EMBnet.journal
Bioinformatics in Action

Sabbatini et al. (2023) EMBnet.journal 28, e1035
http://dx.doi.org/10.14806/ej.28.0.1035

Research Papers

On potential limitations of differential expression analysis with non-linear machine learning models

Gianmarco Sabbatini, Lorenzo Manganaro✉

aizoOn Technology Consulting, Torino, Italy
Competing interests: GS none; LM none

Abstract

Recently, there has been a growing interest in bioinformatics toward the adoption of increasingly complex machine learning models for the analysis of next-generation sequencing data with the goal of disease subtyping (*i.e.*, patient stratification based on molecular features) or risk-based classification for specific endpoints, such as survival. With gene-expression data, a common approach consists in characterising the emerging groups by exploiting a differential expression analysis, which selects relevant gene sets coupled with pathway enrichment analysis, providing an insight into the underlying biological processes. However, when non-linear machine learning models are involved, differential expression analysis could be limiting since patient groupings identified by the model could be based on a set of genes that are hidden to differential expression due to its linear nature, affecting subsequent biological characterisation and validation. The aim of this study is to provide a proof-of-concept example demonstrating such a limitation. Moreover, we suggest that this issue could be overcome by the adoption of the innovative paradigm of eXplainable Artificial Intelligence, which consists in building an additional explainer to get a trustworthy interpretation of the model outputs and building a reliable set of genes characterising each group, preserving also non-linear relations, to be used for downstream analysis and validation.

Introduction

In recent years, high-throughput technologies for molecular data, such as next-generation sequencing (NGS) are getting increasingly cheaper (van Nimwegen *et al.*, 2016) and their use to improve our understanding of complex pathologies, such as cancer, is becoming widespread. This gives rise to an incredible amount of multi-omics data (genomics, transcriptomics, epigenomics, etc.), which can be analysed and exploited in the context of personalised medicine.

One of the main goals of these analyses is disease subtyping, meaning identifying a molecular-based stratification of patients affected by the same pathology, which ideally relates to prognosis. To this end, there is a growing interest in the adoption of state-of-the-art machine learning (ML) algorithms and models. These already proven excellent performances in almost any other field of application due to their ability to catch highly non-linear relations and patterns emerging from the dataset, which seems ideal when studying the biology of such a complex system as cancer (Zhang *et al.*, 2019; Tang *et al.*, 2019).

A very common approach to gene expression-based disease subtyping (see, *e.g.*, Su *et al.*, 2014) is to use a clustering model (an unsupervised ML model) based on molecular data to divide patients into groups, and then validating such grouping from a biological perspective by means of the so-called "downstream analysis". This latter typically consists in performing a differential expression (DE) analysis (Costa *et al.*, 2017; Soneson and Delorenzi, 2013) between the emerging groups to identify a set of genes which are considered determinants to discriminate between subtypes, and then exploiting those genes to perform a gene-set enrichment analysis (GSEA), unveiling the underlying biological processes characterising the emerging subtypes, and laboratory validation, whenever possible. Such an approach is of course valid, especially if a simple and linear clustering model has been used. However, when complex non-linear ML models are involved, DE-based characterisation has potential limitations that may affect the biological interpretability of results and that, to the best of our knowledge, has not been reported so far. In particular, when patient grouping is based on a non-linear relationship with a feature (*i.e.*, gene), such feature may be hidden to DE analysis, leading to a failure of the subsequent downstream analysis and validation and, overall, an incomplete biological characterisation.

Article history
Received: 21 September 2022
Accepted: 10 January 2023
Published: 08 March 2023

