Computer Methods and Programs in Biomedicine 240 (2023) 107719

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Explainable artificial intelligence to predict and identify prostate cancer tissue by gene expression

Alberto Ramírez-Mena[a], Eduardo Andrés-León[b], Maria Jesus Alvarez-Cubero[a,c], Augusto Anguita-Ruiz[d], Luis Javier Martinez-Gonzalez[a,*], Jesus Alcala-Fdez[e]

[a] GENYO, Centre for Genomics and Oncological Research: Pfizer -University of Granada - Andalusian Regional Government, Granada, 18016, Spain
[b] Institute of Parasitology and Biomedicine "López-Neyra" (IPBLN), Spanish National Research Council (CSIC), Granada, 18016, Spain
[c] Department of Biochemistry and Molecular Biology III and Immunology, University of Granada, Granada, 18071, Spain
[d] Barcelona Institute for Global Health, ISGlobal, Barcelona, 08003, Spain
[e] Department of Computer Science and Artificial Intelligence, Andalusian Research Institute in Data Science and Computational Intelligence (DaSCI), University of Granada, Granada, 18071, Spain

ARTICLE INFO

Article history:
Received 21 March 2023
Revised 16 June 2023
Accepted 8 July 2023

Keywords:
Explainable artificial intelligence
Molecular biology
Machine learning
Prostate cancer
Clinical decision support
Biomedical informatics

ABSTRACT

Background and Objective: Prostate cancer is one of the most prevalent forms of cancer in men worldwide. Traditional screening strategies such as serum PSA levels, which are not necessarily cancer-specific, or digital rectal exams, which are often inconclusive, are still the screening methods used for the disease. Some studies have focused on identifying biomarkers of the disease but none have been reported for diagnosis in routine clinical practice and few studies have provided tools to assist the pathologist in the decision-making process when analyzing prostate tissue. Therefore, a classifier is proposed to predict the occurrence of PCa that provides physicians with accurate predictions and understandable explanations.

Methods: A selection of 47 genes was made based on differential expression between PCa and normal tissue, GO gene ontology as well as the literature to be used as input predictors for different machine learning methods based on eXplainable Artificial Intelligence. These methods were trained using different class-balancing strategies to build accurate classifiers using gene expression data from 550 samples from 'The Cancer Genome Atlas'. Our model was validated in four external cohorts with different ancestries, totaling 463 samples. In addition, a set of SHapley Additive exPlanations was provided to help clinicians understand the underlying reasons for each decision.

Results: An in-depth analysis showed that the Random Forest algorithm combined with majority class downsampling was the best performing approach with robust statistical significance. Our method achieved an average sensitivity and specificity of 0.90 and 0.8 with an AUC of 0.84 across all databases. The relevance of *DLX1, MYL9* and *FGFR* genes for PCa screening was demonstrated in addition to the important role of novel genes such as *CAV2* and *MYLK*.

Conclusions: This model has shown good performance in 4 independent external cohorts of different ancestries and the explanations provided are consistent with each other and with the literature, opening a horizon for its application in clinical practice. In the near future, these genes, in combination with our model, could be applied to liquid biopsy to improve PCa screening.

1. Introduction

Prostate cancer (PCa) is one of the most common cancers among men worldwide; in Europe it is the most frequent and the third leading cause of cancer mortality [1]. However, early detection through the use of widespread screening has enabled a significant shift from metastatic to localized disease at the time of

* Corresponding author.
E-mail addresses: alberto.ramirez@genyo.es (A. Ramírez-Mena), eduardo.andres@csic.es (E. Andrés-León), mjesusac@ugr.es (M.J. Alvarez-Cubero), augusto.anguita@isglobal.org (A. Anguita-Ruiz), luisjavier.martinez@genyo.es (L.J. Martinez-Gonzalez), jalcala@decsai.ugr.es (J. Alcala-Fdez).

https://doi.org/10.1016/j.cmpb.2023.107719

initial diagnosis, which is critical for the treatment of the disease and for reducing disease-specific mortality [2]. Screening strategies in PCa are usually focused on the use of serum PSA levels, the combination of different anatomic and functional magnetic resonance imaging (MRI) sequences guided by transrectal ultrasound (TRUS), and digital rectal exam (DRE). However, the serum PSA level is organ-specific but not necessarily cancer-specific and can be elevated for a variety of reasons. Moreover, diagnostic accuracy with MRI is highly dependent on the training and expertise of the radiologist, thus urging the use of objective items in Imaging-Reporting and Data System guidelines (PI-RADS) [3–5]. It should be noted that the invasive nature of DRE and TRUS-guided biopsy increases diagnosis success but causes significant morbidity (pain, fever, bleeding, infection, transient urinary difficulties, or other complications requiring hospitalization).

Many research strategies are focusing on the analysis of extracellular vesicles (which are one source of interesting biomarkers) [6], free miRNAs [7], or, as is the case with other tumors, gene-specific markers such as circulating mRNA molecules [8]. A number of genetic susceptibility markers for PCa have also been identified using different approaches such as family-based studies, candidate gene association studies, genome-wide association studies and RNA-Seq technology. However, due to the heterogeneity of PCa, only a few of these markers have been robustly associated with PCa in contrast with other prevalent tumors, such as breast cancer, in which there is a clear relationship between the alteration of several genes and an increased risk of developing the tumor.

All genetic susceptibility markers for PCa (e.g., *BRCA2, BRCA1, MSH2* among others) are implicated in tumor development or are biomarkers of increased risk for hereditary PCa, but there is no gene reported for PCa diagnosis or screening [9]. Therefore, the identification of new biomarkers in early stages of the disease that allow a better distinction and classification of PCa remains a challenge for researchers. Single-cell analysis will improve detection and prediction models that can assist clinicians in diagnosing or detecting the disease earlier and more accurately, acting as PCa clinical decision support systems (CDSS) to aid decision-making at the primary care level. Previous reports indicate that genomic single-cell analysis could be a good source of quantitative parameters for neoplastic growth and aggressiveness in PCa [10], but its success will depend on the CDSS available to the pathologist when it comes to tissue diagnosis.

Machine Learning (ML) techniques have proven effective in improving the prediction and diagnosis of PCa, due to their capacity to provide automatically descriptive or predictive models from huge amounts of data that can be used to build the inference engines of a CDSS, which can serve as a helping hand for medical practitioners to diagnose or detect the disease earlier and more accurately [11,12]. However, clinicians often do not rely on the latest technical approaches and methodologies (e.g., Deep Learning and Random Forest) despite their high accuracy, because they lack reliable interpretability and explainability of the obtained models, as their predictions cannot be explained in a manner understandable to humans (known as the black box problem [13]). Medical experts do not trust decisions provided by black-box models without comprehensive and easy-to understand explanations because in many cases “how they predict” is more important than “what they predict” [14]. As a result, the ML techniques employed in the clinical domain normally consider simpler and interpretable models (e.g., linear models and rule-induction algorithms) at the expense of accuracy. Many studies have tried to open the black box of complex models and provide an explanation of their decisions, since these models should be considered as CDSS and therefore should be provided with elements that allow them to be properly audited, increasing the practitioner’s confidence. This is especially important in scenarios such as this one, where the decisions of AI-based systems may impact people’s lives, as defined in the “Ethics Guidelines for Trustworthy AI”[1] published by the European Union for the use of Trustworthy AI. A whole field of research, Explainable Artificial Intelligence (XAI), is concerned with studying it further to better understand the system’s underlying mechanisms and to find solutions. XAI recommends the use of transparent models that are self-explanatory, and of post-hoc explainability techniques that aim to reveal understandable information about how a complex model produces its predictions for any given input [15]. Thus, the aim is to improve the use of Artificial Intelligence in an ethical, transparent, fair and responsible manner [16].

Thanks to genomic widespread and the possibility of identifying RNA in fresh or paraffin-embedded tissue, our aim is to develop an accurate and comprehensive model that allows us to predict the appearance of PCa from genes implicated in the proliferation of this type of tumor. To this end, we have performed several differential gene expression analyses comparing PCa-affected tissue versus healthy prostate tissue of PCa patients selected from the TCGA-PRAD (The Cancer Genome Atlas-Prostate Adenocarcinoma) [17] dataset to finally identify 47 genes that may act as biomarkers for PCa. This procedure constituted an expert-knowledge guided feature selection technique that reduced the search space to a subgroup of biomarkers associated with the outcome to be predicted in several populations. The well-known tree ensemble, Random Forest (RF) [18], one of the most accurate ML models currently in use, has been used to generate an accurate predictive ML model from the selected genes. Breiman stated that RF is an outstanding predictor for performance, but fails in interpretability because of the Herculean task of unraveling the complex web formed by the majority vote of more than a hundred trees, which is consistent with other authors suggesting that post-hoc explanatory techniques are needed to understand its behavior [19]. We have used SHapley Additive exPlanations [20] (SHAP) to better understand the underlying mechanisms of the model and analyze the influence of each gene on the prediction. They are useful for explaining various supervised learning ML models by providing an importance value to each input feature for each prediction made, improving the transparency and reliability of a model by understanding the root causes underlying each prediction. The obtained model provides medical practitioners with a trade-off between accuracy and explainability.

