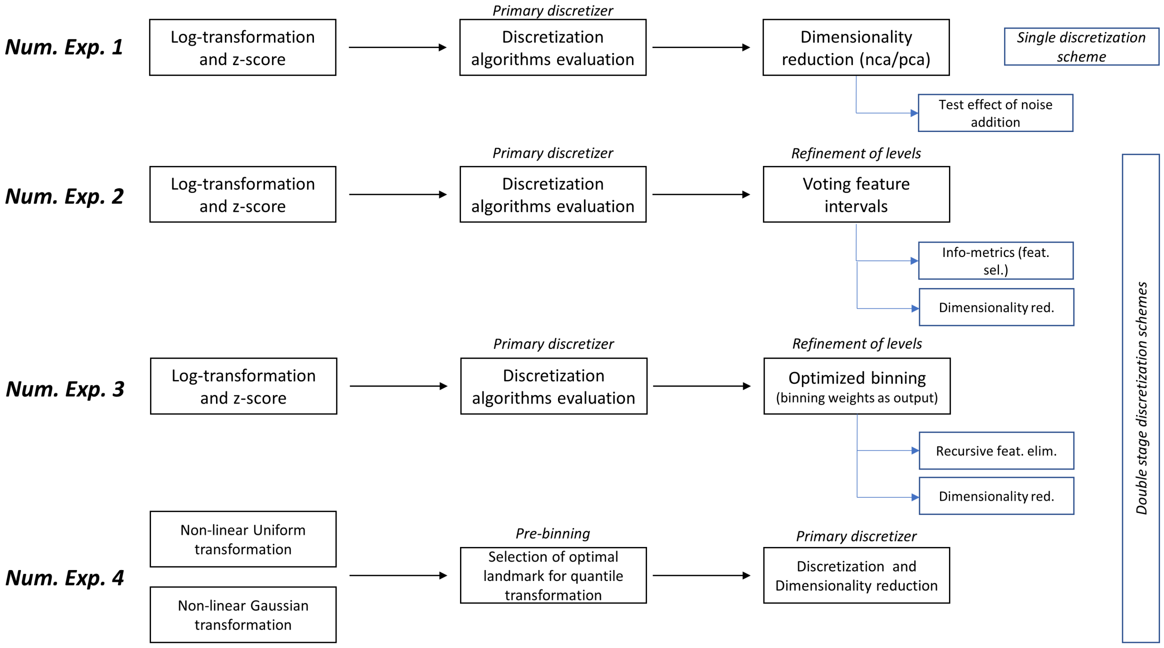

Num. Exp. 1
Log-transformation and z-score
Primary discretizer
Discretization algorithms evaluation
Dimensionality reduction (nca/pca)
Single discretization scheme
Test effect of noise addition
Num. Exp. 2
Log-transformation and z-score
Primary discretizer
Discretization algorithms evaluation
Refinement of levels
Voting feature intervals
Info-metrics (feat. sel.)
Dimensionality red.
Num. Exp. 3
Log-transformation and z-score
Primary discretizer
Discretization algorithms evaluation
Refinement of levels
Optimized binning (binning weights as output)
Recursive feat. elim.
Dimensionality red.
Num. Exp. 4
Non-linear Uniform transformation
Non-linear Gaussian transformation
Pre-binning
Selection of optimal landmark for quantile transformation
Primary discretizer
Discretization and Dimensionality reduction
Double stage discretization schemes

Figure 5. Overview of the four numerical experiments for GED analysis in bladder cancer

3.1. *Numerical Experiment 1*

The first experiment investigated if attributes of discretized data-set could be diminished and how this operation impacts classification. Two of the most popular dimensionality reduction techniques were applied: principal component analysis (PCA) and neighborhood components analysis (NCA). Initially, they were compared to non-linear techniques based on manifold learning and t-distributed stochastic neighbor embedding [38]. This preliminary evaluation showed that non-linear techniques did not overcome PCA and NCA performance. For this reason, NCA and PCA will be used throughout the study as leading methods for dimensionality compression. Discretized and reduced data-sets were graded using five classifiers: Table 2 and Table 3 reported mean ± standard deviation of the area under the receiver operating characteristic curve (ROC AUC) for the best discretizer. In a second attempt to verify both the robustness of outcomes and behavior of classifiers, ±5% uniform or gaussian random noise was added to the discretized levels. It could be asserted that certain discretization methods output a level in the form of an integer after assembling data in intervals, and not all classifiers can correctly handle integer values. For this reason, additional tests after noise injection were carried out together with calculation of the class separability metric (SM Section 4) because, in classification problems, the reconstruction error alone does not measure the quality of the subspace: in supervised learning, class labels are available, and the discriminative ability could be taken into consideration. Tables collecting analysis with the addition of noise are available in the supplement (SM Section 5: Table 11, Table 12 for PCA and Table 13, Table 14 for NCA).

