RESEARCH

On the challenges of predicting treatment response in Hodgkin's Lymphoma using transcriptomic data

André Patrício[1,2*], Rafael S. Costa[1,4†] and Rui Henriques[2,3†]

[*]Correspondence: andremppatricio@tecnico.ulisboa.pt
[1]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Lisbon, Portugal, Portugal
Full list of author information is available at the end of the article
[†]Co-last author

Abstract

Background: Despite the advancements in multiagent chemotherapy in the past years, up to 10% of Hodgkin's Lymphoma (HL) cases are refractory to treatment and, after remission, patients experience an elevated risk of death from all causes. These complications are dependent on the treatment and therefore an increase in the prognostic accuracy of HL can help improve these outcomes and control treatment-related toxicity. Due to the low incidence of this cancer, there is a lack of works comprehensively assessing the predictability of treatment response, especially by resorting to machine learning (ML) advances and high-throughput technologies.

Results: We present a methodology for predicting treatment response after two courses of Adriamycin, Bleomycin, Vinblastine and Dacarbazine (ABVD) chemotherapy, through the analysis of gene expression profiles using state-of-the-art ML algorithms. The presented approach combines dimensionality reduction procedures and hyperparameter optimization of various elected classifiers to retrieve reference predictability levels of refractory response to ABVD treatment using the regulatory profile of diagnostic tumor samples. In addition, and foremost, we propose a data transformation procedure to map the original data space into a more discriminative one using biclustering, where features correspond to discriminative putative regulatory modules. This methodology presents increased performance against reference levels, with the proposed space transformation yielding improvements in the majority of the tested predictive models (e.g. Decision Trees show an improvement of 20pp in both precision and recall).

Conclusion: Taken together, the results reveal improvements for predicting treatment response in HL disease by resorting to sophisticated statistical and ML principles. This work further consolidates the current hypothesis on the structural difficulty of this prognostic task, showing that there is still a considerable gap to be bridged for these technologies to reach the necessary maturity for clinical practice.

Keywords: Hodgkin's lymphoma; cancer; machine learning; gene expression; discriminative patterns; biclustering

Background

Hodgkin's Lymphoma (HL) is a type of blood cancer that originates in the lymphatic system, more precisely in lymphocytes, with the patient age peak of diagnostics occurring at the 20s and 30s. In 2018, HL represented 0.4% of all new

tumors (79990 new cases) and 0.3% of all cancer deaths (26167 deaths) worldwide [1]. Survival of Hodgkin's Lymphoma patients has significantly improved over the past years. Still, after initial remission, patients experience an elevated risk of death from all causes [2], such as cardiotoxicity diseases like myocardial infarction and congestive heart failure [3], and secondary cancers [4], diseases that are often treatment-related [5].

The current prognosis for HL is largely based on the International Prognostic Score (IPS) [6] which predicts for 5-year freedom from progression. Moccia et al. [7] concluded that this scoring does not identify with certainty low or high risk groups, and recommends the use of molecular markers and/or fluorodeoxyglucose Positron Emission Tomography (FDG-PET) scanning for this purpose. Despite the proven relevance of FDG-PET for HL prognostic, this medical exam is: i) intrusive, with the need to inject a radioactive tracer; ii) expensive, estimated at 1020 Eur per exam [8]; and iii) impossible to perform in remote locations and ambulatory settings as it requires large machinery.

The transcriptional activity of tumor cells is a viable proxy candidate to assess regulatory response to treatment, thus being positioned as a possible alternative to the FDG-PET exam. Nevertheless, the role of differential gene expression in HL has not been exhaustively studied as it is a relatively rare cancer (2.86 cases per 100,000 persons annually [9]). Specific approaches, such as hierarchical clustering [10, 11], Cox regression [12] and sparse multinomial logistic regression [13] have been explored in other works, but some of the state-of-the-art machine learning (ML) approaches successfully applied to more common cancers have not yet been comprehensively employed.

In this context, this work proposes a superior methodology to predict the result of an interim FDG-PET performed after two courses of Adriamycin, Bleomycin, Vinblastine and Dacarbazine (ABVD) chemotherapy treatment through the analysis of gene expression profiles using state-of-the-art ML techniques. Transcriptomic data of Hodgkin's Lymphoma patients' diagnostic tumor samples acquired by Luminari et al. [14] are used with the objective of better understanding the predictability of a patient's response to a specific chemotherapy regimen. To this end, we resort to gene expression profiles obtained from Formalin Fixed Paraffin Embedded (FFPE) diagnostic tumor samples [15].

Our work advances the current status quo on this task placed by Luminari et al. [14], which is grounded on a more traditional statistical analysis of the data, resorting to multivariate logistic analysis, filtering by Fold-Change (FC) and False Discovery Rate (FDR) values and multivariate logistic regression. In contrast, we conduct a thorough optimization and assessment of preprocessing and ML techniques to develop a predictor that can, at the moment of diagnosis, classify patients' future interim PET after two courses of ABVD chemotherapy according to treatment response.

In addition to the end-to-end assessment of state-of-the-art predictors, our work proposes the use of biclustering principles to transform the original high-dimensional feature space into one consisting of features given by discriminative gene expression patterns, and shows that the new space yields relevant statistical properties. The gathered results show that this novel transformation yields statistically significant improvements on predictive performance.

Methods

To tackle the introduced research problem, we propose the methodology presented in Figure 1. This methodology starts with essential data preprocessing, followed by a feature analysis stage divided into two phases, resulting in a reduced data space conducive to the subsequent predictive analysis stage. Along the predictive analysis stage, we propose a bicluster-based space transformation that converts the gene-centric space to a pattern-centric one. State-of-the-art ML models can then be applied along the original or pattern-centric data space for the targeted prognostic ends.

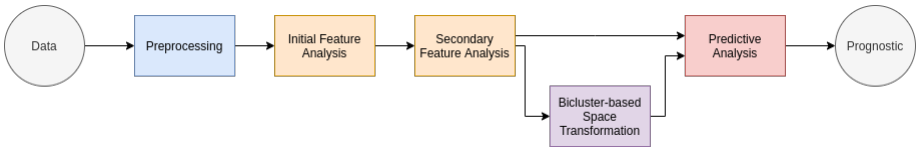

Figure 1: Schematic workflow of proposed solution for the treatment response prediction in HL disease.