With the intention of evaluating the performance of our approach for using novel biomarkers in the prediction of PCa, we have conducted the following studies. First, we compared our proposal with other ML approaches in PCa patients selected from the TCGA-PRAD dataset making use of a stratified repeated 5-fold cross-validation to assess the performance of our model. Different classification metrics and non-parametric statistical tests were used to evaluate the performance of the algorithms analyzed. Second, to validate and contrast the knowledge derived from the TCGA dataset, the generated model has been applied to four independent external cohorts (GSE22260, GTEX, GSE183019 and GSE114740). Finally, a functional enrichment analysis was performed to interpretate the role of the selected genes within the context of PCa. A web page associated with this paper (https://sci2s.ugr.es/PCaXAIRF) has been developed with supplementary material for this study.

2. Material and methods

2.1. General overview of the analysis plan

The main goal of this paper is to construct an accurate and comprehensible classifier capable of predicting the risk of devel-

[1] https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

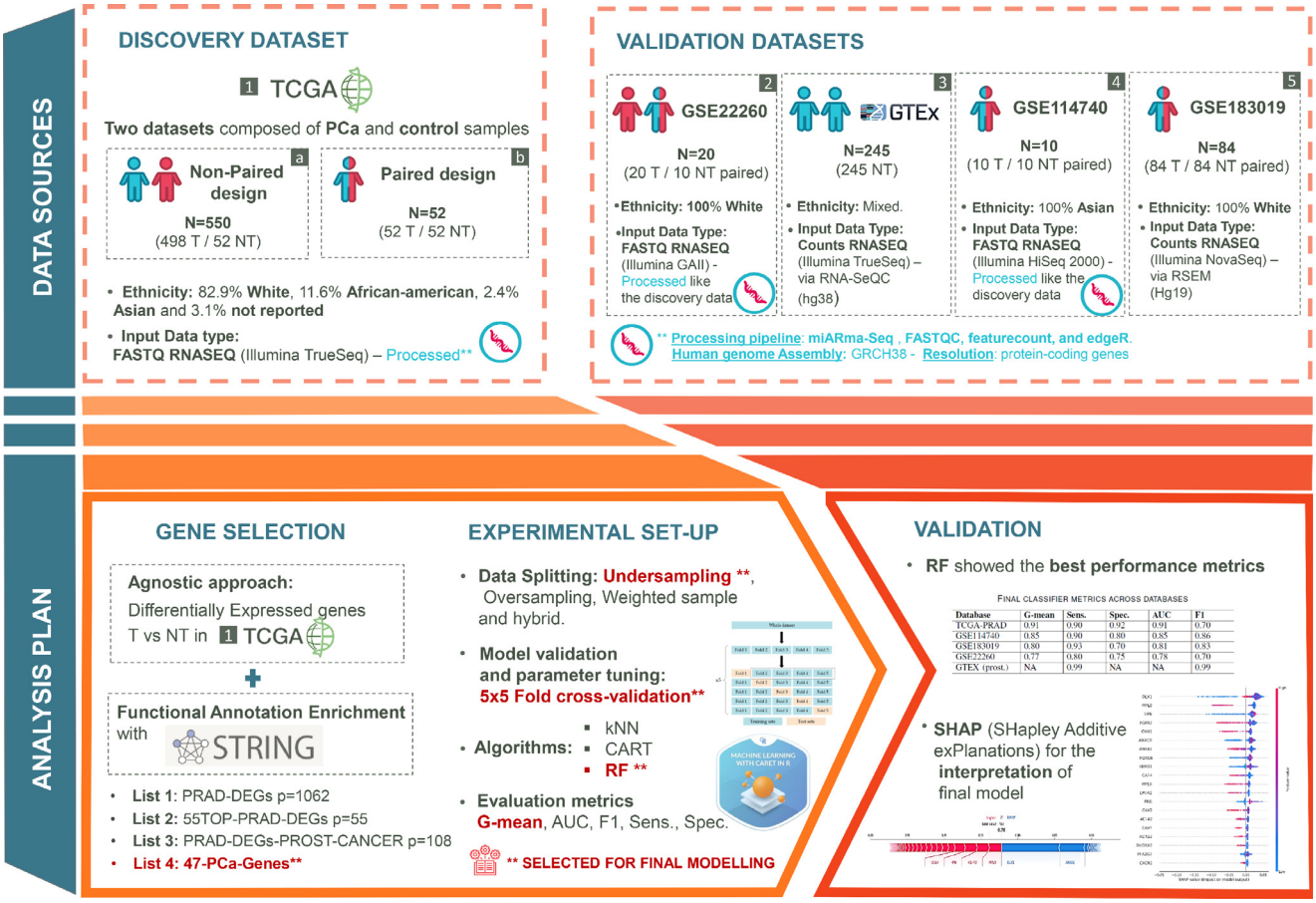

Fig. 1. Graphical abstract. **Data sources:** Datasets used in this work as discovery and validation populations. Each dataset includes information about the number of samples, whether a matched design exists, ethnicity, available data formats and the analysis pipeline applied to the raw data before validation. **Analysis plan:** Gene selection strategy, including the different gene sets generated and tested in this study. The experimental design section includes information about the class balancing strategies applied, model validation and parameter tuning approach, models tested and metrics evaluated for these models (the best-performing strategy is highlighted). **Validation:** Includes information about the best-performing model, results for validation datasets and an overview of the SHAP explanatory analysis.

oping PCa based on gene expression data derived from prostate tissue (RNAseq data/Transrectal biopsy). To this end, we conducted a three-step analytical approach consisting of; 1) Filtering and selection of genetic features, according to their biological relevance in PCa, 2) Construction and analysis of the performance of several ML predictive methods using the genetic features selected in the previous step in the discovery population, and 3) Selection of the best prediction model and validation of that model on a number of independent cohorts, with different ancestries (see Fig. S1 in the supplementary material). All datasets employed in this work correspond to publicly available data in open online repositories. ML models employed in the clinical field do not usually consider complex models because of the ethical implications of making decisions that affect patients' lives without a mechanism to understand black-box models that are usually responsible for these sophisticated techniques. Instead, they often rely on interpretable models at the cost of accuracy [13]. Since these models do not always provide the necessary accuracy, an alternative could be to open the black box and understand the mechanisms behind it.

In the following we will provide explanations and some insights into the drivers of our model from two points of view, including visualization and feature importance based on SHAP. The use of biological-function enrichment analysis and expert knowledge from molecular biologists was of great importance throughout the analytical process, motivating important decisions at the feature-selection stage and helping in the interpretation of the final results. Fig. 1 shows a general overview of this study, including the most relevant details about the data sources used as discovery and validation populations at the top, the gene selection approach used to filter the most biologically relevant genes in PCa and the experimental setup used (data splitting strategies, model validation methodology, methods and evaluation metrics) at the lower-left area, and an overview of the validation results and the post-hoc explanation approach at the lower-right area.

2.2. Data sources overview

The discovery population employed here corresponded to a public dataset from the well-known TCGA consortium. In particular, we focused on the TCGA-PRAD data that was available through the GDC Legacy Archive at the National Cancer Institute (NIH). TCGA-PRAD was chosen as our discovery population because of the availability of raw expression data for download, the diversity of tumour development stages of patients, the presence of paired samples and the presence of patients with different ancestries. This dataset contains data for 550 prostate biopsies, 498 tumoral samples (T) and 52 controls (NT) with International Society of Urological Pathology (ISUP) grade levels ranging from 1 to 5, including different cancer stages. Notably, each control presented a matched tumor sample belonging to the same patient. Therefore, we will refer to the TCGA-PRAD population as two subgroups of samples: paired samples (52 healthy tissue vs 52 tumoral tissue), and non-paired samples (498 tumoral samples vs 52 control samples). After obtaining access to these data from NIH, RNAseq data were avail-

Table 1
Data sources used in this study .