Table 2. Best discretizer using different configurations of PCA components

PCA (ROC AUC)					
Comp.	RF	LR	NB	SVM	Dummy
2	Ext Chi2 0.64±0.1	CAIM 0.68±0.1	CAIM 0.68±0.1	CAIM 0.69±0.1	MDLP 0.54±0.07
3	Ext Chi2 0.62±0.09	CAIM 0.69±0.12	CAIM 0.69±0.1	CAIM 0.69±0.11	CART 0.55±0.09
4	Ext Chi2 0.63±0.1	CAIM 0.69±0.11	CAIM 0.68±0.1	CAIM 0.67±0.1	ChiMerge 0.54±0.06
5	AMEVA 0.64±0.12	CAIM 0.69±0.11	CAIM 0.67±0.1	CAIM 0.68±0.09	XGB 0.52±0.08
6	Ext Chi2 0.64±0.09	CAIM 0.69±0.11	CAIM 0.69±0.1	CAIM 0.68±0.09	CAIM 0.55±0.08
7	Ext Chi2 0.63±0.09	CAIM 0.68±0.12	CAIM 0.69±0.09	CAIM 0.67±0.1	CART 0.55±0.06
8	Ext Chi2 0.64±0.09	CAIM 0.68±0.12	CAIM 0.68±0.1	CAIM 0.66±0.11	Ext Chi2 0.52±0.08
9	Ext Chi2 0.63±0.1	CAIM 0.68±0.12	CAIM 0.68±0.1	CART 0.68±0.12	CART 0.53±0.08
10	Ext Chi2 0.64±0.09	CART 0.67±0.13	CAIM 0.68±0.11	CART 0.67±0.11	CART 0.55±0.06
11	Ext Chi2 0.64±0.1	CART 0.68±0.11	CAIM 0.67±0.11	CART 0.68±0.1	MDLP 0.55±0.1
12	Ext Chi2 0.64±0.09	CART 0.67±0.11	CAIM 0.66±0.1	CART 0.67±0.1	CART 0.53±0.1
13	AMEVA 0.64±0.12	CART 0.68±0.11	CART 0.66±0.12	CART 0.68±0.1	XGB 0.54±0.08
14	CART 0.64±0.08	CART 0.67±0.11	CAIM 0.67±0.12	CART 0.68±0.1	XGB 0.52±0.04
15	AMEVA 0.66±0.12	CART 0.71±0.11	CART 0.7±0.11	CART 0.7±0.11	CART 0.54±0.05
16	CART 0.68±0.09	CART 0.71±0.11	CART 0.69±0.11	CART 0.7±0.1	CART 0.54±0.06
17	CART 0.66±0.08	CART 0.71±0.11	CART 0.69±0.12	CART 0.7±0.11	CART 0.53±0.06
18	CART 0.67±0.09	CART 0.7±0.11	CART 0.69±0.12	CART 0.7±0.11	XGB 0.55±0.1
19	CART 0.68±0.09	CART 0.7±0.11	CART 0.69±0.12	CART 0.69±0.12	Mod Chi2 0.53±0.06
20	CART 0.66±0.05	CART 0.7±0.11	CART 0.69±0.11	CART 0.69±0.12	AMEVA 0.57±0.06
21	CART 0.67±0.09	CART 0.7±0.11	CART 0.7±0.09	CART 0.69±0.12	MDLP 0.55±0.05
22	CART 0.68±0.09	CART 0.69±0.12	CART 0.71±0.09	CART 0.68±0.12	CART 0.54±0.06
23	CART 0.68±0.09	CART 0.7±0.12	CART 0.71±0.09	CART 0.69±0.12	XGB 0.56±0.1
24	CART 0.7±0.08	CART 0.7±0.12	CART 0.71±0.09	CART 0.69±0.12	XGB 0.53±0.07
25	CART 0.66±0.08	CART 0.7±0.12	CART 0.71±0.09	CART 0.69±0.12	Ext Chi2 0.53±0.06

Kruskal-Wallis H-test confirms a significant difference between classifiers and the chance level (exemplified by the Dummy classifier) in the outcomes of CV for all NCA and PCA dimensions. Reduced number of NCA components in combination with the random forest classifier shows a stable performance on Table 3: between 2 to 8 NCA components, AUC ranges between 0.78 and 0.8 on XGB discretized data. Addition of uniform noise keeps AUC score stable (range 0.78 to 0.79, SM Table 13) maintaining the preference for XGB as discretizer. After insertion of gaussian noise (SM Table 14), the best performance (AUC range from the 2nd to the 7th components is between 0.77 and 0.81) is shared by XGB and MDLP as leading discretization algorithms. There is not noticeable oscillation of AUC values between noisy and non-noisy conditions confirmed by the lack of statistical significance at Kruskal-Wallis test (SM Table 45, SM Table 46). Standard deviation stays inside the range 0.06 to 0.09, even with a reduced number of components especially for RF classifier. Statistically, it is also significant the AUC variation between RF and the other classifiers, while there is no difference between NB, LR, and SVM outcomes (Dunn's post-hoc test, SM Table 37 and SM Table 40).

Class separability measure supports the same trend: when NCA is applied as method to decrease the dimensionality of the data-set, separability between two target classes increases as shown in Figure 2.

Table 3. Best discretizer using different configurations of NCA components

	NCA (ROC AUC)				
Comp.	RF	LR	NB	SVM	Dummy
2	XGB 0.79±0.08	CAIM 0.7±0.12	CAIM 0.72±0.1	CAIM 0.7±0.12	XGB 0.53±0.07
3	XGB 0.78±0.07	CAIM 0.7±0.12	CAIM 0.7±0.11	CAIM 0.7±0.12	MDLP 0.53±0.07
4	XGB 0.78±0.07	CAIM 0.7±0.09	CAIM 0.7±0.1	CAIM 0.7±0.09	Mod Chi2 0.55±0.06
5	XGB 0.79±0.08	CAIM 0.7±0.12	CAIM 0.71±0.11	CAIM 0.69±0.1	Mod Chi2 0.56±0.07
6	XGB 0.8±0.09	CAIM 0.7±0.1	CAIM 0.71±0.1	CAIM 0.69±0.1	CART 0.56±0.1
7	XGB 0.79±0.08	CAIM 0.69±0.1	CAIM 0.71±0.1	CAIM 0.69±0.09	AMEVA 0.53±0.08
8	XGB 0.8±0.08	CAIM 0.69±0.1	CAIM 0.7±0.11	CAIM 0.68±0.08	CACC 0.53±0.07
9	XGB 0.76±0.08	CAIM 0.69±0.1	CAIM 0.7±0.11	Ext Chi2 0.68±0.08	CART 0.53±0.07
10	XGB 0.78±0.06	Ext Chi2 0.68±0.08	CAIM 0.7±0.11	CART 0.68±0.11	CAIM 0.55±0.11
11	XGB 0.78±0.07	Ext Chi2 0.68±0.08	CAIM 0.7±0.11	CART 0.68±0.11	CART 0.53±0.07
12	XGB 0.78±0.09	Ext Chi2 0.68±0.08	CAIM 0.7±0.12	Ext Chi2 0.68±0.08	CAIM 0.55±0.08
13	XGB 0.76±0.05	Ext Chi2 0.68±0.08	CAIM 0.7±0.11	CART 0.68±0.1	ChiMerge 0.54±0.06
14	XGB 0.77±0.06	Ext Chi2 0.68±0.08	Ext Chi2 0.69±0.07	CART 0.69±0.1	MDLP 0.55±0.07
15	XGB 0.78±0.07	CART 0.71±0.11	CAIM 0.69±0.11	CART 0.71±0.11	CART 0.52±0.07
16	XGB 0.78±0.08	CART 0.7±0.12	CAIM 0.69±0.11	CART 0.68±0.11	AMEVA 0.54±0.06
17	XGB 0.76±0.06	CART 0.71±0.11	CAIM 0.7±0.12	CART 0.7±0.11	MDLP 0.58±0.06
18	XGB 0.77±0.06	CART 0.7±0.11	CAIM 0.7±0.12	CART 0.7±0.11	CART 0.54±0.06
19	XGB 0.77±0.05	CART 0.7±0.11	CAIM 0.7±0.12	CART 0.7±0.11	CART 0.55±0.05
20	XGB 0.77±0.05	CART 0.7±0.11	CAIM 0.7±0.11	CART 0.69±0.12	ChiMerge 0.54±0.08
21	XGB 0.77±0.07	CART 0.69±0.12	CAIM 0.7±0.12	CART 0.69±0.12	MDLP 0.54±0.08
22	XGB 0.75±0.08	CART 0.7±0.12	CAIM 0.7±0.11	CART 0.69±0.12	CART 0.53±0.07
23	XGB 0.77±0.07	CART 0.7±0.12	CAIM 0.7±0.11	CART 0.69±0.12	CART 0.55±0.08
24	XGB 0.76±0.06	CART 0.71±0.12	CAIM 0.7±0.11	CART 0.69±0.12	CACC 0.53±0.07
25	CART 0.75±0.07	CART 0.7±0.12	CAIM 0.72±0.11	CART 0.69±0.12	XGB 0.54±0.07