Data

This work uses the data from the cohort study conducted by Luminari et al. [14], available at the National Center for Biotechnology Information Gene Expression Omnibus, GSE132348[1]. It consists of 106 samples of patients diagnosed with Classical Hodgkin's Lymphoma. Each individual has associated the normalized expression levels of 765 different genes, obtained using the NanoString's nCounter platform [15] over the RNA extracted from FFPE diagnostic tumor samples. The following clinical variables are also included: gender, age, stage of disease according to the Lugano classification [16], and Lymphocyte-to-Monocyte Ratio (LMR), LMR>2.1. Finally, each record contains the result of an interim PET realized after two courses of ABVD chemotherapy (iPET2), which was classified as "positive" or "negative" according to its classification on the Deauville 5-point scale [17], with PET defined as positive when its ordinal value is greater or equal than 4. More information on the data collection process can be found in the original work [14]. The data is relatively imbalanced, with 84 (80%) iPET2 negatives, and 21 (20%) iPET2 positives.

Data Preprocessing

Samples with missing values were removed, resulting in a new distribution of 82 (79.6%) iPET2 negatives and 21 (20.4%) iPET2 positives. The variable *stage* was encoded as an ordinal variable, ranging from 1 (corresponding to"I A") to 8 (corresponding to "IV B"), indicating that both a larger stage number and the B variant are worse prognostic factors. The mRNA counts were $\log_2$ transformed in order to better handle the variability of expression within and across genes. Since the dataset is considerably imbalanced, with a target's distribution of around 80/20, balancing methods were assessed through a comparative analysis of their effects on predictive performance, with the combination of both oversampling and subsampling techniques being compared. Oversampling using Support Vector Machine

[1] `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE132348` (accessed June 28, 2021)

Synthetic Minority Oversampling Technique (SVM-SMOTE)[2] [19] yield the best result, being the balancing procedure selected for the target experiments.

Feature Analysis

The data at hand is high-dimensional, 770 gene features with variable expression, and has a low number of samples, 103 patients. In this context, the posterior predictive analysis can benefit from dimensionality reduction. Following the practice suggested by Saeys et al. [20], we pre-reduced the feature space using univariate filter methods Wilcoxon Rank Sum Test[3] [22] and Mutual Information[4] [24], and subsequently applied a more complex embedded method Support Vector Machine - Recursive Feature Elimination (SVM-RFE)[4] [25]. This decision is supported by results in the literature where univariate filter methods have similar or better performance than more complex embedded methods [26–28].

Initial Feature Selection. As most preprocessed features do not follow a normal distribution according to Shapiro-Wilk test [29], we resort to the non-parametric Wilcoxon Rank Sum Test [22] and Mutual Information (MI) [24] criteria. For each independent variable y_j and the target response outcome z, the former Wilcoxon statistic tests whether the distributions $y_j|z = 0$ and $y_j|z = 1$ are equal. The latter MI statistic does not test a hypothesis, but a p-value is subsequently generated using a one-sided permutation test. As these approaches are univariate, not taking into account interactions between variables, a less strict than usual significance threshold of 0.1 is considered to prevent the removal of potentially relevant genes.

Secondary Feature Selection. Grounded on empirical evidence, we subsequently employ the embedded Support Vector Machine - Recursive Feature Elimination (SVM-RFE) [25], an adaption of the RFE selection method that replaces an external ranking function with the magnitude of the weights of a Support Vector Machine. This algorithm has been in fact proposed for gene selection in cancer classification tasks and stands out for its good results, especially when considered in combination with Support Vector Machine (SVM) classifiers [30].

Predictive Analysis

The literature on classification tasks in the oncotranscriptomics domain shows a relative predominance of specific machine learning (ML) models [31, 32]. Accordingly, Support Vector Machines (SVM) [33], k-Nearest Neighbors (KNN) [34] and Random Forest [35] are selected as adequate classifier candidates. Given the consistently state-of-the-art performance of XGBoost [36] in other domains, we further disclose its predictability performance. Complementarily, additional predictive models are further considered in our study, including Decision Trees [37] as the learnt associations can reveal important insights about the genetic component of HL, and Naive Bayes [38] to offer a baseline stance on predictive accuracy.[4]

In order to obtain the best classification possible, all the predictors are subjected to parameter optimization through the Tree Parzen Estimator algorithm [39], in addition to an optimized application of the aforementioned preprocessing techniques.

[2] implemented using `imbalanced-learn` [18]
[3] implemented using `SciPy` [21]
[4] implemented using `scikit-learn` [23]

Evaluation Methodology

Given the small population size (103 samples), a nested Cross-Validation (CV) schema is considered using 10 folds. Nested Cross-Validation is used to separate the data used for hyperparameterization from the data used for testing the model [40]. Without this separation, the model would learn its parameters in data in which it would be tested, resulting in an overestimated predictive capability. In a nested CV, inside each CV loop used for model evaluation, another CV must be performed in the training data for parameter tuning and preprocessing.

Performance Metrics

A core contribution of this work is the possibility to create a reliable decision support system that can help decide the intensity of the treatment a patient must undergo, with a positive prediction indicating that a more aggressive regime is necessary. Attending to this observation, the following evaluation metrics were chosen: i) AUC as an overall indicator of how the predictor performs when the decision threshold is not optimized; ii) recall and precision to ensure that the predictor does not skew towards the majority class, especially important due to the imbalanced data nature; and iii) specificity to guarantee the identification of patients who will react well to the standard treatment, avoiding the prescription of an unnecessarily stronger chemotherapy regimen. The metric optimized in the hyperparameterization step is the F1 score, so that the classifiers can attain a balanced performance in both precision and recall. Precision-Recall curves are further plotted to offer a more comprehensive comparison of the predictors.

Bicluster-based Space Transformation

The various biological mechanisms present in our bodies rarely originate due to a single gene's expression values. Instead, the majority of gene regulation is done in a modular way, in which sets of genes interact with each other to enforce a certain mechanism. With this in mind, it is important to not reduce the analysis of transcriptomic data to individual genes but instead to study these values and its contributions in the context of putative gene modules.

Furthermore, although the feature selection process here presented is heavily explored in order to obtain the best possible feature space for classification, this approach has its limitations. It relies on univariate methods that do not take into consideration gene interactions and on an embedded method that can show biases towards specific predictors.