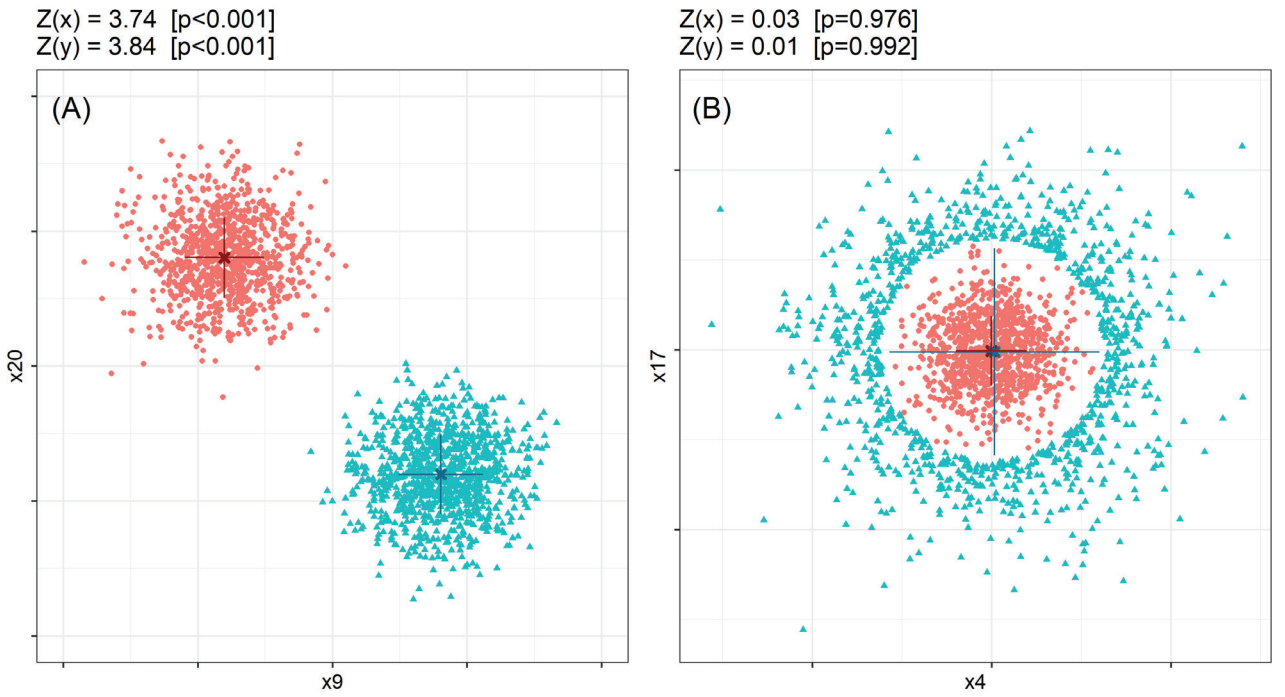

Figure 1. 2D representation of the synthetic datasets A (left) and B (right), where x and y axes represent the 2 relevant variables of each dataset respectively, namely: (A) clouds: x9, x20; (B) circles: x4, x17. Pattern of dataset (C) is not reported to avoid redundancy (similar to panel B). Colours and point shapes help in visualising the 2 emerging groups of each dataset. Thick crosses with error bars represent the mean and standard deviation of the 2 groups. Z(x) and Z(y) in the panel headers report the results of Normal Z tests between the mean values along x and y axes of each plot, respectively, with associated p-values.

Beyond unsupervised disease subtyping, such limitation of DE analysis in general still holds for any problem involving biological characterisation of different groups identified by exploiting non-linear ML models, such as the classification of high/low-risk groups for a specific end-point based on gene expression data (see, *e.g.*, Choi *et al.*, 2020).

In the subset of cases where the model developed provides the possibility to assign labels to new data points (*e.g.*, any supervised classifier or any unsupervised model that creates a partition of the feature space, such as kMeans-based models), we encourage the adoption of the innovative eXplainable Artificial Intelligence (XAI) paradigm (Arrieta *et al.*, 2020). It consists in building an additional explainer model to get a trustworthy interpretation of the model outputs as an alternative to DE analysis to build a reliable set of determinant features (*i.e.*, genes) characterising each group, preserving also non-linear relationships.

The aim of this short paper is to provide a minimal proof-of-concept example of the above-described limitation, which has never been reported so far, suggesting and highlighting the strength of adopting XAI-based alternatives. This is done by considering three well-motivated synthetic datasets (see Discussion), each consisting of two groups, where features mimic the values of a gene-expression matrix, and some of them result in linear or non-linear relationships with groups. Firstly, we analysed the groups with DE analysis, showing that only those features corresponding to a linear separation between groups emerge as significative; secondly, we used simple models distinguishing between the two groups to apply XAI-based explanations and proving that, in this case, also features having non-linear relations with groups are detected as relevant.