Furthermore, when the number of components decreases, class separability is higher. Class separability metric when comparing NCA vs PCA series shows a significant difference in all discretizers, while noise addition reduces score difference between NCA and PCA for certain discretizers only (Mann-Whitney rank test in SM Table 44). When effect of noise injection is evaluated separately for NCA and PCA series, noticeable differences on class separability metric are present in few discretizers: Modified Chi2, Extended Chi2 and MDLP, with Kruskal-Wallis H-test (SM Table 43). However, this effect on class separability measure is not fully embodied by classification outcomes.

In SM Figure 1 it is observable how random forest obtained the highest ROC AUC on NCA components from XGB the discretized data. Topmost AUC score with lowest number of NCA components was selected as simplified model and analyzed using precision, recall and F1 score. Concluding model for this experiment is the RF classifier on XGB data (Table 4).

Table 4. RF classifier on 5 NCA features selected from XGB discretized data

	precision	recall	f1-score	number of occurences
class: "Alive"	0.73	0.75	0.74	226
class: "Dead"	0.67	0.64	0.66	179
accuracy			0.70	405
macro avg	0.70	0.70	0.70	405
weighted avg	0.70	0.70	0.70	405

3.2. *Numerical Experiment 2*

VFI classifier can simultaneously act as a distiller of the discrete levels received as input and as a classifier for bladder cancer outcome modeling. It was implemented to refine the outcomes of the primary discretization algorithm, leading to a uniform leveling of the data, forcing constant interval widths for each feature of the data-set. In this procedure, end-points play a critical role because the algorithm takes advantage of considering only the lower bound as part of the interval, assuming that values increase monotonically. It happens naturally when discrete levels are sorted in ascending order. Initially, VFI discriminative performance was tested on the log-transformed data-set without discretization, but effectiveness increased when associated with discretized log-transformed values. Refinement of the VFI bins was also attempted with another technique, based intervals with the same nearest center as a 1D k-means cluster. However, k-means procedure provided intervals larger than those present in the original discretized levels created by the primary discretizers, increasing complexity of the model rather than simplifying it. For this reason, VFI was associated with uniform leveling, and it was tuned to obtain an optimal number of intervals checking AUC at CV. After tuning phase, prediction of the survival patterns was collected in Table 5, both as AUC and balanced accuracy. Number of discretized levels in the data-set changed according to each discretization algorithm and all feature had its own: in Table 5 it is reported the total number of unique levels.

Table 5. VFI Cross-Validation outcomes and optimal number of bins

Discr.	AUC	Bal. Accuracy	Original levels	VFI levels
CACC	0.99±0.003	97.63±2.65%	199	189
ChiMerge	0.98±0.01	94.39±3.34%	51	43
CART	0.89±0.08	82.55±10.17%	18	15
AMEVA	0.78±0.09	70.00±7.03%	8	6

So far, the analysis involved the whole feature set, but it could be meaningful to reveal how single features contribute to the prediction. For each component of the bladder cancer data-set, to evaluate its contribution in terms of entropy, information theory metrics IG, iv, and IGR (SM Section 6.1) were calculated and used to sort features contribution in descending order. The optimal number of features was identified by selecting those corresponding to the highest F1 score amid 10-Fold stratified CV; adding more features did not improve the predictive performance of the classifier. Results are in SM Figure 4 and also reported in Table 6.

Table 6. VFI Feature selection based on information theory metrics

Metric	Num. Feat.	AUC	Bal. Acc.	Aver. Prec.	F1
			ChiMerge		
IG	16	0.984±0.021	0.951±0.049	0.983±0.023	0.945±0.056
iv	20	0.987±0.017	0.949±0.03	0.986±0.017	0.944±0.033
IGR	24	0.987±0.017	0.946±0.031	0.987±0.016	0.941±0.034
			CACC		
IG	11	0.999±0.002	0.981±0.018	0.999±0.002	0.98±0.019
iv	21	0.998±0.003	0.981±0.021	0.997±0.003	0.98±0.022
IGR	19	0.997±0.004	0.98±0.018	0.997±0.004	0.98±0.019

Application of NCA for dimensionality reduction as an alternative to feature selection did not provide comparable results in performance with information theory metrics outcomes (results on NCA included in the Supplementary Materials Section 6.2). Feature reduction to 16 or 11 based on information gain (IG) preserves the performance of the VFI classifier but at the same time simplifies the model. Concluding, evaluation carried out using the best 11 features selected by IG criterion on the CACC discretized data (it is the procedure with the highest performance in Table 5) is shown in Table 7 after CV.