In this context, we propose a novel transformation procedure in high-dimensional data spaces that can map individual gene expression features with loose discriminative power to discriminative pattern-centric features given by discriminative regulatory modules able to better represent complex interactions between multiple genes.

The biclustering task is suggested towards this end as it has been largely employed for the efficient and effective mining of gene expression patterns [41–43], with pattern-based biclustering showing relevant performance indicators in diverse biological data contexts [44, 45]. Biclustering based on PAttern Mining Software (BicPAMS) [46] integrates dispersed state-of-the-art contributions on pattern-based

biclustering, allowing for a high level of parametrization while performing efficient searches with guarantees of optimality, statistical significance and discriminative power [47, 48]. BicPAMS is applied to find discriminative patterns on the training data. Multiple parameterizations are tested in order to obtain the best space transformation for the classification task. The dataset is then transformed so that each feature corresponds to one gene expression pattern encompassing multiple genes. Each value of the transformed dataset will then represent the similarity between the gene expression expectations associated with a given pattern and the actual levels of gene expression observed for an individual. Figure 2 shows an example of the described process.

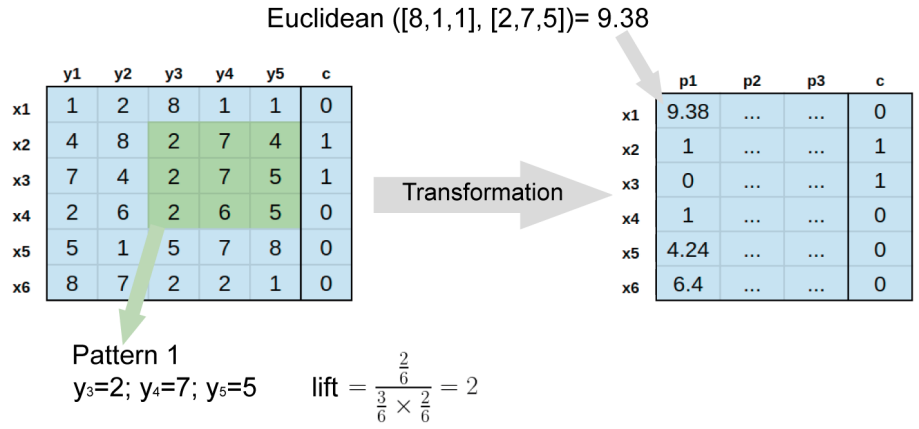

Figure 2: Example of space transformation using biclustering, showing a found bicluster and corresponding pattern (left), and the resultant values on the transformed dataset using Euclidean distance (right).

Results

The proposed methodology is applied on the cohort study conducted by Luminari et. al [14] to assess the limits to the predictability of the quality of HL patient response to ABVD treatment. The gathered results are presented in two major steps: i) a comparison of predictive levels from the optimized application of state-of-the-art ML models against the current reference levels; and ii) an assessment of the improvement yield by the proposed pattern-centric data space transformation. The methodology is implemented in Python 3.8.5 and the experiments run on Intel® Core™ i7-5600U CPU running at 2.60GHz.

Predictive performance under state-of-the-art ML

The initial phase of feature selection by the Mutual Information and Wilcoxon Rank Sum algorithms identified a total of 250 features out of the initial 770. These features correspond to 248 genes and the clinical variables *stage* and *LMR>2.1*. The work of Luminari et al. [14] found, during a first analysis, a 13-gene signature positively correlated with iPET2 in addition to the variable *LMR>2.1*. In comparison, we find that our most influential feature set contains 9 out of the 13 genes and the *LMR>2.1* variable. The selection of the variable *stage* by our algorithm points to

a relation with the target iPET2 not identified through the multivariate logistic analysis performed by Luminari et al. [14]. The second and final phase of feature selection, performed by the algorithm SVM-RFE, identified a set of 14 genes. Out of these 14 genes, only 2 were also found in the 13-gene signature identified by Luminari et al. [14], indicating a discrepancy between the two gene sets and confirming the difficulty of the task of identifying a concise set of discriminative genes.

In order to provide a reference random baseline for the interpretation of classification results, we performed twenty iterations of training and validation of a random classifier in our dataset. It is the convergence value of these metrics that should be interpreted as the reference minimum value for this classification task, namely, a precision of 0.21, and a recall, specificity and AUC of approximately 0.5.

Considering these reference values, we present the results obtained following the previously described methodology using two settings. The first setting purposefully corresponds to an approach similar to the one followed by Luminari et al. [14], so that our results can be compared to the ones obtained in this work. We first apply the two configurations of feature selection introduced in the Methods section. The resulting preprocessed data is then subjected to internal bootstrapping for evaluation purposes. Our results are presented in Figure 3, where each color encodes a given performance metric and the horizontal lines correspond to the random classifier's results in said metric.

We can observe that the combination of our two feature selections steps, data balancing using SVM-SMOTE and the classifier SVM, can achieve a superior mean result of 0.97 AUC (against an AUC of 0.84). The high predictive power of an SVM in this setting is to be expected due to its recurrent good performance in this type of data [49, 50], and the fact that is being paired with the SVM-SMOTE data balancing technique and the SVM-RFE feature selection algorithm, both using an SVM as the base of its decisions.

The second setting guarantees the soundness of the acquired predictability levels by ensuring that feature selection step is performed inside each fold of the nested cross-validation, as previously described. The results obtained in this setting are presented in Figure 4. In this case, the XGBoost algorithm achieves significantly better results than the previous best performer SVM, with an AUC of 0.77, a precision of 0.67, a recall of 0.52 and a specificity of 0.94.