Methods

Datasets

We built three synthetic datasets, named (A) clouds, (B) circles and (C) circles (big), respectively. Datasets A and B (see Figure 1) are made of 20 variables (x1, …, x20 – also referred to as features), mimicking the values of a gene expression matrix where each feature represents a gene. Out of the 20 variables, 18 are built as pure noise, sampling the values from uniform distributions. The remaining two variables for each dataset are instead "significative", allowing to distinguish well-separated groups. "Significative" variables have been set randomly by drawing two numbers between 1 and 20 with uniform probability for each dataset. C is similar to the others in that it is characterised by two "significative" variables, but it is meant to prove that the number of variables involved and correlations between them are actually irrelevant to the issue considered. As such, it is made of additional 800 noisy variables, plus 198 other variables with linear correlations with the others (correlation coefficients have been computed and found to be variable up to 0.99). Overall, the third dataset is made of 1000 variables with a signal-to-noise ratio of 0.002.

Each of the three datasets consists of 2000 points, equally divided into two groups of 1000 that are meant

Sabbatini et al. (2023) EMBnet.journal 28, e1035
http://dx.doi.org/10.14806/ej.28.0.1035

to represent two subtypes or classes of interest within the dataset, in the case of clustering and classification models, respectively. Considering the two significant variables, the groups are generated as follows:

(A) Clouds. Bivariate Normal distributions with different centres and same standard deviation in both directions. This dataset is exemplary of a linear relation between significant features and groups. From a biological perspective, it represents, for example, a couple of genes that are over- and under-expressed in the two groups, respectively.

(B) Circles. The inner cloud is sampled from a bivariate Normal distribution, whereas the outer cloud is sampled from a Gamma distribution with an offset on the radial coordinate. This dataset is exemplary of a purely non-linear relation between significant features and groups. From a biological perspective, it represents, for example, a couple of genes whose expression has to be kept in homeostasis for health conditions, and a disbalance of expression levels causes the behaviour of interest.

(C) Circles (big). The two significant variables are sampled in the same way described for dataset B, thus defining a similar circular pattern that has not been shown in the figure to avoid redundancy. The difference with respect to dataset B lies in the number of noisy variables and the presence of correlations between variables, as previously described.

Moreover, we built a supplementary dataset D (see Supplementary Materials[1]), similar to dataset C but with an increased number of significant variables and synthetic expression values sampled from negative binomial distributions mimicking those of a real RNA-seq dataset (see Supplementary Figure 3[1]). This supplementary dataset is meant to show that the number of significant variables and the underlying distributions are not affecting the results hereafter presented.

Differential expression analysis

For each of the three datasets, we carried out a differential expression analysis between groups. Computations have been performed using two different algorithms, namely DESeq (Love *et al.*, 2014), implemented in the R package DESeq2, which uses shrinkage estimation for dispersions and fold change estimates, and GLM (McCarthy *et al.*, 2012), implemented in the R package edgeR, which applies a kernel transformation to the feature space before regression. Genes are considered significantly differentially expressed in the case of an adjusted p-value below 0.05.

Machine Learning models

For each of the three datasets, we built a simple model distinguishing between the two groups and providing the labels reported in Figure 1 (*i.e.*, tagging "red" and "blue" samples). In particular, for dataset A (clouds) we used a linear model implementing a decision boundary lying on the significant variable plane and perpendicular to a line passing through the centres of mass of the two clouds. For dataset B (circles) and C (circles – big), we used a non-linear model implementing a circular decision boundary lying on the significant variable planes, respectively, centred on the centre of mass of the inner cloud. Implementation is public and available on GitLab (see Data & code availability section).