Table 7. VFI classifier over 11 IG features from CACC discretized data

	precision	recall	f1-score	number of occurences
class: "Alive"	0.97	1.00	0.98	226
class: "Dead"	0.99	0.97	0.98	179
accuracy			0.98	405
macro avg	0.98	0.98	0.98	405
weighted avg	0.98	0.98	0.98	405

3.3. *Numerical Experiment 3*

The analysis focused on binning data values exploiting the weight of evidence (WOE, SM Section 7.1), and evaluating its possible application in the field of gene expression data. The numerical approach tried during experiment 2 suggested further attempts towards a similar methodology because the primary discretizer probably produces sub-optimal intervals. Analysis advanced using primary binning algorithms accompanied by a secondary WOE phase, leading to an optimization of the levels based on the WOE principle. Optimal data split maximizes Jeffreys' divergence [39], also known as information value (IV) metric. At the end of this process, instead of using the numerical levels yield by the optimization procedure, weights computed during optimal split determination were used as input features for the classifiers (predictions reported in Table 8). Among different attempts (SM Tables 16, 17, 18, 19), the most successful approach required a sequence of CART algorithm and WOE optimized stage. WOE methods based on intervals with equal width, CART or intervals enclosing equalized frequency counts gave promising results to discriminate the disease outcome. Out of all classifiers tested, linear SVM was the utmost model.

Table 8. Two steps discretization with final optimized binning scheme

Data Transf.	Init. binning	Binning with opt.	ROC AUC	Classif.	Dummy
Log-transf. z-scores	CART	CART	0.868±0.070	Linear SVM	0.525±0.105
Log-transf. z-scores	CART	Equal-width	0.879±0.065	Linear SVM	0.490±0.072
Log-transf. z-scores	CART	Equal-freq	0.862±0.076	Linear SVM	0.490±0.097
Log-transf. z-scores	CART	MDLP	0.699±0.086	Gaussian Process	0.497±0.068

Finally, features extracted with the procedure as mentioned above were ranked by recursive feature elimination selecting only a subset of relevant ones with a stratified cross-validation (10 folds, SM Figure 5). The operational recipe of this procedure was uncovered to choose genes significant for cancer diagnosis. The optimal number of WOE features after recursive feature elimination is 21. On this reduced data-set, samples were evaluated using the equal-width optimization scheme with CART as the primary binning stage. In Table 8, this sequence leads to the most satisfactory results associated with a linear SVM classifier; thus, this pipeline was appraised separately Table 9 as a conclusive model.

Table 9. Linear SVM classifier with CART pre-binning and optimized discretization based on intervals with equal width

	precision	recall	f1-score	number of occurences
class: "Alive"	0.83	0.84	0.84	226
class: "Dead"	0.80	0.78	0.79	179
accuracy			0.81	405
macro avg	0.81	0.81	0.81	405
weighted avg	0.81	0.81	0.81	405

An alternative example established on a reduced number of NCA components is available in the Supplementary Materials Section 7.2.

3.4. *Numerical Experiment 4*

3.4.1. *Uniform distribution mapping*

In the course of previous experiments, a logarithmic transformation was employed before each empirical investigation. However, a single family of transformations may not approximate gene-specific variance. For example, authors of [40] demonstrated how an asymmetric Laplace distribution shows better goodness-of-fit rather than gaussian for microarray data. Among alternatives to normality, another distribution that found some interest for gene expression statistical analysis is the uniform one [41] both for microarray and RNA-seq data-sets. In numerical experiment 4, gene expression values were transformed according to a uniform mapping in range 0 to 1 using a set of landmarks to subdivide the cumulative distribution function. In this way, the uniform transformation could be interpreted as a pre-binning phase of the data. Mapping to a uniform distribution affects outliers, making them indistinguishable from inliers. In the data-set under investigation, the number of outliers seems negligible (Table 1), between 0% and 2% compared to the total number of samples. In this case, the transformation will not create bias by shrinking the distance between outliers and inliers. Initially, an exploratory study was prepared to determine the number of landmarks for the transformation (results in Supplementary Materials, Section 8). Five different values were selected as landmarks to determine the optimal cutting points of the cumulative distribution function (N/2, N/4, N/8, N/16, N/32). For landmark optimization, the CART algorithm was the preferred choice and used in later analysis stages (uppermost AUC among all discretizer, SM Table 31, SM Table 33). During the landmark tuning phase, NCA was favored over PCA because the class separability metric (SM Figure 7) showed a statistically significant difference between NCA and PCA with the Mann-Whitney rank test ($p \ll 0.01$ for all landmarks, SM Table 48). The NCA study that identified the optimal number of landmarks (SM Section 8) to develop a pre-binning stage in the final model is summarized in Figure 6. Number of landmarks corresponding to maximal AUC is 51 with 5 NCA components.