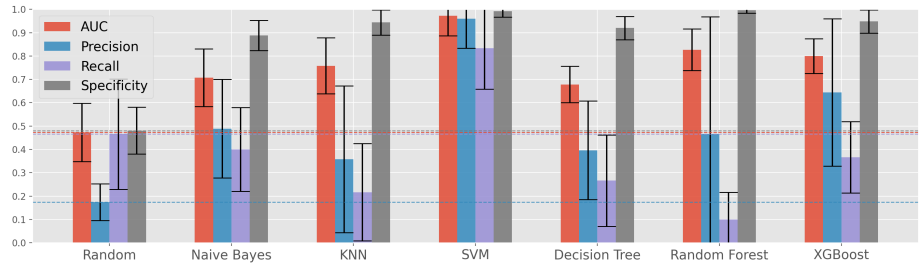

Figure 3: Optimized classifier's performance in the prediction of iPET2 (setting I)

In contrast, KNN has notably poor performance in comparison with the defined baseline, which can be explained by two major factors: i) a high percentage of

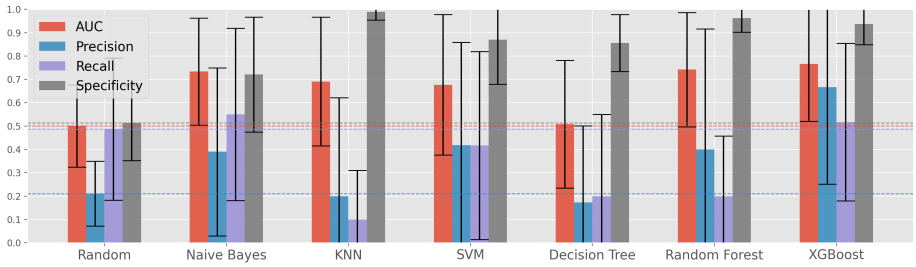

Figure 4: Optimized classifier's performance in the prediction of iPET2 (setting II)

outliers in our dataset; and ii) high gene expression variability, leading to inflated differences between individuals that belong to the same class. Two out of the three associative classifiers, Decision Tree (DT) and Random Forest (RF), also present comparable performance to the random classifier, suggesting that a decision tree is unable to fully take advantage of all the discriminative gene interactions present in the dataset. Since the RF and XGBoost classifiers are both ensembles of DTs, the discrepancy between their results must originate due to the embed feature engineering capabilities of XGBoost and differences on the pursued bagging and boosting strategies. Since bagging uses multiple independently trained DTs to make predictions, the fact that a DT cannot assimilate the knowledge in the data will lead to an overall lack in performance for the RF. Boosting on the other hand trains the trees iteratively, allowing for each consequent model to improve where the previous one failed, resulting in a classifier that can better model more complex interactions.

Overall, it is observable that all the classifiers can attain a high specificity, but at the cost of a reduced precision and recall. In other words, classifiers have an easier time correctly classifying patients with negative iPET2, guaranteeing that the patients that will react well to the ABVD regimen are correctly identified. The main difficulty with this predictive problem is in the correct classification of positive patients, possibly due to the low number of samples of this class. In addition, treatment response is being assessed with regards to iPET2 results, which may not be an optimal representative of the true quality of patient response to ABVD chemotherapy.

The predictive results can be further analyzed through the study of the classifiers' Precision-Recall curves presented in Figure 5. KNN is omitted from these graphics due to its dependence on an ineffective method to calculate the appropriate thresholds.

It is in our interest to better understand what leads a certain patient to be misclassified by our models. We plot some of the characteristics of the correctly classified individuals against wrongly classified ones by the best predictor in the second configuration, XGBoost. This analysis is provided for the four clinical variables, *age*, *gender*, *stage* and *LMR>2.1* in Figure 6. Starting with the variable *gender*, no significant trend is noted. *LMR>2.1* on the other hand shows a more clear inclination for correctly classifying positive cases when this variable is "False", with the percentage of True Positives (TP) being substantially higher than the False Positives (FP) percentage. Regarding the *stage* variable, the values "I A" and "III B" are

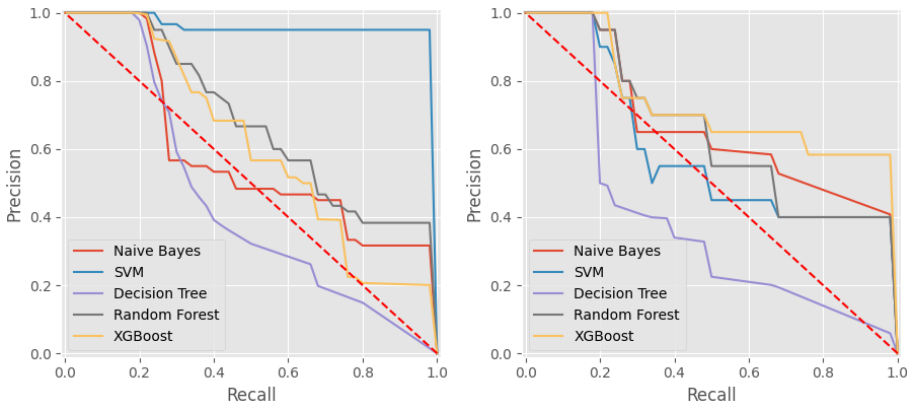

Figure 5: Precision-Recall curves for configuration one (left) and configuration two (right) for each machine learning model.

omitted due to the low number of samples corresponding to each one. In the observation of the remaining values only "III A" shows a significant deviation from the others, with all the positive cases correctly predicted but with a low performance in the false cases. The final plot is dedicated to the variable *age* and is presented in a stacked view, where the bins encompassing a 10-year period from each class (TN, FN, TP or FP) are stacked to facilitate a comparative analysis between them. We can then recognize that the majority of False Negatives (FN) occur in patients between 30 and 40 years, and the False Positives (FP) are more evenly distributed with a slightly higher concentration in patients between 20 to 30 years.

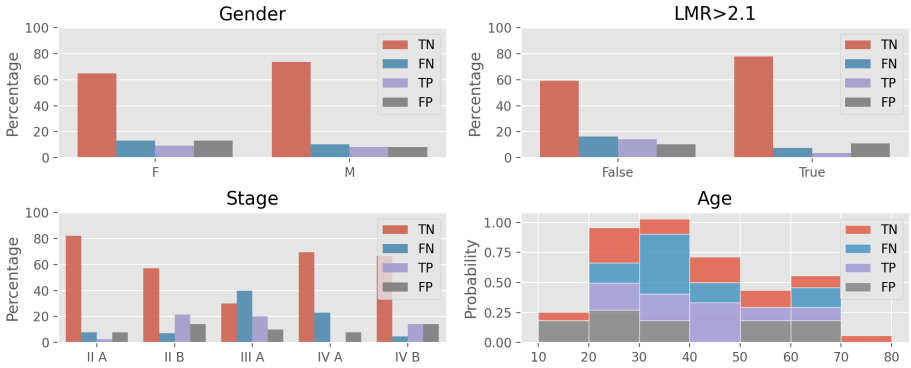

Figure 6: Distribution of predictions for the clinical variables

As some of our results indicate, the majority of useful information about how the patient will react to treatment is contained on gene expression features, and therefore it is imperative that the analysis of the factors inducing wrong classifications be extended towards these features. In Figure 7 we can see the distributions of the nine more discriminative genes according to their Mutual Information [24] with the response outcome. The red vertical lines highlight the expression associated with the six patients that were misclassified by all our predictive models.