Data ID	Description	Ancestry	Sequencing Platform	Data format
TCGA	Discovery population. 498T vs 52NT (52 paired). ISUP grades: 1 (46 samples), 2 (146 samples), 3 (101 samples), 4 (64 samples), 5 (141 samples).	82.9% white, 11.6% black or African-American, 2.4% Asian and 3.1% not reported	Illumina TrueSeq RNA sequencing	FASTQ RNA-SEQ (controlled access)
GSE22260	Validation dataset. 20T vs 10NT (10 paired) 20 prostate cancer tumors and 10 matched normal tissues. Gleason grades: 6 (7 samples), 7 (11 samples), 8 (1 sample), 9 (1 sample).	Caucasian	Illumina GAII platform	FASTQ RNA-SEQ
GTEX	Validation dataset. 245NT. The Genotype-Tissue Expression (GTEx) project stores samples from 54 non-diseased tissue sites across nearly 1000 individuals.	84.6% white, 12.9% African-American, 1.3% Asian, 1.1% not reported	Illumina TrueSeq RNA sequencing	Read counts obtained via RNA-SeQC. FASTQ files were aligned to hg38.
GSE183019	Validation dataset. 84T vs 84NT (paired). mRNA profiles of 84 pairs of localized primary prostate cancer samples and the matched normal tissues were generated by deep sequencing. ISUP grades: 1 (8 samples), 2 (54 samples), 3 (18 samples), 4 (1 samples), 5 (3 samples).	White	Illumina NovaSeq 6000.	Read counts obtained via RSEM. FASTQ files were aligned to hg19.
GSE114740	Validation dataset. 10T vs 10NT (paired). mRNA profiles of 10 pairs of localized primary prostate cancer samples and the matched normal tissues were generated by deep sequencing.	Chinese/Han	Illumina HiSeq 2000	FASTQ RNA-SEQ.

able in the form of FASTQ files for each TCGA sample. The ethnicity of the participants was classified as: 82.9% white, 11.6% black or African-American, 2.4% Asian and 3.1% not reported. To validate our final model, we used independent external populations, corresponding to public datasets, with different ancestries. These populations included GSE22260 [21], GTEX (RNA-seq data from healthy prostate tissue), GSE183019 and GSE114740. The datasets beginning with the GSE identifier corresponded to paired sample studies accessed via the Gene Expression Omnibus database [22], and belong to PCa patients. The GTEx (Genotype-Tissue Expression) dataset was obtained from the GTEx Portal, filtering only prostate tissue belonging to individuals not affected by PCa. These datasets were intended for the validation of the selected predictive models and corresponded to different ethnic groups. More details on each population can be found in Table 1.

2.3. Transcriptomic analysis

The cohorts included in this work were sequenced using different platforms, and data was available in a variety of formats. For TCGA-PRAD, GSE22260 and GSE114740 RNA-Seq FASTQ files were retrieved for each patient in order to run a standardized RNA-Seq pipeline for the purpose of extracting raw counts expression data for each sample. miARma-Seq [23] was used to carry out this process. First, the raw data was analyzed with FASTQC[2] in order to assess the quality of the reads. Then, the reads were mapped to the GRCh38 human genome using the star [24] aligner. Finally, read counts were calculated with featureCounts [25]. The counts matrix was then filtered, keeping only protein-coding genes and Counts Per Million (CPM) were calculated for each gene after normalizing all samples by using EdgeR's [26] TMM implementation. Normalized data was analyzed using hierarchical clustering and Principal Component Analysis (PCA) but all samples were kept, since no clear outliers were found. GTEX and GSE183019 datasets only had publicly available expression count matrices. These matrices were computed using different pipelines and, in the case of the GSE183019 dataset, the reads were aligned to a different version of the human genome: GRCh37. Finally, CPMs for each gene were scaled and centered, calculating their z-score according to Eq. 1, where z_j is the z-score for gene j, x_j is the original value for gene j, μ_j is the mean of gene j across samples and σ_j is its standard deviation. Due to the nature of RNA-seq data, expression levels can differ substantially between genes, so this procedure is crucial in order to harmonize the relevance of each gene before running the algorithms.

$$z_j = \frac{(x_j - \mu_j)}{\sigma_j} \tag{1}$$

2.4. Gene selection

RNA-seq count matrices are populated with thousands of genes per sample. Reducing the gene dimensionality is key, not only to discard genes lacking biological interest for the purpose of this study, but also to make sure that ML algorithms are computationally affordable avoiding the so-called "curse of dimensionality" [27,28]. As a first step in our strategy for selecting the best candidate genes to be used as predictors for the ML algorithms that would be run at a later stage, we ran two differential expression analyses using EdgeR. For the first study, only paired samples (52NT vs 52T) were considered, generalized linear models (GLM) were fitted to account for tissue type (T/NT) and patient effect, and finally quasi-likelihood F-tests were applied to find differentially expressed genes (DEGs). For the second study, all samples were considered (52NT vs 498T) and quantile-adjusted conditional maximum likelihood (qCML) was used because only the tissue type factor was accounted for. After fitting negative binomial models and obtaining dispersion estimates, genewise exact tests were employed to find DEGs. In both cases, the raw count matrices were normalized and those genes that were expressed at less than 1 CPM in more than 52 samples, which is the smallest group in both studies, were filtered out (52, NT) [29].

The final results were filtered again, keeping only those genes with a False Discovery Rate (FDR) lower than 0.05 and a |logFC| of at least 1. The final results showed 1991 DEGs for the whole set of samples and 1332 for the paired dataset. 1065 genes (PRAD-DEGs) were shared in both sets, representing 47.17% of the DEGs found. Since these 1065 genes were differentially expressed in both populations and were also found to be significant, we decided to keep them as an initial set of predictors for our work. Similarly, we also calculated the intersection of both sets of DEGs, but this time considering only the top 200 genes in each set according to their logFC value, obtaining a new gene set of interest made up

[2] https://www.bioinformatics.babraham.ac.uk/projects/fastqc

Table 2
Gene sets considered in this study .

Gene set	Description	Num. genes
PRAD-DEGs	Intersection of DEGs satisfying FDR<0.05 and |logFC| of at least 1, for paired and unpaired populations in TCGA-PRAD.	1065
55TOP-PRAD-DEGs	The same procedure was used with the with PRAD-DEGs gene set, but only the top 200 genes according to their |logFC| value were used for paired and unpaired populations before computing their intersection.	55
PRAD-DEGs-PROST-int-CANCER	Genes in the PRAD-DEGs gene set annotated with both "prostate" and "cancer".	108
47-PCa-Genes	Final gene set for our tool. Obtained by merging the gene sets 55TOP-PRAD-DEGs an PRAD-DEGs-PROST-int-CANCER, using STRING and k-means clustering and choosing the most representative gene in each cluster.	47

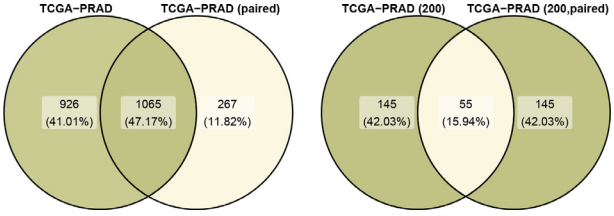

Fig. 2. Intersection of differentially expressed genes in TCGA paired samples and TCGA non-paired samples. On the left, all DEG satisfying FDR<0.05 and |logFC|>1 are considered, and their intersection is the PRAD-DEGs gene set consisting of 1065 genes. On the right, only the top 200 DEGs, according to their |logFC| are considered for each population and their intersection amounts to 55 genes: 55TOP-PRAD-DEGs gene set.

of 55 genes: 55TOP-PRAD-DEGs. Fig. 2 shows how these gene sets were obtained.

Next, we carried out a functional enrichment analysis on our PRAD-DEGs gene set using STRING [30] and looking for the annotations "cancer" and "prostate" in Biological Process (GO), Molecular Function (GO), Cellular Component (GO), KEGG, Tissue expression (TISSUES) and Reference publications (PUBMED). As a result of the enrichment analysis 114 genes were enriched with both the terms "cancer" and "prostate". We will refer to this gene set as PRAD-DEGs-PROST-int-CANCER.

In an effort to further reduce the number of genes of interest and extend our search space to other genes that may be relevant to our study, we merged the 55TOP-PRAD-DEGs and PRAD-DEGs-PROST-int-CANCER genesets, obtaining 157 genes. Subsequently, we classified them using the STRING tool, grouping them in 45 clusters using kmeans clustering. For each cluster, the gene with more connections was selected, assuming the hypothesis that each gene would provide the biological information of its cluster. When this key gene did not exist in a cluster, two were selected. Finally, we obtained 47 candidate genes for our tool: 47-PCa-Genes.

Table 2 shows the different gene sets considered as potential predictors of interest for this study.

2.5. Pre-processing and experimental set-up

In order to build our classifier to differentiate healthy prostate tissue from tumoral prostate tissue, we trained different models using the caret [31] (Classification and Regression Training) package available for R[3].