XAI-based explanation analysis

We applied an XAI-based approach, training two different explainers for each dataset to interpret the model's output. To this end, we used both LIME (Ribeiro *et al.*, 2016) and kernel-SHAP (Lundberg and Lee, 2017), which are the two most popular explainers for models built on tabular data, and they are both local explainers, meaning that they provide local explanations for each point of the dataset. In particular, the explainers result in a value associated with the importance of each feature in explaining the output of each point. Intuitively, this is done by: **(i)** creating a neighbourhood of the data point to be explained. Such process is different between the two explainers since LIME applies a Gaussian perturbation to the point, whereas kernel-SHAP substitutes some of the values of the features with those sampled from a background set provided as input (we used 20 random points of the dataset). **(ii)** Assigning labels to the neighbours applying the model and; **(iii)** fitting a linear model on the whole neighbourhood that represents a local linear approximation of the global decision boundary between groups in the neighbourhood of the data point to be explained. The weights of the linear approximation are used to assign a local importance score to each feature. The overall importance of the features can be obtained either by averaging the contributions over all the points of the datasets or by considering the median value of the distribution. It should be noted that absolute importance values provided by LIME and kernel-SHAP are not directly comparable.

Data and code availability

The data generated to support the presented findings, as well as the code used for data generation, data analysis and plots, are publicly available on GitLab at: https://gitlab.com/deflect-public/differential-expression/.

Results

The variables randomly selected as significant were x9 and x20 for dataset A, x4 and x17 for dataset B and x23 and x83 for dataset C. Results of the DE analysis for all three datasets are shown in Figure 2. From the adjusted p-value, it results that DE analysis (both DESeq and GLM) is effective in identifying important features in the case of dataset A (clouds), in which x9 and x20 are characterised by $p<0.001$ and noisy features are rejected. Instead, it fails when non-linearity is introduced, *i.e.*, with datasets B (circles) and C (circles – big) in which no significative features are detected.

[1]http://journal.embnet.org/index.php/embnetjournal/article/downloadSuppFile/1035/1035_supp_1

Sabbatini et al. (2023) EMBnet.journal 28, e1035
http://dx.doi.org/10.14806/ej.28.0.1035