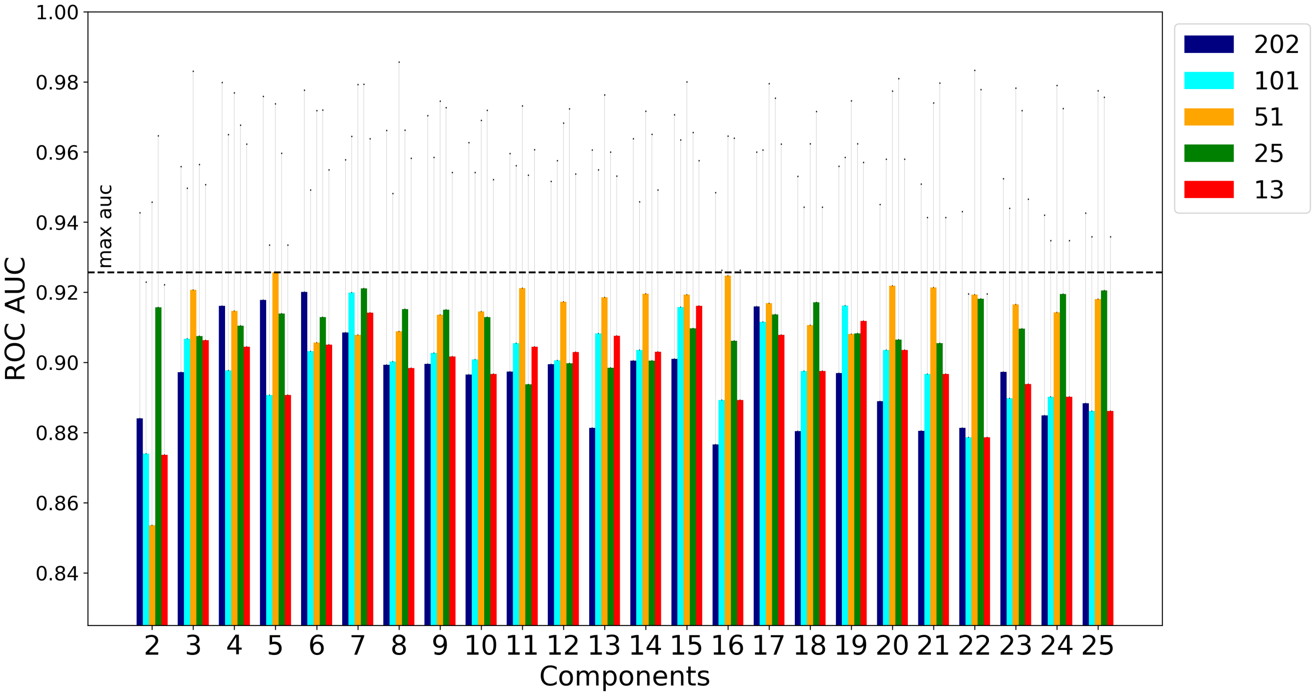


Figure 6. Optimal landmark selection: Uniform transf. NCA dim. red.

3.4.2. *Normal distribution mapping*

Biological data (if the sample is large enough) usually approximate a normal distribution [42] that embodies the natural inter-subject variability occurring for expression levels of all genes. In this section, each gene expression was mapped to a gaussian distribution with standardized outputs (mean value of 0 and a standard deviation of 1.0) to reproduce the theoretical normal distribution expected in biological samples. This methodology was similar to the one applied during uniform distribution mapping, including evaluating the number of landmarks needed to discretize the cumulative distribution function. Class separability measure showed a significant difference between landmarks at Mann-Whitney rank test mimicking the pattern seen for the uniform distribution (SM Table 50). For landmark selection, we applied a CART discretizer as primary algorithm together with NCA decomposition (SM Table 32, for PCA SM Table 34). CART algorithm was selected being the best one among others at exploratory study both for PCA and NCA. Top performance at cross-validation is summarized in Figure 7 with standard deviation added as a thin line with a marker on the endpoint. As found during uniform transformation, NCA maintains higher performance with fewer components than PCA: it could be possible to choose five NCA components with a reduced number of landmarks to obtain fair results and keep variance low (below ±5% of AUC). Moreover, reducing the landmarks for distribution mapping from 51 to 13 did not modify the classification outcomes in both cases. The concluding model consisted of 51 landmark points as estimated ranks of the cumulative distribution function needed to map a normal distribution, collecting five NCA components. These parameters were the same found during Uniform mapping, facilitating comparative analysis between models due to their congruity.

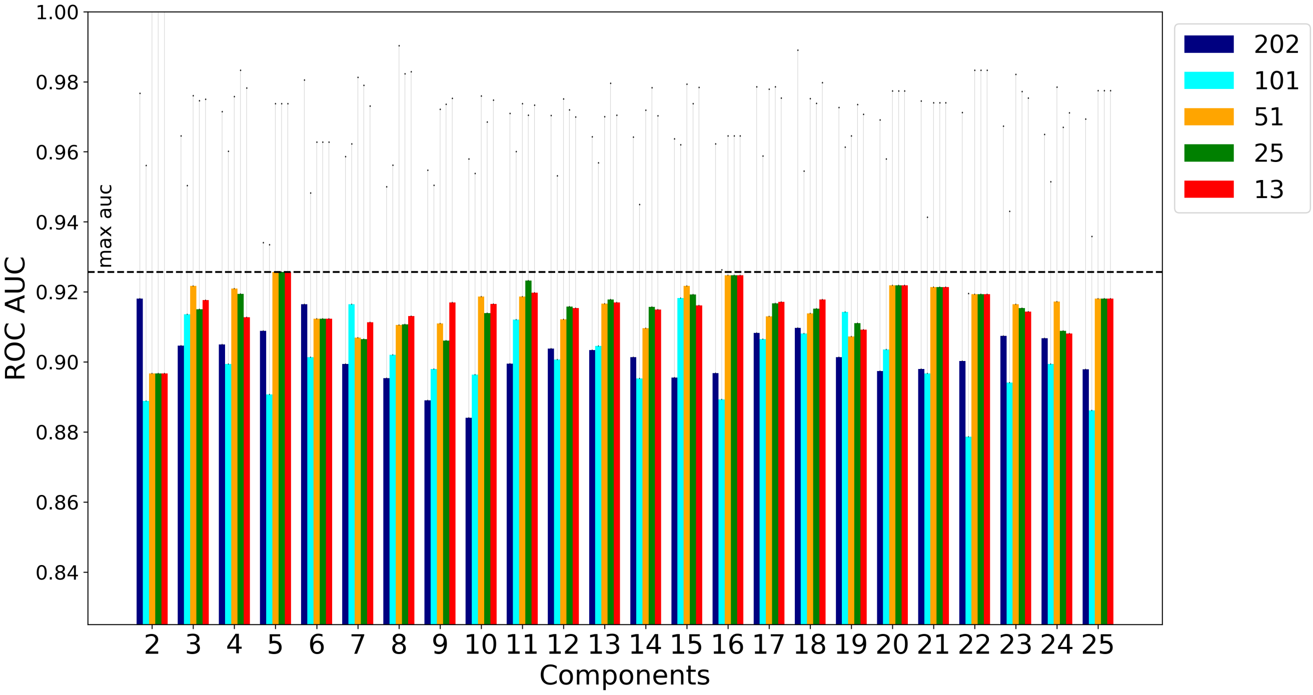


Figure 7. Optimal landmark selection: Normal transf. NCA dim. red.

3.4.3. *Models of numerical experiment 4*

CART discretized data from uniform or normal mappings reached top-notch AUC values (SM Tables 31, 32, 33, 34) both at NCA and PCA transformations. Five NCA components were enough to reach the peak AUC, and this parameter was selected for the conclusive Uniform (Table 10) or Normal (Table 11) models. Classifier with the top AUC is the k-NN, with little difference between k-NN outcomes and RF ones (SM Tables 23, 28). In conclusion, choosing 51 landmarks for both Uniform and Normal mappings seem an acceptable compromise to obtain an accurate data transformation and an initial pre-binning simultaneously.