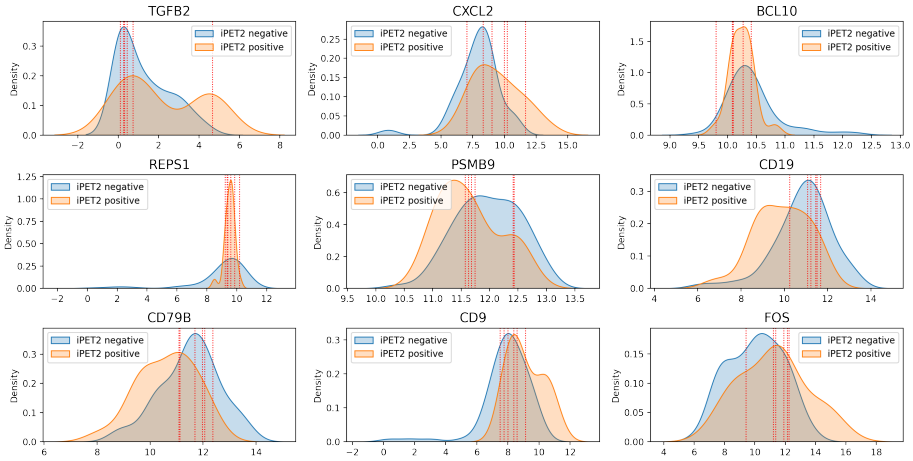

Figure 7: Distribution of top discriminative genes according to Mutual Information with highlighted values (red vertical lines)

We can then observe that, as expected, the majority of the highlighted values are found in the intersection of both distributions, where the classification is harder to perform.

Bicluster-based Space Transformation

The results presented until here correspond to the classification task in a feature space reduced by a composition of feature selection procedures. This approach lacks the ability to effectively represent the complex gene interactions responsible for the outcome of the patient. In accordance with the introduced methodology, we capture these interactions through discriminative biclusters using BicPAMS [46]. The found biclusters are then used to create new features representing discriminative and statistically significant gene expression patterns.

With the goal of better understanding the impact that BicPAMS' parameters have on predictive performance, we performed a comparative analysis of how distinct parameterizations affect the behavior of a baseline Naive Bayes predictor. The evaluated parameters are: i) *number of iterations*, indicating how many times the mining process is repeated, masking the found biclusters in each new iteration and forcing the mining process to find other less trivial biclusters but resulting in greater computational cost; ii) *minimum lift*, a placed threshold to determine whether a given bicluster is sufficiently discriminative [48]; iii) *number of labels*, corresponding to the number of overlapping gene expression levels [45]; and iv) *maximum number of biclusters*, the number of mapped features in the transformed data space by the postprocessing filtering of the bottom discriminative biclusters according to their lift. The results for each of these parameters are shown in Figure 8. We can see in the plotted results that in respect to most of the parameters there is a gain in performance by increasing its respective values, but only until a certain threshold is reached, from which there are no further advantages. The minimum lift is an exception to this, showing almost no effect in this specific problem and pointing to the importance of less discriminative patterns in the learning process.

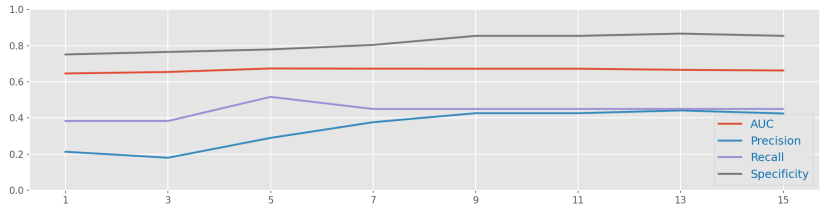

Figure 8a: Variation of performance according to the number of iterations

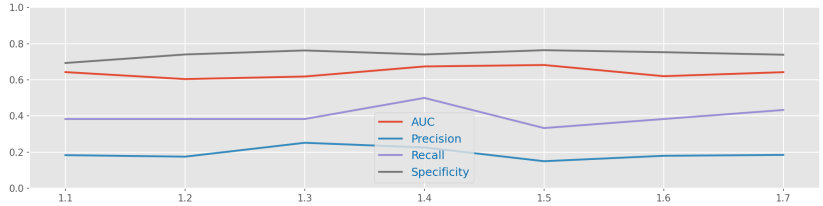

Figure 8b: Variation of performance according to the minimum lift

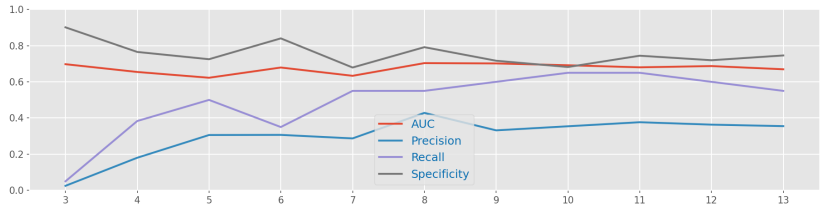

Figure 8c: Variation of performance according to the number of labels

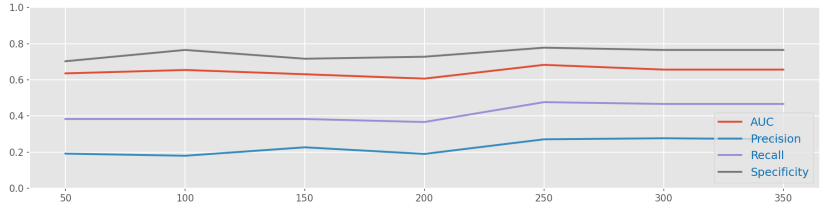

Figure 8d: Variation of performance according to the maximum number of biclusters

Finally, we further assessed the impact of different dissimilarity functions to assess the how likely is a given gene expression pattern for a specific patient, determining how the values present in the transformed dataset are computed. To this end, we consider both the Euclidean distance and a binary value indicating if a patient possesses a given pattern or not. In the second case, a tolerance threshold can be included to accommodate for noise. Figure 9 shows the performance of a Naive Bayes classifier in three different settings: the binary transformation with thresholds of 0.5 and 1, and the transformation using Euclidean distance. As expected, by using a numeric representation instead of a binary one, there is less loss of information, and consequently, we can achieve better results in the most difficult metrics for this classification task, precision and recall.