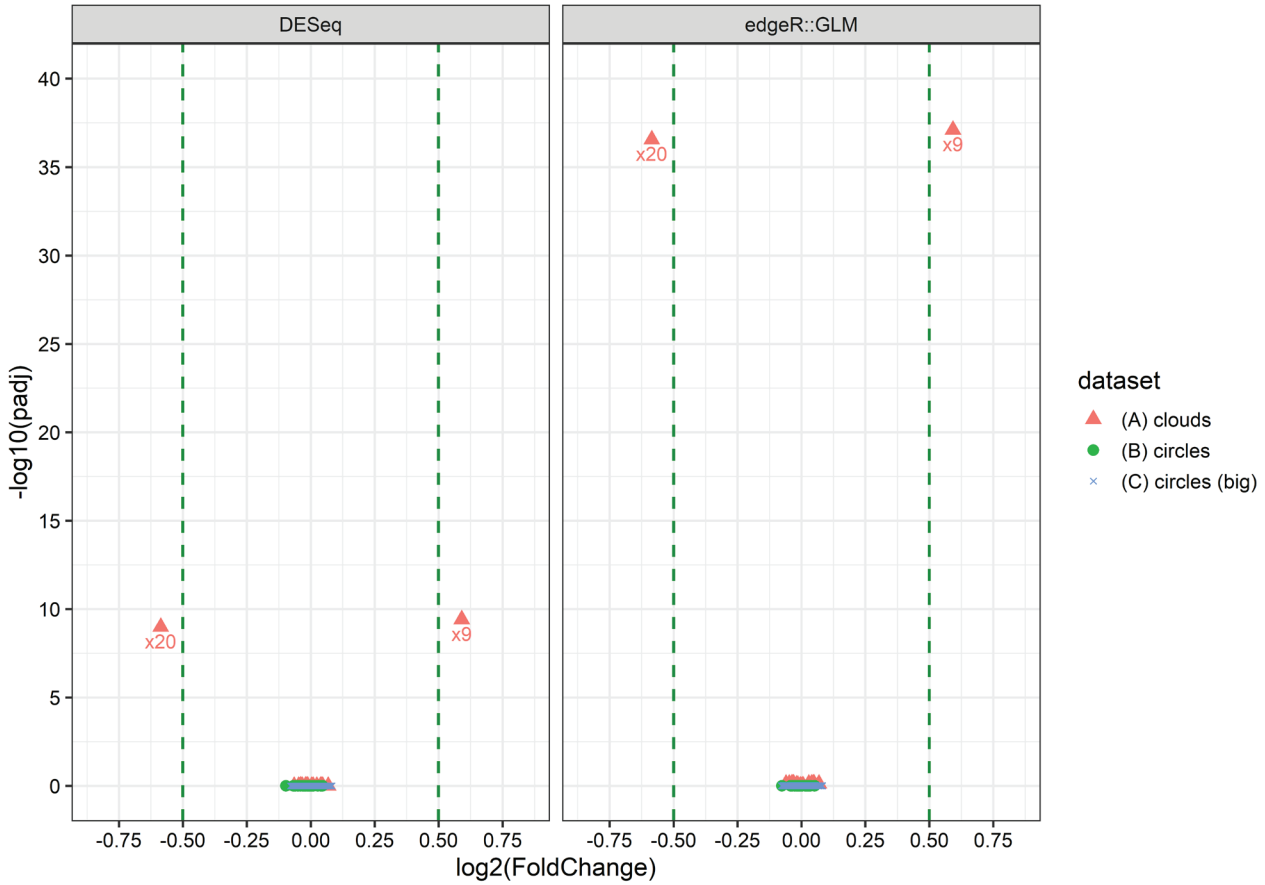

Figure 2. Volcano plot representing the results of differential expression analysis for the three datasets (different colours and shapes) obtained with DESeq (left) and edgeR-GLM (right), respectively. Vertical dashed lines represent significance thresholds. Adjusted p-values have been computed using the Benjamini-Hochberg correction.

Results of XAI-based feature importance analysis are summarised in Figure 3 for both tested explainers, namely LIME and kernel-SHAP. As expected, both the explainers proved to be effective in highlighting the relevant features for all the datasets considered by assigning importance scores way above those of noisy variables.

In the chosen examples, LIME seems to result in distributions that are a bit better separated from noisy variables with respect to kernel-SHAP. On the other hand, kernel-SHAP seems to be more efficient in recognising noisy variables whose contribution is set to zero, whereas LIME typically assigns negligible but non-zero contributions to those variables.

Discussion

The present study focuses on a general issue arising any time ML models are applied to gene expression data to stratify patients based on their molecular profile. In particular, the clustering model or classifier assigns labels to the samples, and researchers have to understand whether the resulting grouping is biologically meaningful or not. This latter process, often referred to as "downstream analysis", involves many further analyses such as GSEA (or pathway analysis) and wet lab validation, and it is possibly followed by clinical validation if the results are considered significant and robust enough. All these analyses, however, rely on the common issue of identifying the subset of genes that have been determinant for the model in assigning the labels or, in other words, the set of genes characterising the groups to be used for subsequent pathway analysis and validation. Such characterisation is very often performed with DE analysis between groups; however, notably, the above-presented findings highlight an intrinsic limitation of DE analysis. In particular, it is very effective if the groups under consideration are (nearly) linearly separable, whereas it fails in identifying relevant features when non-linearity is introduced. This does not mean that DE analysis is wrong, nor it is the intention of the authors to make criticisms of specific previous literature, but that its use as a group characterisation and gene selection method for downstream analysis should be considered with extreme care if coupled with non-linear ML models. Such an aspect, to the best of our knowledge, has not been addressed so far in the literature.