Since the TCGA-PRAD dataset is significantly imbalanced, 498 tumor samples vs 52 healthy samples, we applied different strategies to prevent the algorithms from being biased towards classifying samples as tumoral just for being the majority class, including undersampling, oversampling, weighted sample and hybrid approaches. The upsampling technique creates synthetic samples similar to the real ones belonging to the minority class; the undersampling approach reduces the number of samples in the majority class by removing some of them in the population; weighted sample technique aims to assign a weight to every sample depending on the class they belong to, so that weights are higher for the minority class, which would make the failures when predicting that class more significant. The last strategy consists in combining the upsampling and undersampling techniques to balance the population size for each class using procedures such as SMOTE [32]. Note that, prior to applying these techniques to our discovery population, training and testing partitions were created for each fold, and while the training partitions were affected by these procedures, the test data remained unaffected and the samples retained their original values and proportions.

We have used a stratified 5-fold cross-validation, repeated 5 times, to assess the performance, amounting to a total of 25 executions. This technique is appropriate in cases where the population size is limited, reducing estimation errors, and providing a good bias-variance tradeoff, apart from being a computationally efficient process [33]. This technique may lead to worse but more realistic results because the outcome of the algorithms is not influenced by the seed chosen when splitting the dataset.

The following methods were selected for our study: k-nearest neighbors [34] (KNN), rpart[4] (Classification and Regression Trees - CART) and RF [35]. KNN and CART techniques generate models that are understandable to experts, and are expected to provide sufficient information about the relationship between input features and predictions, while allowing clinicians to answer questions related to which genes are playing a key role in predictions. Breiman stated that RF [18] is an excellent alternative when performance is key, pointing out that it is an excellent predictor that fails in terms of interpretability, which makes it necessary to use post-hoc techniques to understand its behavior as described in the next subsection.

Different parameters were trained depending on the algorithm: *k* for KNN: the number of the instances closest to the query to be considered; *CP* (complexity parameter) for rpart: its role is pruning any split in the tree that doesn't improve the fit by at least *CP*; and *mtry* for RF: the number of predictors randomly selected as candidates at each tree split. Any other parameters for the previous algorithms have been set to their default values, as recommended by their authors, in an effort to facilitate comparisons and take advantage of using settings that perform well in the majority of cases instead of searching for very specific values. To evaluate the performance of the algorithms, several metrics that have been extensively described in the literature were used, complementing the information provided by each one: F1, G-mean, AUC, sensitivity and specificity [36].

Metrics that consider classes individually enable the analysis of that specific class. Performance measures, such as "G measures"

[3] https://www.R-project.org

[4] https://CRAN.R-project.org/package=rpart

that combine basic metrics, were designed to summarize different trade-offs between the individual performance measure for each class. G-mean is the geometric mean of sensibility and specificity, and this metric measures how balanced sensitivity and specificity are regardless of the majority/minority classes. This metric is intended to balance the success rate between the majority and minority classes: in our case, a low performance in predicting controls will imply a low value for the g-mean metric even if all cases with PCa are correctly classified. This measure helps us avoid overfitting the majority class while underfitting the minority group [37]. All the possible combinations using KNN, rpart and RF with the class balancing approaches discussed earlier and the dataset 47-PCa-Genes were tested and ranked according to their g-mean.

2.6. Explainability

In this study we have used a number of ML algorithms, including RF, which belongs to the so-called ensemble classifiers. It is difficult to obtain understandable explanations for these classifiers, given their complexity. To understand the mechanisms behind RF models, which are more complex than the individual models from which they are derived, post-hoc explanatory techniques are needed [19]. Global explanation methods, such as the traditional feature importance for RF, can be used to explain the overall behavior of the model. However, global explanations lack the ability to explain individual predictions and do not allow the magnitude and direction of the effect of each feature on the final outcome to be determined. Feature relevance explanation technique SHAP, which we used for this work, can provide local explanations that allow us to fairly explain the underlying reasons behind individual predictions in terms of the contribution of each predictor to the final outcome. In addition, SHAP can also provide global explanations by building a matrix of Shapley values with one row per data instance and one column per feature, allowing predictors to be ranked according to their average contribution. These global and local explanations provide complementary information about the behavior of a model, which is key for experts to understand its mechanisms, especially in highly sensitive areas such as health.

SHAP is based on the idea of Shapley value, used in game theory, which assumes that a prediction can be explained by the assumption that every feature (in our case, a gene) is a "player" in a game where the prediction is the "payoff". The magnitude and sign of the attribution of each feature to the final result provided by the model are computed based on Shapley values, which allow the payoff to be allocated equally among the features.

We employ SHAP to compute the importance of each feature in each prediction, so that we have a more detailed idea of the mechanisms behind each of these predictions. Our model generates an output between 0 and 1 so that values below 0.5 are predicted as non-PCa-affected tissue and those equal to or above that threshold are classified as PCa tissue. In this context, we calculate the Shapley value for each feature in each prediction, which can be thought of as the effect of a specific gene on the final output and can be calculated as shown in Equation 2, where ϕ_i is the Shapley value for feature i, S is a feature subset, F is the set of all features, $f_{S\cup\{i\}}$ is the model trained with the feature i present, f_S is the model trained with the feature i withheld and x_S represents the values of the input features in the set S. A Shapley value is basically the marginal average contribution of a feature considering all possible combinations, which requires retraining the model on all feature subsets with and without including feature i.

It has been shown that these mechanisms can provide clinicians with accurate and reliable explanations, which will make medical experts more comfortable with RF decisions [38].

Table 3
Average quality metrics across the 25 test sets, for each class balancing strategy (Class Bal.) in the training sets for the TCGA-PRAD population.

Method	Class Bal.	G-mean	Sens.	Spec.	AUC	F1
RF	Undersampling	0.91	0.90	0.92	0.95	0.69
RF	Hybrid	0.90	0.85	0.95	0.96	0.74
RF	Upsampling	0.84	0.71	0.99	0.96	0.76
RF	Weight	0.80	0.65	0.99	0.96	0.73
RF	-	0.79	0.63	0.99	0.96	0.71
KNN	Undersampling	0.89	0.92	0.86	0.94	0.58
KNN	Hybrid	0.89	0.91	0.88	0.93	0.61
KNN	Upsampling	0.88	0.92	0.84	0.93	0.54
KNN	Weight	0.70	0.50	0.99	0.93	0.60
KNN	-	0.70	0.50	0.99	0.92	0.60
rpart	Undersampling	0.87	0.85	0.88	0.87	0.57
rpart	Hybrid	0.86	0.82	0.89	0.86	0.58
rpart	Upsampling	0.85	0.83	0.88	0.85	0.55
rpart	Weight	0.85	0.82	0.88	0.85	0.56
rpart	-	0.73	0.56	0.97	0.77	0.59

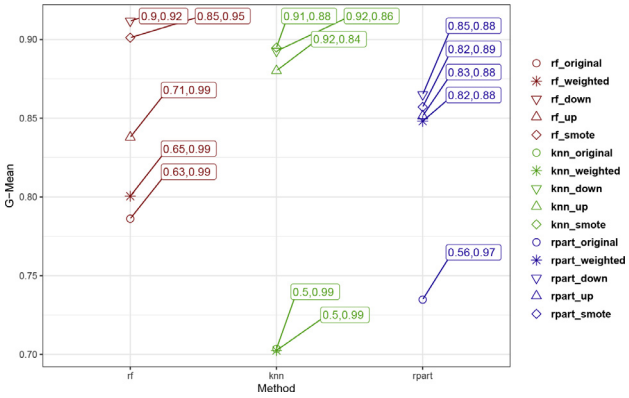

Fig. 3. G-mean, sensitivity and specificity values (the last two in the box) for the different models based on RF, KNN and rpart methods and different class-balancing approaches.

Gene importance was then calculated as the mean of the absolute value of these Shapley values for each feature in each sample belonging to a given dataset, as shown in Eq. 3, where I_j is the importance for feature j, n is the number of samples in the population and $\phi_j^{(i)}$ is the Shapley value for sample i and feature j.

$$\phi_i = \sum_{S\subseteq F\setminus\{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!}[f_{S\cup\{i\}}(x_{S\cup\{i\}}) - f_S(x_S)] \quad (2)$$

$$I_j = \frac{1}{n}\sum_{i=1}^{n} |\phi_j^{(i)}| \quad (3)$$

3. Results and discussion

3.1. Results

Table 3 shows the average results obtained for each quality metric considered in the 25 test sets, with each of the strategies used to balance the sample classes in the training sets, for each method in our discovery population TCGA-PRAD. Fig. 3 shows a graphical overview of these results. Non-parametric tests were used to compare these results in order to choose the best performing algorithm for G-mean, F1, AUC, Sensitivity and Specificity values. Despite the heterogeneity of the methods used in this study, we applied Friedman's test [39], rejecting the null hypothesis. Friedman's ranking for each measure can be found in Table 4, with RF ranking first in 3 of the 4 measures. The Shaffer test [40] was then considered for pairwise comparisons between methods. For

Table 4
Statistical test results for the measures G-mean, F1, AUC, Sensitivity and Specificity .