Table 10. k-NN classifier on CART discretized data with 5 NCA components and 51 landmarks for quantile transformation as pre-binning (Uniform distr.)

	precision	recall	f1-score	number of occurences
class: “Alive”	0.87	0.90	0.88	226
class: “Dead”	0.87	0.83	0.85	179
accuracy			0.87	405
macro avg	0.87	0.87	0.87	405
weighted avg	0.87	0.87	0.87	405

Table 11. k-NN classifier on CART discretized data with 5 NCA components and 51 landmarks for quantile transformation as pre-binning (Normal distr.)

	precision	recall	f1-score	number of occurences
class: "Alive"	0.84	0.92	0.88	226
class: "Dead"	0.88	0.78	0.82	179
accuracy			0.85	405
macro avg	0.86	0.85	0.85	405
weighted avg	0.86	0.85	0.85	405

4. Discussion

In the present work, four different experiments were carried out to determine if quantitative gene alterations can effectively predict patient outcome after bladder cancer diagnosis. We experimented with distinct analysis pipelines throughout the study to forecast disease prognosis while trying to balance model complexity and simplification. Moreover, the involvement of different techniques and methodologies delivered four distinct approaches exposed in the numerical experiments as outlined in Figure 5.

During the first experiment (Section 3.1), a single data discretization step was taken into account. This first pipeline is the one that shows lower results in terms of predictive capacity. It can be considered a preliminary step to build up evidence towards using different strategies for analyzing the bladder cancer data-set. During this session, injection of ±5% uniform or gaussian random noise was attempted to evaluate classifiers' behavior after converting discrete levels into floats. It should be noted that not all discretization algorithms tested imply transforming the input values into integers. Among classifiers, the naïve bayes are specifically selected for their capacity to handle data converted in levels. However, the addition of noise didn't increase the performance of classifiers, implying it is possible to use primary discretizer outputs directly as classifiers inputs. Also, statistical test proved the absence of effect provided by noise addition. The last evaluation showed that RF performed better than other classifiers (accuracy 70%), especially on XGB discretized data.

A specific classifier was selected for the second experiment (Section 3.2): the VFI algorithm works on singles features (gene expressions), creating sets of values (intervals) for each feature grouped by classes (target variable). Combining all bound values of each feature, it could be possible to build consecutive intervals all along the training phase. During testing, counts of class instances falling into each interval were computed by voting to distinguish class membership. It was possible to notice how VFI slightly reduces the number of bins pre-computed by the primary discretization algorithm, with higher classification scores obtained in combination with ChiMerge and CACC (accuracy 98%). It is proper to mention how VFI acted as a refinement of the levels created by primary discretization, at the same time leading to an accurate prediction of the disease outcome.

The pipeline of the third numerical experiment (Section 3.3) involved a preliminary discretization of the data followed by an optimized binning based on the weight of evidence technique. Again, the optimized binning could be seen as a kind of refinement or adjustments of the levels produced by the primary discretization algorithm to fit the target variable better. Optimization of an initial granular discretization is a problem already addressed by commercial math software; however, we based our analysis on custom solution borrowed from [43] for credit risk modeling. In financial engineering, optimal binning is a problem primarily investigated for credit loss modeling to predict high or low-risk operations with credit cards [44]. In this study, the dichotomic nature of disease outcome is compatible with the "event" or "non-event" scheme used in financial modeling. During experiment three, after a primary discretization process, CART was chosen as the primary discretizer because it showed effective results along the whole analysis pipeline. A secondary binning phase tried to optimize CART levels using four methodologies (a tuned CART, MDLP, equal frequency intervals or equal width intervals). The weights associated with the optimized binning intervals were used as input features for classification. Weights associated with optimal intervals are in linear relation with them, allowing us to use weights instead of discrete levels to reduce computational times by exploiting floating-point arithmetic on commodity hardware. Final model of this section reached an accuracy of 81% with linear SVM classifier.

Forth numerical experiment (Section 3.4) tried to corroborate if other non-linear conversions could be successful alternatives to the log-transformation of GED attributes. GED was mapped to a uniform or a normal distribution, offering good strategies to substitute logarithmic transformation with the purpose of GED preparation before starting a machine learning workflow. In numerical experiment four, uniform and normal mapping were shaped to act as initial pre-binning, discretizing the cumulative distribution function by a certain number of landmarks (accuracy 87% and 85%, respectively). When log-transformation was combined with discretization algorithms (Section 3.1), it didn't provide better results than uniform or normal data mapping applied on the same data. In Section 3.4, AUC of the most exemplary discretization schemes have noticeable differences in performance across NCA dimensions. Kruskal-Wallis test confirms this observation (H=50.41, $p \ll 0.01$), and Dunn post-hoc test substantiates that performance across NCA dimensions is different between log-transformation and uniform or normal ones (SM Section 10). In GED analysis with positively skewed values, if scientists plan to apply a discretization scheme during the pre-processing phase, a uniform or normal transformation could be appraised as alternatives to the logarithmic one.

In experiment two (Section 3.2), the pipeline involving CACC and VFI offered the best modeling of the GED under exam. However, it should be mentioned that the present study has some limitations. All analyses were carried out on a single GED data-set. While potentially it is a helpful benchmark for other researchers working on the same kind of data, the generalizability of the results is restricted. In addition, genes were pre-selected among those that demonstrated solid prognostic value in previous investigations. Compared to biological data collected in laboratories, the data-set under investigation could be less noisy and redundant. Usually, noise decreases performance and increases the complexity of machine learning models; for this reason, it is a common practice to pre-process GED to clean out unnecessary noise [45]. Our models took advantage of this pre-selected pool of GEDs, excluding disturbance sources, hence focusing on robust algorithm evaluation. Lastly, data under investigation involves exposed individuals at the time of outcome status evaluation. This kind of dataset is suitable for machine learning modeling and hypothesis generation, but such models require further experimental studies for clinical validation.