To assess the effects of the pattern-centric feature space mapping, we applied the target transformation using BicPAMS algorithm with the following parameters (placed according to previously gathered empirical evidence): number of iterations = 9; minimum lift = 1.3; number of labels = 10; maximum biclusters = 250; and distance criterion = Euclidean distance. Under this configuration, it is possible

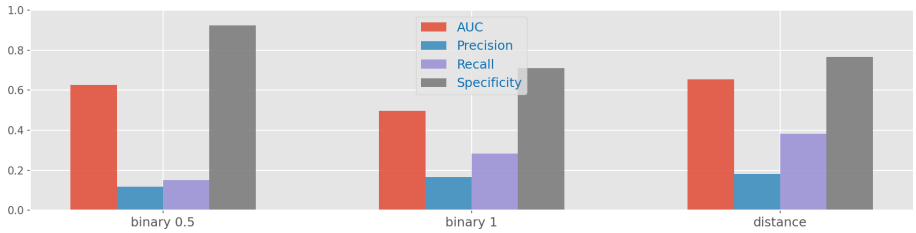

Figure 9: Variation of performance according to the distance criterion

to carry out a comprehensive exploration of the discriminative biclusters by performing multiple iterations while maintaining a relatively low minimum lift so that the computational requirements do not get too high. The high number of labels guarantees that the found patterns discriminate fine levels of expression while the relatively high number of biclusters ensures that most discriminative and statistically significant biclusters are retrieved. Figure 10 provides a direct comparison of each individual metric between these results and the ones previously obtained by our second setting (Figure 4).

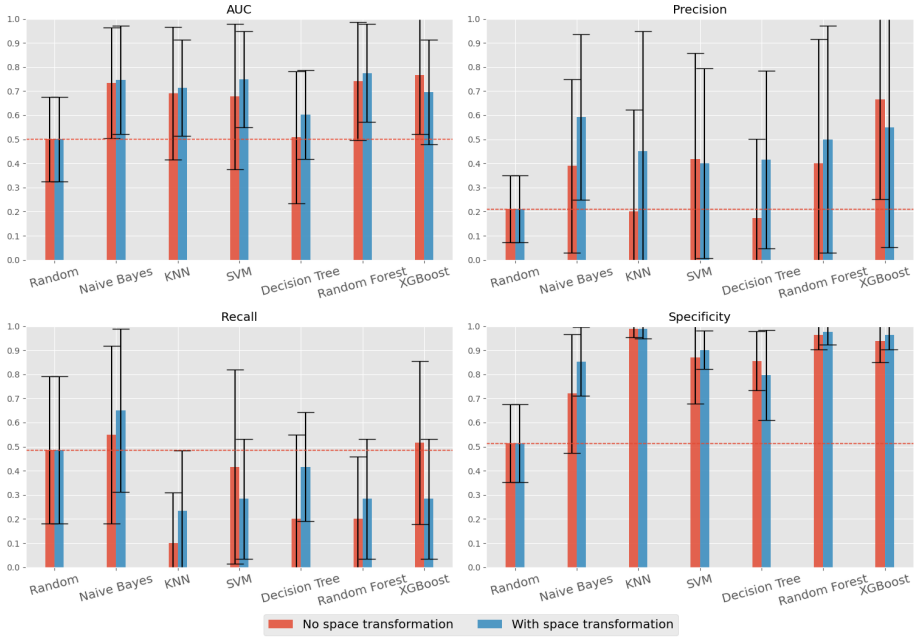

Figure 10: Direct comparison of results with (blue) and without (red) bicluster-based space transformation

The gathered results show that the transformed feature space has statistically significant impact on the behavior of the classifiers. SVM and XGBoost, classifiers that already presented good results, are not as significantly affected by this transformation, but all the others benefit from it and present an increase in performance in all the studied metrics.

Discussion

The ability to better discriminate how a patient will respond to a treatment is essential, especially in the domain of cancer therapy where the majority of treatments are associated with high toxicity and the prognostic exams can be intrusive and expensive. We comprehensively assess the predictability guarantees achieved by state-of-the-art ML models. These models are carefully optimized through the Tree Parzen Estimator algorithm and evaluated in a controlled manner resorting to nested cross-validation. Despite the placed optimization principles, the obtained results still fall short on the predictability power necessary to translate decisions in real-world practice. Transcriptional and iPET2 activity are structurally different, with the former being better positioned to model regulatory responses to treatment, even at the cost of iPET2 discordance.

The high specificity attained by most models (0.94) indicates that the classification models can correctly identify most of the patients that show disease regression after the treatment, but are more susceptible to recognize the positive patients. One possible reason for this bias is the low percentage of positive cases, representing only 20% of the total patients. To correct for this imbalance, we resorted to the use of balancing with SVM-SMOTE and hyperparameterization of the predictive models according to F1 score, an evaluation metric sensitive to the positive samples. Despite these efforts, the best classifier achieved a precision of 0.67 and a recall of 0.52, confirming the impact the difficulty of finding a transcriptional exam resembling the nature of iPET2 activity.

The nature of the target cohort data further introduces generalization challenges to the target prognostication, with low number of samples and high-dimensionality. In addition, the transcriptome profiling is susceptible to the infiltration of non-cancer cells and arbitrarily-high variations to the composition of the target cell population, further contributing to generalization difficulties. Finally, the biological mechanisms underlying diseases such as HL are immensely complex and dependent of interactions between many genes at multiple omics levels.

In order to better represent the complex interactions between genes that originate biological processes, the proposed space transformation offers an elegant way of shifting the learning from individual genes towards patterns of gene expression. By doing this, we group statistically significant and discriminative sets of genes that partake in regulatory modules correlated with a specific outcome of interest. This creates more straightforward conditions to guide the learning of predictive models. We observed that the efficacy of this transformation is proven by the increase in performance of the majority of classifiers in all the studied metrics.

Conclusion

This work introduces a novel methodology to improve the predictive accuracy of HL treatment response after two courses of ABVD chemotherapy against reference predictive levels [14]. This is achieved through a biclustering-based data space transformation that creates a shift from gene-centric to pattern-centric organization of expression data, combined with the thorough optimization of preprocessing procedures and state-of-the-art ML models.