Algorithm	Ranking G-mean	Ranking F1	Ranking AUC	Ranking Sens.	Ranking Spec.
RF	**1.60**	**1.44**	**1.28**	1.88	**1.54**
KNN	1.84	2.06	1.90	**1.84**	2.24
rpart	2.56	2.5	2.92	2.28	2.22
Algorithm	**APV G-mean**	**APV F1**	**APV AUC**	**APV Sens.**	**APV Spec.**
RF	-	-	-	0.888	-
KNN	0.396	0.028	0.05	-	0.039
rpart	0.002	0.001	0	0.359	0.039

Table 5
Final classifier metrics across databases .

Database	G-mean	Sens.	Spec.	AUC	F1
TCGA-PRAD	0.91	0.90	0.92	0.91	0.70
GSE114740	0.85	0.90	0.80	0.85	0.86
GSE183019	0.80	0.93	0.70	0.81	0.83
GSE22260	0.77	0.80	0.75	0.78	0.70
GTEX (prost.)	NA	0.99	NA	NA	0.99

each observation, adjusted p-values (APVs) were calculated to assure the lowest degree of significance prior to rejecting the equality hypothesis. The numerical output for the comparisons are also shown in Table 4. Significant differences of at least 0.05 with the other methods are observed for the measurements in which RF is the best method, except for KNN and g-mean, in which there are no significant differences. As for the sensitivity measure, we observed that there are no significant differences between the methods when KNN is the method best classified by Friedman, showing that all the methods present a similar behavior on the positive class (NT). According to the statistical results obtained, we can state that RF is the method with the best statistical performance in the 25 runs on the test sets.

We tested the classifier with validation populations to see to what extent it was able to generalize its predictions when different ancestries, sequencing technologies and analysis pipelines came into play. We trained our final classifier using RF, in conjunction with the undersampling strategy to generate a model over the complete discovery dataset. As described in Table 5, the results also look promising, with g-mean, AUC and F1 values always above 0.7 and the following sensibility/specificity values in the different populations: GSE22260 (0.8, 0.75), GSE114740 (0.9, 0.8), GSE183019 (0.93, 0.7) and GTEx (0.99, NA). There were no tumor samples for GTEx, so the specificity value could not be calculated, and for GSE183019 we had to rely on the array counts provided, and could not apply our RNASeq pipeline to the raw sequencing data.

Once the results had been obtained, we focused on uncovering the mechanisms underlying our algorithm by using SHAP. Fig. 4 shows a graphical representation of the importance of our top 20 predictors, according to their importance, in our classifier for the TCGA-PRAD dataset. Genes are ordered according to their overall importance. Each dot represents the attribution of a given gene in the classifier's final prediction for a specific patient, and its color is determined by that gene expression value for each patient (red=high, blue=low). Red values represent higher expression values, while blue tones are associated with poorly expressed genes. On the left, only samples with PCa are represented, while on the right only healthy samples are shown. At the bottom, the classifier's explanation for a single patient is displayed; the numerical output of the classifier is explained with the individual contribution of each predictor, in some cases adding and in others subtracting until the final value is obtained. The most influential genes are labeled.

Note that blue and red tones are often separated by the zero-impact value, which means that, for most genes, their contribution to the final outcome is strongly linked to their expression level.

3.2. Discussion

The heterogeneity present in the way that the different datasets were obtained, including a variety of analysis pipelines, sequencing technologies and different reference genomes when aligning reads, leads us to believe that the results, which will be discussed later, could have been even better if we had been able to apply our RNA-SEQ pipeline to each dataset. Moreover, it also shows that our classifier tolerates some flexibility regarding how the input data is processed. In addition to the input data format and the reference genome build used, this system is also able to correctly classify samples of very different ethnicities (see supplementary data, section Datasets).

The strength of the present classifier is based on the final genes included in the classifer, such as, *DLX1* (Distal-Less Homeobox 1), which is currently included in SelectMDX urine test as a diagnostic risk biomarker for classification in negative and misclassified PCa biopsies [41].

HPN (Hepsin) has also been shown to distinguish normal tissue from PCa lesions through single-cell RNA sequencing [42]; the same is true for *CNN1* (Calponin 1) with differences demonstrated between tumor and normal tissues [43]. Moreover, *ANXA2* (Annexin A2 or Annexin II) has been suggested as a prognostic biomarker of PCa because of its association between high expression patterns and higher grade and stage PCa [44].

There are also results suggesting an oncogenic role of *AMACR* (alpha-methylacyl-CoA racemase) in PCa and indicating its role as a potential biomarker for its diagnosis [45]. It has also been shown to be a marker of recurrence after radical prostatectomy [46], but currently has no application in clinical screening. Moreover, *MYL9* (Myosin light chain 9) is closely associated with poor prognosis in several tumors such as PCa, lung, breast and melanoma. Its role as a molecular marker and potential target for early diagnosis, prognostic prediction and selective treatment of malignant tumors has been proposed [47].

Similar to the genes that rank first in terms of SHAP importance, those ranked last are also biologically relevant in PCa. *DCN* has been previously reported as a prognostic marker of PCa in tissue [48]. *MYO6* (Myosin VI) is suggested to play an essential role in PCa progression and has promising therapeutic effects [49]. There are not many reports citing *TFF3* (trefoil factor 3), but it has been suggested to play a role in the stratification of PCa in combination with *HOXB13* (Homeobox B13). *SVIL* (Supervillin) has also been mentioned as a possible methylation-specific marker of PCa, but with low sensitivity (75.4%) [50]. With regard to *TIMP3* and *KRT7*, they have previously been linked to a therapeutic implication in PCa, but not to a screening target [51]. Another important gene is *FGFR2*, which is a fibroblast growth factor receptor and a membrane receptor that promotes cell proliferation and differentiation. As we see in our results, the downregulation of *FGFR2* is associated with poor prognosis in PCa [52] but not in other types of cancers, thus, this gene seems a very specific marker and therefore relevant for the specificity of our classifier. *EPHA2* is also very interesting, as it is the most extensively studied EphA receptor in PCa. Initial studies identified *EPHA2* protein overexpression in PCa cell lines related with metastatic potential. However, normal and benign prostate tumor cells showed weak or no staining with the EphA2 antibody[53]. As seen in Table 6, another relevant gene is *TDRD1*, which is thought to function in the suppression of transposable elements during spermatogenesis. It has been observed