The major implication of this DE limitation is that features that are determinant for the model to distinguish

Sabbatini et al. (2023) EMBnet.journal 28, e1035
http://dx.doi.org/10.14806/ej.28.0.1035

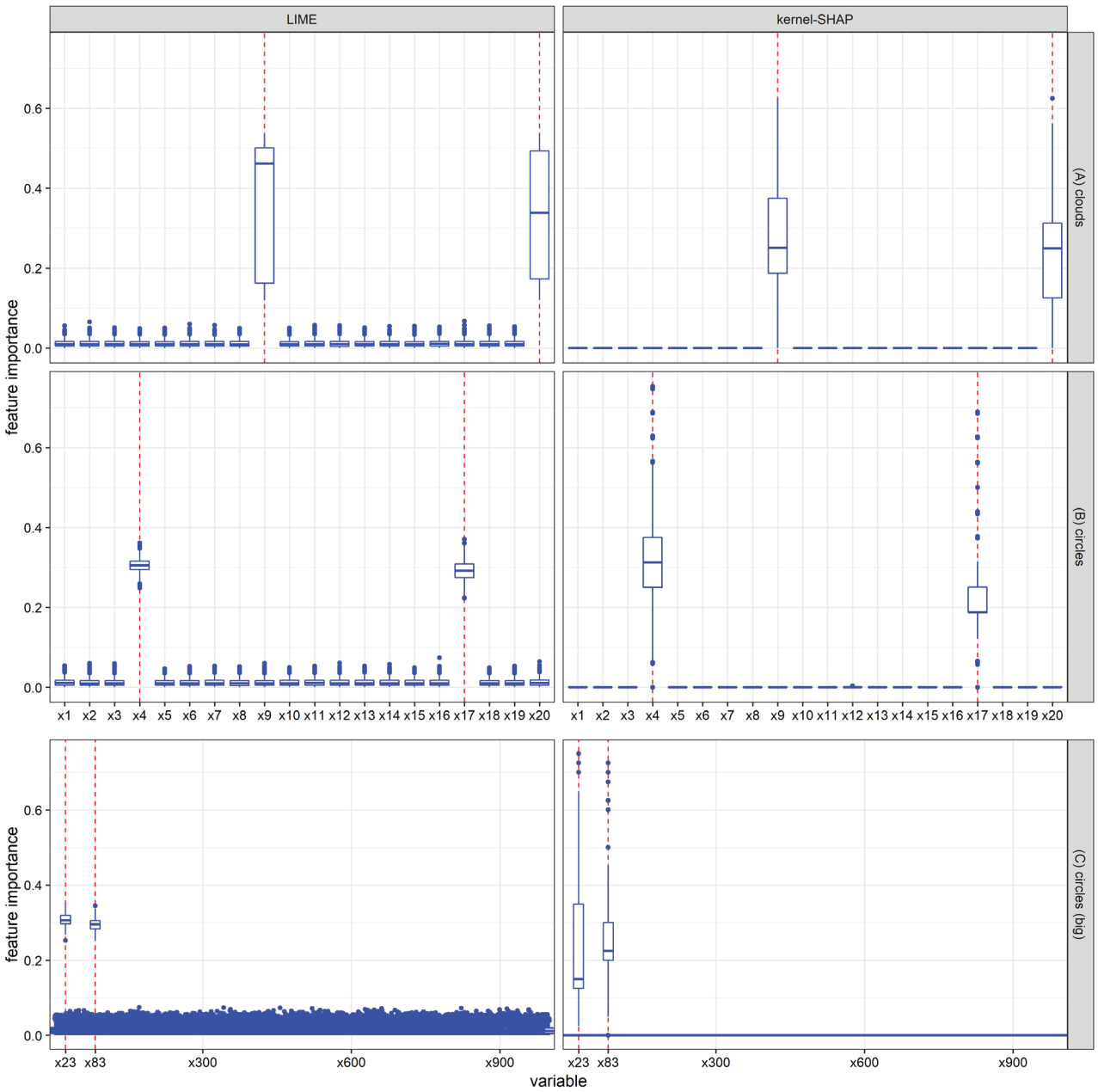

Figure 3. Feature importance of the three datasets (rows) computed using two different explainers, namely LIME (left column) and kernel-SHAP (right column). Boxplots report feature importance based on local explanations of each sample. Vertical dashed lines indicate the relevant features of each dataset.

between the two groups could potentially pass undetected. This has the consequence of affecting the list of genes used for lab validation or pathway analysis, thus potentially compromising the significance of such validation techniques. In other words, if the hypothesis developed is good, it may be that relevant biological characteristics of the groups under consideration evade the attention of researchers and limit the interpretability of results or, worse, that the hypothesis is rejected because results are erroneously considered not biologically meaningful.