Despite the yield improvements, the gathered results are indicative of the innate difficulty of the target predictive task, claiming for further contributions in this

domain able to translate high-dimensional regulatory profiles into actionable and reliable results.

In order to deal with this challenge, we suggest the following directions of research: i) strengthen this methodology by completing the current regulatory stances with complementary omic layers; ii) further combine the transcription of non-coding RNAs, recently shown to play an important role in HL [51]; iii) assess the potential increase in performance by using more specialized classification principles best suited to deal with the inherent overlapping class-conditional distributions of expression per gene (Figure 7); iv) place a finer description on the quality of treatment response, translating the classification task into an (ordinal) regression task; and v) further assess the impact of alternative pattern-based feature space transformations on predictive accuracy, namely by resorting to ensembles of biclusters with different characteristics.

Competing interests
The authors declare that they have no competing interests.

Funding
This work was supported by Fundação para a Ciência e a Tecnologia (FCT), through IDMEC, under LAETA project (UIDB/50022/2020), IPOscore with reference (DSAIPA/DS/0042/2018) and ILU (DSAIPA/DS/0111/2018). This work was further supported by the Associate Laboratory for Green Chemistry (LAQV), financed by national funds from FCT/MCTES (UIDB/50006/2020 and UIDP/50006/2020), INESC-ID plurianual (UIDB/50021/2020) and the contract CEECIND/01399/2017 to RSC.

Authors contributions
All authors contributed to the design of the methodology. AP implemented the methodology, conducted the experimental analysis, and produced the first draft of the manuscript. RH and RSC validated the work and revised the manuscript. All authors declare that they read and approved the final manuscript.

Acknowledgements
Not applicable.

Author details
[1]IDMEC, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais 1, 1049-001, Lisbon, Portugal, Portugal. [2]INESC-ID, Lisboa, Portugal. [3]Instituto Superior Técnico, Universidade de Lisboa, Portugal. [4]LAQV-REQUIMTE, DQ, NOVA School of Science and Technology, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.

References

1. Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.
2. Berthe MP Aleman, Alexandra W van den Belt-Dusebout, Willem J Klokman, Mars B van't Veer, Harry Bartelink, and Flora E van Leeuwen. Long-term cause-specific mortality of patients treated for hodgkin's lymphoma. *Optimizing treatment of patients with Hodgkin's lymphoma*, 21:95, 2007.
3. Berthe MP Aleman, Alexandra W van den Belt-Dusebout, Marie L De Bruin, Mars B van't Veer, Margreet HA Baaijens, Jan Paul de Boer, Augustinus AM Hart, Willem J Klokman, Marianne A Kuenen, Gabey M Ouwens, et al. Late cardiotoxicity after treatment for hodgkin lymphoma. *Blood*, 109(5):1878–1886, 2007.
4. Graça M Dores, Catherine Metayer, Rochelle E Curtis, Charles F Lynch, E Aileen Clarke, Bengt Glimelius, Hans Storm, Eero Pukkala, Flora E Van Leeuwen, Eric J Holowaty, et al. Second malignant neoplasms among long-term survivors of hodgkin's disease: a population-based evaluation over 25 years. *Journal of clinical oncology*, 20(16):3484–3494, 2002.
5. RT Hoppe. Hodgkin's disease: complications of therapy and excess mortality. *Annals of oncology*, 8: S115–S118, 1997.
6. Dirk Hasenclever, Volker Diehl, James O Armitage, David Assouline, Magnus Björkholm, Ercole Brusamolino, George P Canellos, Patrice Carde, Derek Crowther, David Cunningham, et al. A prognostic score for advanced hodgkin's disease. *New England Journal of Medicine*, 339(21):1506–1514, 1998.
7. Moccia A Moccia, Jane Donaldson, Mukesh Chhanabhai, Paul Hoskins, Richard Klasa, Kerry J Savage, Tamara Shenkier, Brian Skinnider, Randy D Gascoyne, Joseph M Connors, et al. The international prognostic factor project score (ips) in advanced stage hodgkin lymphoma has limited utility in patients treated in the modern era., 2009.
8. Paul Verboom, Harm van Tinteren, Otto S Hoekstra, Egbert F Smit, Jan HAM Van Den Bergh, Ad JM Schreurs, Roland ALM Stallaert, Piet CM van Velthoven, Emile FI Comans, Fred W Diepenhorst, et al. Cost-effectiveness of fdg-pet in staging non-small cell lung cancer: the plus study. *European journal of nuclear medicine and molecular imaging*, 30(11):1444–1449, 2003.