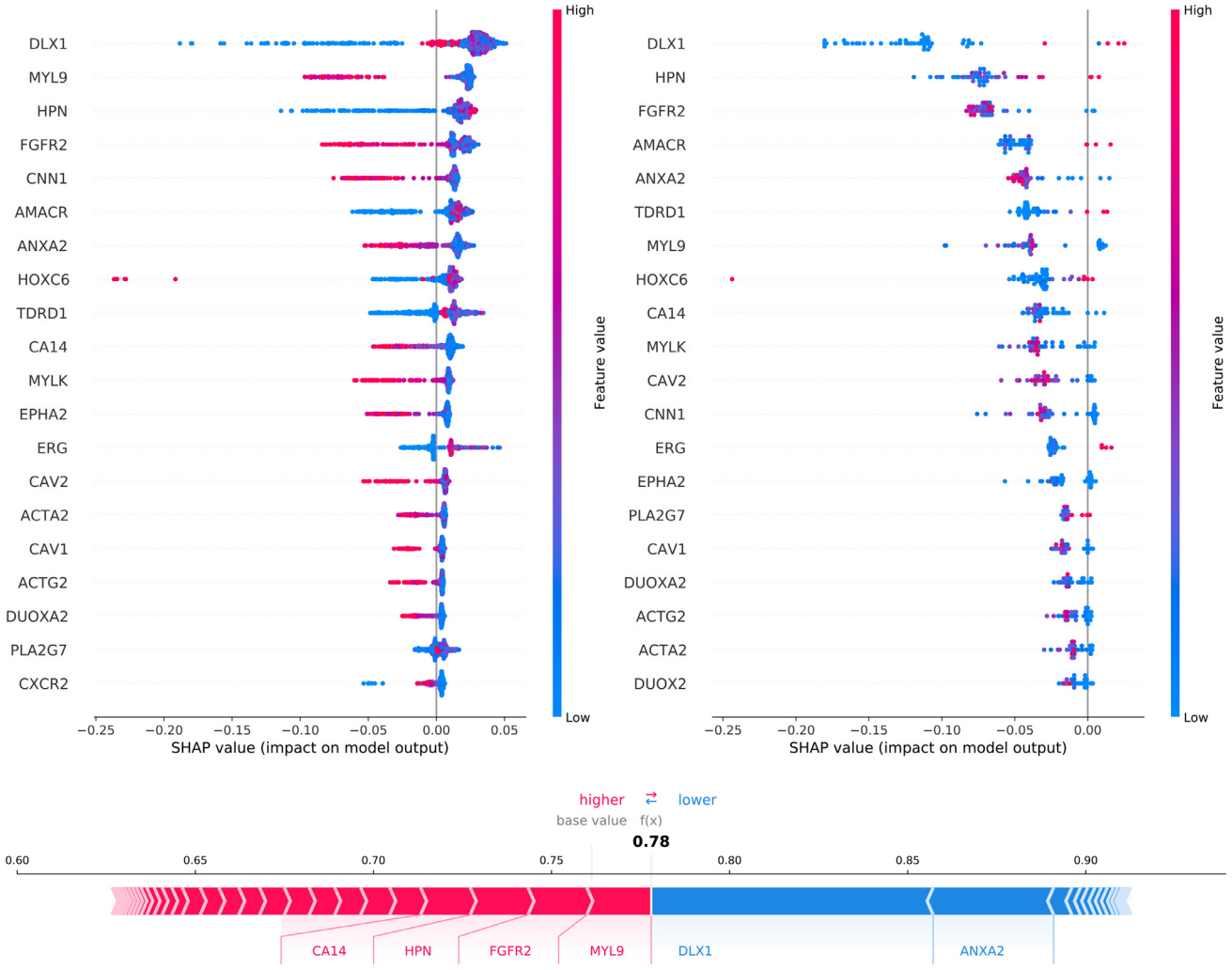

Fig. 4. SHAP analysis for our RF model. Top 20 genes, according to the importance are shown for T samples (on the left) and NT samples (on the right). Each point represents the impact on the model output for a specific patient and gene. Color shows expression value. At the bottom, the prediction for a single patient is shown, the impact on the final result of the most influential genes is represented.

Table 6
Inferred patterns in classifier prediction .

Class	Pattern
T	↑ *DLX1*, ↑ *HPN*, ↓ *FGFR2*, ↓ ***MYL9***, ↓ *CNN1*, ↑ *TDRD1*
NT	↓ *DLX1*, ↓ ***HPN***, ↑ *FGFR2*, ↑ *AMACR*, ↓ *TDRD1*

Up and down arrows indicate high and low expression levels, respectively. Genes in bold represent the most relevant genes in each class, after *DLX1*.

that the TDRD1 protein is expressed in the majority of human prostate tumors, but not in normal prostate tissue, so in this regard it has been proposed as a novel PCa biomarker [54].

The most influential gene for T samples is *DLX1*, followed by *MYL9, HPN, FGFR2* and *CNN1*, while in the NT class the most relevant gene is still *DLX1* but this time followed by *HPN, FGFR2, AMACR* and *ANXA2*. After the predictor *DLX1*, shared by both classes, the most influential genes for the T and NT classes are *MYL9* and *HPN*, respectively. Table 6 summarizes these patterns in relation to the expression level of the most relevant genes.

The expression patterns of the main genes included in Table 6 help us to classify T and NT samples, mainly with decreased expression in T samples, with *FGFR2, CNN1* and *ANXA2* playing an important role. Gene silencing is also relevant in the T samples in genes such as *CA14* or *EPHA2*. Furthermore, in the T samples, there is an increased expression in 14 genes, among which the major contributors to the present algorithm are *DLX1, HPN, AMACR, HOXC6* and *TDRD1*. Therefore, these 5 genes represent a set of relevant markers because they share an increasing expression pattern in all populations. Consequently, they may be the most reasonable choice for application in the detection of liquid biopsies, as is the case with *DLX1* [41].

It is important to highlight the contribution of the present proposal. Although some of the genes included in our classifier have been previously reported individually as related to prognostic tumoral tissue classification in PCa; they have never been used together until now. In addition, the present work also includes genes that are barely or not at all described in PCa, that are key for decision making in both our discovery and validation populations, such as *MYLK* (Myosin Light Chain Kinase), *CAV2* (Caveolin 2), *TDRD1* (Tudor Domain Containing 1) as well as other genes that have never been previously described in PCa such as *RNF112* (Ring Finger Protein 112), *APOF* (Apolipoprotein F) and *MYOCD* (Myocardin). These genes are related to the tumor microenvironment, but with different roles. *MYLK* and *MYOCD* genes promote tumor formation and vascularization [55,56], *RNF112* gene is related to cellular differentiation, *CAV2* modulates mitotic pathways and *APOF* is involved in the regulation of cellular transport. However, the most

interesting part is that several of them have been previously identified as biomarkers in other tumors, but never in PCa, which reinforces their utility as biomarkers. For example, *APOF* regulates cell transport and has been described as a biomarker or target for hepatocellular carcinoma [57,58] or cervical cancer [59]. In the case of RNF112, it has been described as a prognostic biomarker in oral cancer [60].

In the near future, by combining our classifier and single-cell strategies we will be able to identify a tumor when there is only one malignant cell, by detecting the signature of these cells from among the rest of the healthy cells, which will improve the accuracy of the imaging techniques currently used for diagnosis [61].

Finally, current genomic methodologies could provide expression analysis in tissue (fresh or paraffin-embedded) with a high success rate and sequence coverage. Therefore, the application of this algorithm in current medicine and clinical practice for PCa classification is feasible at a low cost.

4. Conclusions

In this work, we have addressed the development of a classifier to predict the risk of PCa in prostate tissue based on a set of biologically relevant genes that could provide explanatory power to its predictions, using the well-known SHAP algorithm. This classifier showed good results considering several quality metrics widely used in ML, not only in the discovery population but also in external populations with a wide range of ancestries. The fact that biomarkers for PCa screening are not currently used in clinical practice highlights the interest of this work, in which we have demonstrated the relevance of *DLX1, MYL9* and *FGFR* genes, in addition to novel genes for PCa screening such as *CAV2* and *MYLK*. The lowest ranked predictors of our classifier complement the remaining most relevant genes to achieve the good accuracy demonstrated by the algorithm, as they are involved in metabolic pathways and biological processes of interest for this disease.

To the best of our knowledge, this is the first time that a classifier combining gene expression and ML has been used for PCa detection and screening. With the help of this tool, the misclassification rates of anatomopathological analysis could be decreased, thus reducing the need for repeated biopsies. Thanks to the development of this tool, fundamental genes in the development and evolution of PCa have been identified for evaluation in the clinical practice. Finally, the application of this algorithm to other sample types, such as urine or blood, could allow for its use as part of the liquid biopsy strategy in PCa in the future.

Funding

This work was supported by the ERDF and the Ministry of Economy, Innovation and Science of the Regional Government of Andalusia (grant number P18-RT-2248).

Declaration of Competing Interest

The authors of this study declare that they have no conflicts of interest.

Acknowledgements

The results published here are based on data generated by the TCGA Research Network: https://www.cancer.gov/tcga, GTEX, GSE22260, GSE183019 and GSE114740. For this reason, we thank all the donors who contributed samples to these projects.

Appendix A. Supplementary data

A website with supplementary data has been developed for this study and is available at the following URL: https://sci2s.ugr.es/PCaXAIRF.

References

[1] T. Dyba, G. Randi, F. Bray, C. Martos, F. Giusti, N. Nicholson, A. Gavin, M. Flego, L. Neamtiu, N. Dimitrova, R. Negrã Carvalho, J. Ferlay, M. Bettio, The european cancer burden in 2020: incidence and mortality estimates for 40 countries and 25 major cancers, Eur. J. Cancer 157 (2021) 308–347, doi:10.1016/j.ejca.2021.07.039.

[2] N. Mottet, R.C.N. van den Bergh, E. Briers, T. Van den Broeck, M.G. Cumberbatch, M. De Santis, S. Fanti, N. Fossati, G. Gandaglia, S. Gillessen, N. Grivas, J. Grummet, A.M. Henry, T.H. van der Kwast, T.B. Lam, M. Lardas, M. Liew, M.D. Mason, L. Moris, D.E. Oprea-Lager, H.G. van der Poel, O. Rouvière, I.G. Schoots, D. Tilki, T. Wiegel, P.-P.M. Willemse, P. Cornford, EAU-EANM-ESTRO-ESUR-SIOG Guidelines on prostate cancer-2020 update. part 1: screening, diagnosis, and local treatment with curative intent, Eur. Urol. 79 (2) (2021) 243–262, doi:10.1016/j.eururo.2020.09.042.