Concerning the model, we would like to insist that the considerations made throughout the paper are valid both for classifiers and clustering models. In fact, for our purpose, the model might be considered as a black-box tool that, given the input data, provides labels as output, establishing a linear or non-linear relationship between inputs and output. Instead, we propose that the focus should lie on methods that allow to understand which features were actually relevant in determining such labels. As a consequence, details on classification accuracy or model training are not provided since they are not pertinent for clustering models and, overall, irrelevant to the results and conclusion presented.

As a complement to DE analysis, we propose to adopt (whenever possible) an XAI-based approach, which we

demonstrated to be able to overcome limitations related to non-linearity. On the other hand, such an approach has some drawbacks that should be considered. First, it is not always applicable, especially because it requires that the AI model selected can be used to predict labels of new data points. But, this is only the case of supervised models and the subset of unsupervised models that create a partition of the feature space, such as kMeans-based models. For the other cases, DE analysis still remains the only option. Second, research on XAI models is a cutting-edge topic in the field of ML applications, and it is still in its early years. Thus, little guidance exists in order to help researchers in choosing the best configuration of parameters to get reliable estimations of feature importance from the explainers, which at this point requires tuning of hyperparameters and their combinations. In this direction, it is worth mentioning the work of (Amparore *et al.*, 2021) in providing reliable metrics to quantify the quality of XAI explanations, which may be helpful in guiding hyperparameter tuning. Finally, while DE analysis provides p-values associated with fold change estimates, a major limitation of the XAI-based approach is that it only provides a number whose absolute value is associated with feature importance. As a result, relevant and not relevant features have to be defined by means of a threshold, typically applied to the average importance value, which may not be straightforward to set either since there is not a general analytical rule. A possible method would be to look for gaps or "knees" in the ordered feature importance plot.

A further element that is worth discussing is our choice of simulated data instead of real data, which is generally preferred in methodological studies related to gene expression. Indeed, in this case, the choice is well motivated by several considerations. First, simulated data allow us to control which genes are relevant for classification and to verify if they are actually detected by the approaches considered without any dependencies on the biological interpretation of results, which instead would not have been possible with a real dataset. Plus, many biological interpretations are based on existing methodological results and thus subject to their shortcomings: using them would have resulted in a self-feeding vicious circle. Second, it is true that real datasets are characterised by many genes interacting with each other, but we show with dataset C that the number of variables and correlations between features does not affect the methods considered, apart from increasing the computing time, and that the conclusions derived from dataset A and B, still hold for dataset C. In fact, results for dataset C, characterised by numerous and correlated features, are equivalent to those obtained with dataset B (see Figures 2 and 3), characterised by few features without correlations. Moreover, with supplementary dataset D[1], we also show that the number of significant variables or the underlying distribution of the synthetic expression values do not affect the results. Third, and probably most importantly, we highlighted how DE limitations arise in the case of non-linear relations between gene expression data and the resulting groups. It is important to note that such grouping is the output of the model (*i.e.*, labels), so that non-linearities are introduced by the model itself and do not necessarily coincide with the underlying "ground truth". In other words, what is actually relevant is not the structure of the dataset, but rather the shape of the decision boundary defined by the model. If the model results in a non-linear decision boundary, which is likely to happen when complex ML models are used, DE analysis may not be effective in identifying the relevant variables for group characterisation. Conversely, as we have shown, the XAI-based approach is better suited, independently of the dataset. Finally, we stress that the focus of this study is purely methodological; thus, although a real dataset would have been illustrative of a full downstream analysis leading to a biological interpretation based on the relevant features (genes) detected, in this case, we are just assessing the capability of each method to detect those relevant features under different conditions.

It should be noted that while the present study focuses on differential expression, the underlying idea has a broader application, and there are no conceptual limitations in extending it to similar analyses under the same assumptions, such as the case of group characterisation based on differentially methylated genes (see, *e.g.*, Kolbe *et al.*, 2014).