4.1. *Algorithm derived from Numerical Experiment 2*

To sum up our analysis, the leading pipeline obtained during numerical experiment two (Section 3.2) is described in the form of a single algorithm, merging two different approaches (class-attribute contingency coefficient and voting feature intervals) in one routine. Given a feature f_j in the data-set composed of x_i samples and C classes, Algorithm 4.1 calculates *cacc* variable as in Section 2.5 together with VFI classifier intervals. After each attribute is discretized, routine counts the number of instances of each class on particular intervals using this information to build the voting scheme for that feature. Class membership of test set instances is determined separately for each feature, and then all contributions are summed to select predicted class by majority voting.

5. Conclusions

In this study, a newly released bladder cancer dataset was evaluated to provide machine learning models addressing survival rate prediction. For gene expression data analysis, a common practice is to log-transform raw gene expressions and work on discretized data, but in Sections 3.2, 3.3, and 3.4, we checked three expansions of this procedure. According to the present investigation, two-stage binning schemes had higher forecasting effectiveness compared to stand-alone discretization. This study gathers evidence towards using a double discretization approach on gene expression data analysis applying a primary discretizer followed by a refinement of the levels or a pre-binning employing data transformation before the main discretization algorithm. The usage of ChiMerge or CACC obtained the most satisfactory results as primary discretization algorithms followed by the voting feature intervals classifier acting as drainer of original levels and at the same time as a predictor. In the future, this double discretization approach will be furtherly extended to other datasets and types of omics data.

Algorithm 4.1 Combination of CACC and VFI in a single procedure

```
globalcacc=0
for all features $f_j$ in data-set do
    $D = [min(f_j); max(f_j)]$
    sort($f_j$)
    for $x$ in $f_j$ do
        $B = B \cup (x_{i-1} + x_i)/2$
    end for
    $D' = D \cup B$
    $k = 2$
    while cacc>globalcacc and k≤n do
        $cacc = \sqrt{\frac{y'}{y'+M}} \, \forall \, k \, intervals \, in \, D'$
        where $max(cacc) \rightarrow D \cup (x_{i-1} + x_i)/2$
        $k + +$
        globalcacc=$max(cacc)$
    end while
    $train_x, test_x$ = split($f_j$)
    for $c_i$ in C do
        $G = G \cup [min(train_x \in c_i); max(train_x \in c_i)]$
    end for
    $sort(G)$
    $\forall \, [g_{i-1}; g_i]$ in G and $\forall \, c_i$ in C, → voting scheme=count($train_x \in c_i$)/count($c_i$)
    for $x$ in $test_x$ do
        which $[g_{i-1}; g_i] \in x \rightarrow vote_x \leftrightarrow$ voting scheme
    end for
end for
$\sum$ vote $\forall$ features $\rightarrow \forall \, x \in test_x$, $max(vote)$ = predicted class
```

List of abbreviations

AUC Area Under Curve
CACC Class-Attribute Contingency Coefficient
CAIM Class-Attribute Interdependence Maximization
CART Classification And Regression Tree
Comp Components
CV Cross-Validation
GED Gene Expression Data
IG Information Gain
IGR Information Gain Ratio
iv Intrinsic value
IV Information Value
k-NN k-Nearest Neighbors
LR Logistic Regression
MDLP Minimum Description Length Principle
MRMR Minimum Redundancy Maximum Relevance
NB Naïve Bayes
NCA Neighborhood Components Analysis
PCA Principal Components Analysis
RF Random Forest
ROC Receiver Operating Characteristic

SVM Support Vector Machine
SM Supplementary Materials (Appendix file)
VFI Voting Feature Intervals
WOE Weight of evidence
XGB Extreme Gradient Boosting

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska Curie grant agreement No 860462. This manuscript reflects only the author's view and the Agency and the Commission are not responsible for any use that may be made of the information it contains. The authors thank the two anonymous reviewers whose comments helped improve and clarify this manuscript.

Author contributions

MN (formal analysis, investigation, methodology, visualization, manuscript writing), MV (project supervision, manuscript writing), LR (funding acquisition, project supervision, manuscript writing, project administration)

References

1. D. Wu, C. M. Rice, and X. Wang, Cancer bioinformatics: A new approach to systems clinical medicine, 2012.
2. S. Zheng, L. Yang, Y. Dai, L. Jiang, Y. Wei, H. Wen, and Y. Xu, Screening and survival analysis of hub genes in gastric cancer based on bioinformatics, *Journal of Computational Biology*, vol. 26, no. 11, pp. 1316–1325, 2019.
3. C. Zhang, M. Berndt-Paetz, and J. Neuhaus, Identification of key biomarkers in bladder cancer: Evidence from a bioinformatics analysis, *Diagnostics*, vol. 10, no. 2, p. 66, 2020.
4. P. Kutwin, T. Konecki, M. Cichocki, P. Falkowski, and Z. Jabłonowski, Photodynamic diagnosis and narrow-band imaging in the management of bladder cancer: a review, *Photomedicine and Laser Surgery*, vol. 35, no. 9, pp. 459–464, 2017.
5. I. Erb and C. Notredame, How should we measure proportionality on relative gene expression data?, *Theory in Biosciences*, vol. 135, no. 1-2, pp. 21–36, 2016.
6. C. A. Gallo, R. L. Cecchini, J. A. Carballido, S. Micheletto, and I. Ponzoni, Discretization of gene expression data revised, *Briefings in bioinformatics*, vol. 17, no. 5, pp. 758–770, 2016.
7. P. Domingos, The role of occam's razor in knowledge discovery, *Data mining and knowledge discovery*, vol. 3, no. 4, pp. 409–425, 1999.
8. C. Zhang, M. Berndt-Paetz, and J. Neuhaus, Bioinformatics analysis identifying key biomarkers in bladder cancer, *Data*, vol. 5, no. 2, p. 38, 2020.
9. S. v. Buuren and K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in r, *Journal of statistical software*, pp. 1–68, 2010.
10. B. V. Church, H. T. Williams, and J. C. Mar, Investigating skewness to understand gene expression heterogeneity in large patient cohorts, *BMC bioinformatics*, vol. 20, no. 24, pp. 1–14, 2019.
11. Y. Chen, S. Tu, and L. Xu, The prognostic role of genes with skewed expression distribution in lung adenocarcinoma, in *International Conference on Intelligent Science and Big Data Engineering*, pp. 631–640, Springer International Publishing, 2017.
12. J. R. Holland, J. D. Baeder, and K. Duraisamy, Towards integrated field inversion and machine learning with embedded neural networks for rans modeling, in *AIAA Scitech 2019 Forum*, p. 1884, American Institute of Aeronautics and Astronautics, 2019.
13. D. George and M. Mallery, *Using SPSS for Windows step by step: a simple guide and reference.* Boston, MA: Allyn & Bacon, 2003.