[3] S. Mehralivand, D. Yang, S.A. Harmon, D. Xu, Z. Xu, H. Roth, S. Masoudi, D. Kesani, N. Lay, M.J. Merino, B.J. Wood, P.A. Pinto, P.L. Choyke, B. Turkbey, Deep learning-based artificial intelligence for prostate cancer detection at biparametric MRI, Abdomin. Radiol. 47 (4) (2022) 1425–1434, doi:10.1007/s00261-022-03419-2.

[4] J. Yang, S. Xiao, A. Li, W. Lu, X. Gao, Y. Li, MSTA-Net: forgery detection by generating manipulation trace based on multi-Scale self-Texture attention, IEEE Trans. Circuits Syst. Video Technol. 32 (7) (2022) 4854–4866, doi:10.1109/tcsvt.2021.3133859.

[5] Y. Zhao, S. Xiao, J. Yang, W. Lu, X. Gao, No-reference quality index of tone-mapped images based on authenticity, preservation, and scene expressiveness, Signal Process. 203 (2023) 108782, doi:10.1016/j.sigpro.2022.108782.

[6] A.M. Grimaldi, M. Salvatore, C. Cavaliere, Diagnostic and prognostic significance of extracellular vesicles in prostate cancer drug resistance: a systematic review of the literature, Prostate Cancer Prostatic Dis. (2022) 1–12, doi:10.4274/jcrpe.1471.

[7] L.J. Martínez-González, V. Sánchez-Conde, J.M. Gonzälez-Cabezuelo, A. Antunez-Rodríguez, E. Andrés-León, I. Robles-Fernandez, J.A. Lorente, F. Vázquez-Alonso, M.J. Alvarez-Cubero, Identification of mirnas as viable aggressiveness biomarkers for prostate cancer, Biomedicines 9 (6) (2021), doi:10.3390/biomedicines9060646.

[8] Y. Zhu, M. Mo, Y. Wei, J. Wu, J. Pan, S.J. Freedland, Y. Zheng, D. Ye, Epidemiology and genomics of prostate cancer in Asian men, Nat. Rev. Urol. 18 (5) (2021) 282–301, doi:10.1038/s41585-021-00442-8.

[9] E.D. de la Guardia-Bolívar, R. Barrios-Rodríguez, I. Zwir, J.J. Jiménez-Moleón, C. del Val, Identification of novel prostate cancer genes in patients stratified by gleason classification: role of antitumoral genes, Int. J. Cancer (2022) 1–10, doi:10.1002/ijc.33988.

[10] J. Alexander, J. Kendall, J. McIndoo, L. Rodgers, R. Aboukhalil, D. Levy, A. Stepansky, G. Sun, L. Chobardjiev, M. Riggs, H. Cox, I. Hakker, D.G. Nowak, J. Laze, E. Llukani, A. Srivastava, S. Gruschow, S.S. Yadav, B. Robinson, G. Atwal, L.C. Trotman, H. Lepor, J. Hicks, M. Wigler, A. Krasnitz, Utility of single-cell genomics in diagnostic evaluation of prostate cancer, Cancer Res. 78 (2) (2018) 348–358, doi:10.1158/0008-5472.CAN-17-1138.

[11] L.-Y. Ye, X.-Y. Miao, W.-S.C. Wan Jiang Xu, Medical image diagnosis of prostate tumor based on PSP-Net+VGG16 deep learning network, Comput. Methods Programs Biomed. 221 (2022) 106770, doi:10.1016/j.cmpb.2022.106770.

[12] A. Balagopal, H. Morgan, M. Dohopolski, R. Timmerman, J. Shan, D.F. Heitjan, W. Liu, D. Nguyen, R. Hannan, A. Garant, N. Desai, S. Jiang, PSA-Net: deep learning-based physician style-aware segmentation network for postoperative prostate cancer clinical target volumes, Artif. Intell. Med. 121 (2021), doi:10.1016/j.artmed.2021.102195.

[13] A. Adadi, M. Berrada, Peeking inside the black-Box: a survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160, doi:10.1109/ACCESS.2018.2870052.

[14] E. Tjoa, C. Guan, A survey on explainable artificial intelligence (XAI): toward medical XAI, IEEE Trans. Neural Netw. Learn. Syst. 32 (11) (2021) 4793–4813, doi:10.1109/TNNLS.2020.3027314.

[15] H.W. Loh, C.P. Ooi, S. Seoni, P.D. Barua, F. Molinari, U.R. Acharya, Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011–2022), Comput. Methods Programs Biomed. 226 (2022), doi:10.1016/j.cmpb.2022.107161.

[16] D. Shin, Y.J. Park, Role of fairness, accountability, and transparency in algorithmic affordance, Comput. Human Behav. 98 (2019) 277–284, doi:10.1016/j.chb.2019.04.019.

[17] A. Abeshouse, et al., The molecular taxonomy of primary prostate cancer, Cell 163 (4) (2015) 1011–1025, doi:10.1016/j.cell.2015.10.025.

[18] L. Breiman, Statistical modeling: the two cultures (with comments and a rejoinder by the author), Stat. Sci. 16 (3) (2001), doi:10.1214/ss/1009213726.

[19] A.B. Arrieta, N. Díaz-Rodríguez, J.D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and chal-