As a final remark, we would like to point out that, with this study, we are not presenting an original XAI model and that there may be better XAI-based approaches to use as a complement to DE analysis, depending on the specific application. Nonetheless, the limitation of DE analysis that we are highlighting still holds and should be considered when characterising the biology of groups identified with non-linear ML models.

Conclusions

In conclusion, in the present study we identified a potential limitation of DE analysis used as a gene selection method for subsequent enrichment analysis and lab validation when patient grouping is obtained with the application of complex non-linear ML models. We provided a proof-of-concept example of such limitation by exploiting three synthetic datasets. To overcome the issue, we suggest using XAI-based alternatives that can be effective on the cases considered.

Funding

This study has been conducted as part of a large scientific project called DEFLeCT (Digital tEchnology For Lung Cancer Treatment) funded by Regione Piemonte (P.O.R. F.E.S.R. 2014/2020 technological platform "Health and Wellness").

Acknowledgements

We would like to thank Sushant Parab and Davide Corà for the fruitful discussions and the whole aizoOn health

R&D group, particularly Paolo Falco and Selene Bianco, for their support.

Key Points

- DE has limited capability to detect non-linear relationships between features and target.
- DE limitation becomes relevant when coupled with complex (non-linear) ML clustering or classification models.
- For the subset of ML models that can predict labels for new data points, DE limitation can be overcome by applying XAI to interpret ML output and detect the relevant features.

References

Amparore E, Perotti A and Bajardi P (2021) To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. PeerJ Comput Sci **7**, e479. http://dx.doi.org/10.7717/peerj-cs.479

Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S *et al.* (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion **58**, 82–115. http://dx.doi.org/10.1016/j.inffus.2019.12.012

Choi Y, Qu J, Wu S, Hao Y, Zhang J *et al.* (2020) Improving lung cancer risk stratification leveraging whole transcriptome RNA sequencing and machine learning across multiple cohorts. BMC Medical Genom **13**(10), 1–15. http://dx.doi.org/10.1186/s12920-020-00782-1

Costa-Silva J, Domingues D and Lopes FM (2017) RNA-Seq differential expression analysis: An extended review and a software tool. PLoS ONE **12**(12), e0190152. http://dx.doi.org/10.1371/journal.pone.0190152

Kolbe DL, DeLoia JA, Porter-Gill P, Strange M, Petrykowska HM *et al.* (2012) Differential analysis of ovarian and endometrial cancers identifies a methylator phenotype. PLoS ONE **7**(3), e32941. http://dx.doi.org/10.1371/journal.pone.0032941

Love MI, Huber W and Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol **15**(12), 1–21. http://dx.doi.org/10.1186/s13059-014-0550-8

Lundberg SM and Lee SI (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst **30**. http://dx.doi.org/10.1145/2939672.2939778

McCarthy DJ, Chen Y and Smyth GK (2012) Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res **40**(10), 4288–97. http://dx.doi.org/10.1093/nar/gks042

Ribeiro MT, Singh S and Guestrin C (2016) "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. http://dx.doi.org/10.1145/2939672.2939778

Soneson C and Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics **14**(91). http://dx.doi.org/10.1186/1471-2105-14-91

Su X, Malouf GG, Chen Y, Zhang J, Yao H *et al.* (2014) Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. Oncotarget **5**(20). http://dx.doi.org/10.18632/oncotarget.2454

Tang B, Pan Z, Yin K and Khateeb A (2019) Recent advances of deep learning in bioinformatics and computational biology. Front Genet **10**, 214. http://dx.doi.org/10.3389/fgene.2019.00214

van Nimwegen KJ, van Soest RA, Veltman JA, Nelen MR, van der Wilt GJ *et al.* (2016) Is the $1000 genome as near as we think? A cost analysis of next-generation sequencing. Clin Chem **62**(11), 1458–1464. http://dx.doi.org/10.1373/clinchem.2016.258632

Zhang Z, Zhao Y, Liao X, Shi W, Li K *et al.* (2019) Deep learning in omics: a survey and guideline. Brief Funct Genomics **18**(1), 41–57. http://dx.doi.org/10.1093/bfgp/ely030