14. T. Speed, Always log spot intensities and ratios, *Speed Group Microarray Page, at http://www. stat. berkeley. edu/users/terry/zarray/Html/log. html*, 2000.
15. C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, Analysis of microarray data using z score transformation, *The Journal of molecular diagnostics*, vol. 5, no. 2, pp. 73–81, 2003.
16. R. D'Agostino and E. S. Pearson, Tests for departure from normality. empirical results for the distributions of b^2 and $\sqrt{b}$, *Biometrika*, vol. 60, no. 3, pp. 613–622, 1973.
17. C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median, *Journal of Experimental Social Psychology*, vol. 49, no. 4, pp. 764–766, 2013.
18. F. E. Harrell and C. Davis, A new distribution-free quantile estimator, *Biometrika*, vol. 69, no. 3, pp. 635–640, 1982.
19. Z. Gu, L. Gu, R. Eils, M. Schlesner, and B. Brors, circlize implements and enhances circular visualization in r, *Bioinformatics*, vol. 30, no. 19, pp. 2811–2812, 2014.
20. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, Lightgbm: A highly efficient gradient boosting decision tree, in *Advances in neural information processing systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
21. M. Beleut, R. Soeldner, M. Egorov, R. Guenther, S. Dehler, C. Morys-Wortmann, H. Moch, K. Henco, and P. Schraml, Discretization of gene expression data unmasks molecular subgroups recurring in different human cancer types, *PloS one*, vol. 11, no. 8, p. e0161514, 2016.
22. S. Kotsiantis and D. Kanellopoulos, Discretization techniques: A recent survey, *GESTS International Transactions on Computer Science and Engineering*, vol. 32, no. 1, pp. 47–58, 2006.
23. L. Peng, W. Qing, and G. Yujia, Study on comparison of discretization methods, in *2009 International Conference on Artificial Intelligence and Computational Intelligence*, vol. 4, pp. 380–384, IEEE, 2009.
24. L. A. Kurgan and K. J. Cios, Caim discretization algorithm, *IEEE transactions on Knowledge and Data Engineering*, vol. 16, no. 2, pp. 145–153, 2004.
25. C.-J. Tsai, C.-I. Lee, and W.-P. Yang, A discretization algorithm based on class-attribute contingency coefficient, *Information Sciences*, vol. 178, no. 3, pp. 714–731, 2008.
26. L. Gonzalez-Abril, F. J. Cuberos, F. Velasco, and J. A. Ortega, Ameva: An autonomous discretization algorithm, *Expert Systems with Applications*, vol. 36, no. 3, pp. 5327–5332, 2009.
27. U. Fayyad and K. Irani, Multi-interval discretization of continuous-valued attributes for classification learning, in *Proceedings of the 13th international joint conference on artificial intelligence*, pp. 1022–1027, IJCAI, 1993.
28. R. Kerber, Chimerge: Discretization of numeric attributes, in *Proceedings of the tenth national conference on Artificial intelligence*, pp. 123–128, AAAI Press, 1992.
29. F. E. Tay and L. Shen, A modified chi2 algorithm for discretization, *IEEE Transactions on knowledge and data engineering*, vol. 14, no. 3, pp. 666–670, 2002.
30. C.-T. Su and J.-H. Hsu, An extended chi2 algorithm for discretization of real value attributes, *IEEE transactions on knowledge and data engineering*, vol. 17, no. 3, pp. 437–441, 2005.
31. L. Reiman, J. Friedman, R. Olshen, and C. Stone, *Classification and Regression trees (Belmont, California: Wadsworth Ind. Group)*. Wadsworth Ind. Group, 1984.
32. T. Chen and C. Guestrin, Xgboost: A scalable tree boosting system, in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, New York, NY, USA: Association for Computing Machinery, 2016.
33. C. Ding and H. Peng, Minimum redundancy feature selection from microarray gene expression data, *Journal of bioinformatics and computational biology*, vol. 3, no. 02, pp. 185–205, 2005.
34. G. Figueroa, Y.-S. Chen, N. Avila, and C.-C. Chu, Improved practices in machine learning algorithms for ntl detection with imbalanced data, in *2017 IEEE Power & Energy Society General Meeting*, pp. 1–5, IEEE, 2017.

35. A. Martino, A. Rizzi, and F. M. F. Mascioli, Supervised approaches for protein function prediction by topological data analysis, in *2018 International joint conference on neural networks (IJCNN)*, pp. 1–8, IEEE, 2018.

36. G. Demiröz and H. A. Güvenir, Classification by voting feature intervals, in *European Conference on Machine Learning*, pp. 85–92, Springer, 1997.

37. F. Ali and M. Hayat, Classification of membrane protein types using voting feature interval in combination with chou pseudo amino acid composition, *Journal of theoretical biology*, vol. 384, pp. 78–83, 2015.

38. L. v. d. Maaten and G. Hinton, Visualizing data using t-sne, *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

39. H. Jeffreys, An invariant form for the prior probability in estimation problems, *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

40. E. Purdom and S. P. Holmes, Error distribution for gene expression data, *Statistical applications in genetics and molecular biology*, vol. 4, no. 1, 2005.

41. Z. Fang, R. Du, and X. Cui, Uniform approximation is more appropriate for wilcoxon rank-sum test in gene set analysis, *Plos One*, vol. 7, no. 2, p. e31505, 2012.

42. M. C. Whitlock and D. Schluter, *The analysis of biological data.* Roberts and Company Publishers, 2009.

43. G. Navas-Palencia, Optimal binning: mathematical programming formulation, *arXiv preprint arXiv:2001.08025*, 2020.

44. R. Anderson, *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation.* Oxford University Press, 2007.

45. G. L. Libralon, A. C. P. de Leon Ferreira, A. C. Lorena, *et al.*, Pre-processing for noise detection in gene expression classification data, *Journal of the Brazilian Computer Society*, vol. 15, no. 1, pp. 3–11, 2009.