lenges toward responsible AI, Inf. Fusion 58 (2020) 82–115, doi:10.1016/j.inffus.2019.12.012.
[20] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nature Mach. Intell. 2 (1) (2020) 56–67, doi:10.1038/s42256-019-0138-9.
[21] K. Kannan, L. Wang, J. Wang, M.M. Ittmann, W. Li, L. Yen, Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing, Proc. Natl. Acad. Sci. 108 (22) (2011) 9172–9177, doi:10.1073/pnas.1100489108.
[22] R. Edgar, Gene expression omnibus: NCBI gene expression and hybridization array data repository, Nucl. Acids Res. 30 (1) (2002) 207–210, doi:10.1093/nar/30.1.207.
[23] E. Andrés-León, R. Núñez-Torres, A.M. Rojas, MiARma-Seq: a comprehensive tool for miRNA, mRNA and circRNA analysis, Sci. Rep. 6 (2016), doi:10.1038/srep25749.
[24] A. Dobin, C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, T.R. Gingeras, STAR: Ultrafast universal RNA-seq aligner, Bioinformatics 29 (1) (2012) 15–21, doi:10.1093/bioinformatics/bts635.
[25] Y. Liao, G.K. Smyth, W. Shi, Featurecounts: an efficient general purpose program for assigning sequence reads to genomic features, Bioinformatics 30 (7) (2013) 923–930, doi:10.1093/bioinformatics/btt656.
[26] M.D. Robinson, D.J. McCarthy, G.K. Smyth, Edger: a bioconductor package for differential expression analysis of digital gene expression data, Bioinformatics 26 (1) (2009) 139–140, doi:10.1093/bioinformatics/btp616.
[27] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J.M. Benítez, F. Herrera, A review of microarray datasets and applied feature selection methods, Inf. Sci. (Ny) 282 (2014), doi:10.1016/j.ins.2014.05.042.
[28] J. Yang, Z. Zhang, S. Xiao, S. Ma, Y. Li, W. Lu, X. Gao, Efficient data-driven behavior identification based on vision transformers for human activity understanding, Neurocomputing 530 (2023) 104–115, doi:10.1016/j.neucom.2023.01.067.
[29] S. Anders, D.J. McCarthy, Y. Chen, M. Okoniewski, G.K. Smyth, W. Huber, M.D. Robinson, Count-based differential expression analysis of RNA sequencing data using r and bioconductor, Nat. Protoc. 8 (9) (2013) 1765–1786, doi:10.1038/nprot.2013.099.
[30] L.J. Jensen, M. Kuhn, M. Stark, S. Chaffron, C. Creevey, J. Muller, T. Doerks, P. Julien, A. Roth, M. Simonovic, P. Bork, C. von Mering, STRING 8–A global view on proteins and their functional interactions in 630 organisms, Nucl. Acids Res. 37 (Database) (2009) D412–D416, doi:10.1093/nar/gkn760.
[31] M. Kuhn, Building predictive models in r using the caret package, J. Stat. Softw. 28 (5) (2008) 1–26, doi:10.18637/jss.v028.i05.
[32] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357, doi:10.1613/jair.953.
[33] M. Kuhn, K. Johnson, Feature engineering and selection: a practical approach for predictive models, 2019, doi:10.1201/9781315108230.
[34] N.S. Altman, An introduction to kernel and nearest-Neighbor nonparametric regression, Am. Stat. 46 (3) (1992) 175–185, doi:10.2307/2685209.
[35] T.K. Ho, Random decision forests, in: Proceedings of 3rd international conference on document analysis and recognition, volume 1, IEEE, 1995, pp. 278–282.
[36] A. Fernández, S. García, M. Galar, R.C. Prati, B. Krawczyk, F. Herrera, Performance measures, in: Learning from Imbalanced Data Sets, Springer International Publishing, 2018, pp. 47–61, doi:10.1007/978-3-319-98074-4.
[37] J.S. Akosa, Predictive accuracy : a misleading performance measure for highly imbalanced data, 2017.
[38] S. El-Sappagh, J.M. Alonso, S.M.R. Islam, A.M. Sultan, K.S. Kwak, A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease, Sci. Rep. 11 (1) (2021), doi:10.1038/s41598-021-82098-3.
[39] M. Friedman, The use of ranks to avoid the assumption of normality implicit in the analysis of variance, J. Am. Stat. Assoc. 32 (200) (1937) 675–701, doi:10.1080/01621459.1937.10503522.
[40] J.P. Shaffer, Modified sequentially rejective multiple test procedures, J. Am. Stat. Assoc. 81 (395) (1986) 826–831, doi:10.1080/01621459.1986.10478341.
[41] M. Maggi, F.D. Giudice, U.G. Falagario, A. Cocci, G.I. Russo, M.D. Mauro, G.S. Sepe, F. Galasso, R. Leonardi, G. Iacona, P.R. Carroll, M.R. Cooperberg, A. Porreca, M. Ferro, G. Lucarelli, D. Terracciano, L. Cormio, G. Carrieri, E. De Berardinis, A. Sciarra, G.M. Busetto, SelectMDx and multiparametric magnetic resonance imaging of the prostate for men undergoing primary prostate biopsy: aprospective assessment in a multi-Institutional study, Cancers (Basel) 13 (9) (2021) 2047, doi:10.3390/cancers13092047.
[42] X. Ma, J. Guo, K. Liu, L. Chen, D. Liu, S. Dong, J. Xia, Q. Long, Y. Yue, P. Zhao, F. Hu, Z. Xiao, X. Pan, K. Xiao, Z. Cheng, Z. Ke, Z.-S. Chen, C. Zou, Identification of a distinct luminal subgroup diagnosing and stratifying early stage prostate cancer by tissue-based single-cell RNA sequencing, Mol. Cancer 19 (1) (2020), doi:10.1186/s12943-020-01264-9.
[43] X. Chen, J. Wang, X. Peng, K. Liu, C. Zhang, X. Zeng, Y. Lai, Comprehensive analysis of biomarkers for prostate cancer based on weighted gene co-expression network analysis, Medicine (United States) 99 (14) (2020), doi:10.1097/MD.0000000000019628.
[44] S.-H. Tan, D. Young, Y. Chen, H.-C. Kuo, A. Srinivasan, A. Dobi, G. Petrovics, J. Cullen, D.G. Mcleod, I.L. Rosner, S. Srivastava, I.A. Sesterhenn, Prognostic features of annexin A2 expression in prostate cancer, Pathology 53 (2) (2021) 205–213, doi:10.1016/j.pathol.2020.07.006.
[45] P. Fu, C. Bu, B. Cui, N. Li, J. Wu, Screening of differentially expressed genes and identification of AMACR as a prognostic marker in prostate cancer, Andrologia 53 (6) (2021), doi:10.1111/and.14067.
[46] E. Gökmen, O. Özman, M. Kars, S. Gönültaş, B. Arslan, Relationship between biopsy core α-methylacyl-Coa racemase positivity and five-year biochemical recurrence in D'amico low- and intermediate-risk prostate cancer, Üroonkoloji bül. 20 (2) (2021) 92–95, doi:10.4274/uob.galenos.2020.1721.
[47] Y. You, T. Liu, J. Shen, Research progress in myosin light chain 9 in malignant tumors, J. Central South Univ. (Med. Sci.) 46 (10) (2021) 1153–1158, doi:10.11817/j.issn.1672-7347.2021.200814.
[48] R. Rezaie, Z. Falakian, S. Mazloomzadeh, M. Ayati, A. Morakabati, M.R.T. Dastjerdan, M. Zare, M. Moghimi, T. Shahani, A. Biglari, While urine and plasma decorin remain unchanged in prostate cancer, prostatic tissue decorin has a prognostic value, Iran. Biomed. J. 24 (4) (2020) 229–235, doi:10.29252/ibj.24.4.229.
[49] D. Wang, L. Zhu, M. Liao, T. Zeng, W. Zhuo, S. Yang, W. Wu, MYO6 Knockdown inhibits the growth and induces the apoptosis of prostate cancer cells by decreasing the phosphorylation of ERK1/2 and PRAS40, Oncol. Rep. 36 (3) (2016) 1285–1292, doi:10.3892/or.2016.4910.
[50] D.K. Vanaja, K.V. Ballman, B.W. Morlan, J.C. Cheville, R.M. Neumann, M.M. Lieber, D.J. Tindall, C.Y.F. Young, PDLIM4 Repression by hypermethylation as a potential biomarker for prostate cancer, Clin. Cancer Res. 12 (4) (2006) 1128–1136, doi:10.1158/1078-0432.CCR-05-2072.
[51] X. Deng, S. Bhagat, Z. Dong, C. Mullins, S.R. Chinni, M. Cher, Tissue inhibitor of metalloproteinase-3 induces apoptosis in prostate cancer cells and confers increased sensitivity to paclitaxel, Eur. J. Cancer 42 (18) (2006) 3267–3273, doi:10.1016/j.ejca.2006.07.003.
[52] J.E. Lee, S.-H. Shin, H.-W. Shin, Y.-S. Chun, J.-W. Park, Nuclear FGFR2 negatively regulates hypoxia-induced cell invasion in prostate cancer by interacting with HIF-1 and HIF-2, Sci. Rep. 9 (1) (2019), doi:10.1038/s41598-019-39843-6.
[53] J. Walker-Daniels, K. Coffman, M. Azimi, J. Rhim, D. Bostwick, P. Snyder, B. Kerns, D. Waters, M. Kinch, Overexpression of the epha2 tyrosine kinase in prostate cancer, Prostate 41 (2000), doi:10.1002/(SICI)1097-0045(19991201)41:43.0.CO;2-T. 275–80
[54] L. Xiao, R.B. Lanz, A. Frolov, P.D. Castro, Z. Zhang, B. Dong, W. Xue, S.Y. Jung, J.P. Lydon, D.P. Edwards, M.A. Mancini, Q. Feng, M.M. Ittmann, B. He, The germ cell gene TDRD1 as an ERG target gene and a novel prostate cancer biomarker, Prostate 76 (14) (2016) 1271–1284, doi:10.1002/pros.23213.
[55] C. Liu, H. Pei, F. Tan, Matrix stiffness and colorectal cancer, Onco. Targets Ther. 13 (2020) 2747–2755, doi:10.2147/OTT.S231010.
[56] C.-L. Lu, M.-T. Liao, Y.-C. Hou, Y.-W. Fang, C.-M. Zheng, W.-C. Liu, C.-T. Chao, K.-C. Lu, Y.-Y. Ng, Sirtuin-1 and its relevance in vascular calcification, Int. J. Mol. Sci. 21 (2020), doi:10.3390/ijms21051593.
[57] Y.-B. Wang, B.-X. Zhou, Y.-B. Ling, Z.-Y. Xiong, R.-X. Li, Y.-S. Zhong, M.-X. Xu, Y. Lu, H. Liang, G.-H. Chen, Z.-C. Yao, M.-H. Deng, Decreased expression of apof associates with poor prognosis in human hepatocellular carcinoma, Gastroenterol. Rep. (Oxf.) 7 (2019) 354–360, doi:10.1093/gastro/goz011.
[58] H. Sharifi, H. Safarpour, M. Moossavi, M. Khorashadizadeh, Identification of potential prognostic markers and key therapeutic targets in hepatocellular carcinoma using weighted gene co-Expression network analysis: A Systems biology approach, Iran. J. Biotechnol. 20 (2022) e2968, doi:10.30498/ijb.2022.269817.2968.
[59] S. Han, J. Zhang, Y. Sun, L. Liu, L. Guo, C. Zhao, J. Zhang, Q. Qian, B. Cui, Y. Zhang, The plasma DIA-Based quantitative proteomics reveals the pathogenic pathways and new biomarkers in cervical cancer and high grade squamous intraepithelial lesion, J. Clin. Med. 11 (2022), doi:10.3390/jcm11237155.
[60] S.K. Kuk, J.I. Lee, K. Kim, Prognostic genomic markers of pathological stage in oral squamous cell carcinoma, Head Neck Pathol. (2022), doi:10.1007/s12105-022-01516-8.
[61] C. Chen, J. Luo, X. Wang, Identification of prostate cancer subtypes based on immune signature scores in bulk and single-cell transcriptomes, Med. Oncol. 39 (9) (2022), doi:10.1007/s12032-022-01719-7.