9. N Howlader, AM Noone, M Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, et al. Seer cancer statistics review, 1975–2017. *National Cancer Institute*, 2020.
10. Elisabeth Devilard, François Bertucci, Pascal Trempat, Reda Bouabdallah, Béatrice Loriod, Aurélia Giaconia, Pierre Brousset, Samuel Granjeaud, Catherine Nguyen, Daniel Birnbaum, et al. Gene expression profiling defines molecular subtypes of classical hodgkin's disease. *Oncogene*, 21(19):3095–3102, 2002.
11. Ralf Küppers, Ulf Klein, Ines Schwering, Verena Distler, Andreas Bräuninger, Giorgio Cattoretti, Yuhai Tu, Gustavo A Stolovitzky, Andrea Califano, Martin-Leo Hansmann, et al. Identification of hodgkin and reed-sternberg cell-specific genes by gene expression profiling. *The Journal of clinical investigation*, 111(4): 529–537, 2003.
12. David W Scott, Fong Chun Chan, Fangxin Hong, Sanja Rogic, King L Tan, Barbara Meissner, Susana Ben-Neriah, Merrill Boyle, Robert Kridel, Adele Telenius, et al. Gene expression–based model using formalin-fixed paraffin-embedded biopsies predicts overall survival in advanced-stage classical hodgkin lymphoma. *Journal of Clinical Oncology*, 31(6):692, 2013.
13. Christian Steidl, Tang Lee, Sohrab P Shah, Pedro Farinha, Guangming Han, Tarun Nayar, Allen Delaney, Steven J Jones, Javeed Iqbal, Dennis D Weisenburger, et al. Tumor-associated macrophages and survival in classic hodgkin's lymphoma. *New England Journal of Medicine*, 362(10):875–885, 2010.
14. Stefano Luminari, Benedetta Donati, Massimiliano Casali, Riccardo Valli, Raffaella Santi, Benedetta Puccini, Sofya Kovalchuk, Alessia Ruffini, Angelo Fama, Valentina Berti, et al. A gene expression–based model to predict metabolic response after two courses of abvd in hodgkin lymphoma patients. *Clinical Cancer Research*, 26(2):373–383, 2020.
15. Gary K Geiss, Roger E Bumgarner, Brian Birditt, Timothy Dahl, Naeem Dowidar, Dwayne L Dunaway, H Perry Fell, Sean Ferree, Renee D George, Tammy Grogan, et al. Direct multiplexed measurement of gene expression with color-coded probe pairs. *Nature biotechnology*, 26(3):317–325, 2008.
16. Bruce D Cheson, Richard I Fisher, Sally F Barrington, Franco Cavalli, Lawrence H Schwartz, Emanuele Zucca, and T Andrew Lister. Recommendations for initial evaluation, staging, and response assessment of hodgkin and non-hodgkin lymphoma: the lugano classification. *Journal of clinical oncology*, 32(27):3059, 2014.
17. Michel Meignan, Andrea Gallamini, Michel Meignan, Andrea Gallamini, and Corinne Haioun. Report on the first international workshop on interim-pet scan in lymphoma. *Leukemia & lymphoma*, 50(8):1257–1260, 2009.
18. Guillaume Lemaître, Fernando Nogueira, and Christos K. Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017. URL `http://jmlr.org/papers/v18/16-365.html`.
19. Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, 3(1):4–21, 2011.
20. Yvan Saeys, Iñaki Inza, and Pedro Larrañaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007.
21. Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. .
22. Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.
23. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
24. Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.
25. Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3):389–422, 2002.
26. Enrico Glaab, Jaume Bacardit, Jonathan M Garibaldi, and Natalio Krasnogor. Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PloS one*, 7(7):e39932, 2012.
27. Gaurav Pandey, Om P Pandey, Angela J Rogers, Mehmet E Ahsen, Gabriel E Hoffman, Benjamin A Raby, Scott T Weiss, Eric E Schadt, and Supinda Bunyavanich. A nasal brush-based classifier of asthma identified by machine learning analysis of nasal rna sequence data. *Scientific reports*, 8(1):1–15, 2018.
28. Likai Wang, Yanpeng Xi, Sibum Sung, and Hong Qiao. Rna-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC genomics*, 19(1):546, 2018.
29. Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.
30. Mehdi Pirooznia, Jack Y Yang, Mary Qu Yang, and Youping Deng. A comparative study of different machine learning methods on microarray gene expression data. *BMC genomics*, 9(1):1–13, 2008.
31. Maxim D Podolsky, Anton A Barchuk, Vladimir I Kuznetcov, Natalia F Gusarova, Vadim S Gaidukov, and Segrey A Tarakanov. Evaluation of machine learning algorithm utilization for lung cancer classification based on gene expression levels. *Asian Pacific Journal of Cancer Prevention*, 17(2):835–838, 2016.
32. Reinel Tabares-Soto, Simon Orozco-Arias, Victor Romero-Cano, Vanesa Segovia Bucheli, José Luis Rodríguez-Sotelo, and Cristian Felipe Jiménez-Varón. A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data. *PeerJ Computer Science*, 6:e270, 2020.
33. Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

34. Evelyn Fix. *Discriminatory analysis: nonparametric discrimination, consistency properties*. USAF School of Aviation Medicine, 1951.
35. Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
36. Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.
37. J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.
38. David D Lewis. Naive (bayes) at forty: The independence assumption in information retrieval. In *European conference on machine learning*, pages 4–15. Springer, 1998.
39. James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In *25th annual conference on neural information processing systems (NIPS 2011)*, volume 24. Neural Information Processing Systems Foundation, 2011.
40. André Patrício, Rafael S Costa, Rui Henriques, et al. Predictability of covid-19 hospitalizations, intensive care unit admissions, and respiratory assistance in portugal: Longitudinal cohort study. *Journal of Medical Internet Research*, 23(4):e26075, 2021.
41. Yizong Cheng and George M Church. Biclustering of expression data. In *Ismb*, volume 8, pages 93–103, 2000.
42. Robert B Bentham, Kevin Bryson, and Gyorgy Szabadkai. Mcbiclust: a novel algorithm to discover large-scale functionally related gene sets from massive transcriptomics data collections. *Nucleic acids research*, 45(15): 8712–8730, 2017.
43. Andrew Williams and Sabina Halappanavar. Application of biclustering of gene expression data and gene set enrichment analysis methods to identify potentially disease causing nanomaterials. *Beilstein journal of nanotechnology*, 6(1):2438–2448, 2015.
44. Rui Henriques, Claudia Antunes, and Sara C Madeira. A structured view on pattern mining-based biclustering. *Pattern Recognition*, 48(12):3941–3958, 2015.
45. Rui Henriques and Sara C Madeira. Bicpam: Pattern-based biclustering for biomedical data analysis. *Algorithms for Molecular Biology*, 9(1):1–30, 2014.
46. Rui Henriques, Francisco L Ferreira, and Sara C Madeira. Bicpams: software for biological data analysis with pattern-based biclustering. *BMC bioinformatics*, 18(1):1–16, 2017.
47. Rui Henriques and Sara C Madeira. Bsig: evaluating the statistical significance of biclustering solutions. *Data Mining and Knowledge Discovery*, 32(1):124–161, 2018.
48. Leonardo Alexandre, Rafael S Costa, Lucio Lara Santos, and Rui Henriques. Mining pre-surgical patterns able to discriminate post-surgical outcomes in the oncological domain. *IEEE Journal of Biomedical and Health Informatics*, 2021.
49. Michael PS Brown, William Noble Grundy, David Lin, Nello Cristianini, Charles Walsh Sugnet, Terrence S Furey, Manuel Ares, and David Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences*, 97(1):262–267, 2000.
50. Yoonkyung Lee and Cheol-Koo Lee. Classification of multiple cancer types by multicategory support vector machines using gene expression data. *Bioinformatics*, 19(9):1132–1139, 2003.
51. Anna Cordeiro, Mariano Monzó, and Alfons Navarro. Non-coding rnas in hodgkin lymphoma. *International journal of molecular sciences*, 18(6):1154, 2017.