Exploring explainable machine learning to predict genetic factors associated with survival in renal cell carcinoma using transcriptomic data

Grace dos Santos Feijó[1], **Mariana Recamonde-Mendoza** [1,2]

[1]Institute of Informatics, Universidade Federal do Rio Grande do Sul (UFRGS), Brazil

[2]Núcleo de Bioinformática, Hospital de Clínicas de Porto Alegre (HCPA), Brazil

{grace.feijo@inf.ufrgs.br, mrmendoza@inf.ufrgs.br

Abstract. *The integration of machine learning (ML) with transcriptomic data offers a promising path for advancing precision oncology by improving predictive performance and identifying clinically relevant molecular signatures. This study analyzes gene signatures associated with survival in Renal Cell Carcinoma subtypes (KIRC and KIRP) using TCGA data, combining predictive modeling with three Explainable AI techniques (SHAP, permutation importance, and surrogate decision trees). Among five evaluated algorithms, CatBoost achieved the best performance (AUC-ROC: 0.82 for KIRC). The XAI analysis identified both established and novel biomarker candidates related to therapeutic resistance and inflammatory pathways. These results highlight the potential of explainable AI to bridge predictive accuracy and biological interpretability.*

1. Introduction

Transcriptomic data, particularly RNA sequencing (RNA-seq), has provided vital information about tumor progression and patient prognosis by offering a comprehensive overview of gene expression levels at a given time [Shah and Lodhi 2025, Tran et al. 2025]. However, the high dimensionality and inherent noise of genomic data represent significant challenges for traditional statistical methods. While Machine Learning (ML) algorithms have demonstrated superior performance in capturing non-linear relationships within these datasets, their clinical adoption is often hindered by their ”black-box” nature [Seo et al. 2025].

In the medical domain, the rationale behind a prediction is as important as the predi ction itself. Clinicians require interpretable models to understand which biological mechanisms drive a patient's risk profile. As argued by [Sharma et al. 2021], the lack of transparency in ML models can lead to ethical concerns and a lack of trust in automated diagnostic tools. To address this, the field of Explainable Artificial Intelligence (XAI) has emerged, providing methodologies to decompose complex model decisions into human-understandable insights.

This study focuses on survival prediction for Renal Cell Carcinomas, specifically the Kidney Renal Clear Cell Carcinoma (KIRC) and Kidney Renal Papillary Cell Carcinoma (KIRP) subtypes. These subtypes exhibit both distinct outlooks and metabolic and immunological profiles, demanding specialized prognostic models [Albertson et al. 2025]. By integrating state-of-the-art ML classifiers with XAI techniques like SHAP, permutation importance and surrogate decision tree models, we aim to identify gene signatures that influence patient survival. This approach not only enhances predictive performance but also identifies potential biomarkers that could serve as targets for future therapeutic interventions.

2. Related Work

Renal tumors have been widely investigated using transcriptomic data and ML techniques due to their biological heterogeneity and variability in clinical outcomes [Tabibu et al. 2019]. Among renal cell carcinomas, KIRC and KIRP present distinct molecular profiles, histopathological characteristics, and clinical behaviors, which influence diagnosis, prognosis, and treatment strategies [Tabibu et al. 2019]. KIRC is the most prevalent subtype, accounting for approximately 70–75% of cases and typically exhibiting higher aggressiveness and metastatic potential. In contrast, KIRP represents about 15–20% of cases and generally shows a less aggressive clinical course, although advanced stages still pose important therapeutic challenges.

Several ML approaches have been applied to transcriptomic data analysis. Ensemble methods such as Random Forest, CatBoost, and gradient boosting algorithms (e.g., XGBoost and LightGBM) are well suited for handling the high dimensionality and noise typical of RNA-seq data [Peduzzi et al. 2025]. These models have shown strong predictive performance in clinical risk modeling tasks, including mortality prediction and the identification of prognostic genomic biomarkers [Su et al. 2024]. In addition, neural network-based models, such as Multi-Layer Perceptrons and Deep Neural Networks, have gained attention for their ability to capture complex nonlinear relationships in high-dimensional and multi-omic datasets [Tran et al. 2025].

Despite their predictive power, many ML models operate as black boxes, limiting their interpretability in biomedical applications. To address this challenge, explainable artificial intelligence (XAI) techniques have been increasingly adopted to improve transparency and facilitate biological interpretation [Seo et al. 2025]. Methods such as SHAP (SHapley Additive exPlanations), Permutation Feature Importance (PFI), and Global Interpretation via Recursive Partitioning (GIRP) provide complementary insights into model behavior, enabling the identification of prognostically relevant genes and supporting the translation of ML predictions into interpretable knowledge that may assist clinical decision-making.

3. Methodology

3.1. Data Sources

Transcriptomic and clinical information were acquired from The Cancer Genome Atlas (TCGA) portal. Data was already normalized using the $\log_2(x + 1)$ transformation. The dataset comprised 530 samples (173 for the positive class) for KIRC and 288 (44 for the positive class) for KIRP. The target variable was defined as overall survival, binarized for a classification approach, where participants were categorized as survivors (0) and non-survivors (1), with the latter considered the positive class in our analyses.

3.2. Dimensionality reduction

To reduce dimensionality, 10% of genes with the highest interquartile range (IQR) were selected, ensuring high variability while minimizing the influence of outliers. It was then narrowed down to include only genes that showed a statistically significant difference ($p < 0.05$) between the survivor and non-survivor groups. Finally, Spearman's correlation was applied, one gene was removed from any pair with a correlation coefficient greater than 0.98 to eliminate redundant information. The final dataset for KIRC was composed by 3454 genes and for KIRP by 951 genes.

3.3. ML models and performance analysis

Five supervised ML algorithms were selected to ensure a comprehensive evaluation: Random Forest (RF), Gradient Boosting (GB), Logistic Regression (LR), Multilayer Perceptron (MLP), and Categorical Boosting (CatBoost). The dataset was divided into training (80%) and test (20%) and a 5-fold cross-validation adopted on the training data dor hyperparameters optimization. All data splits were conducted in a stratifier manner.

For hyperparameter adjustment, Bayesian optimization was applied (30 optimization trials for each model). To address class imbalance, model weights were adjusted during training, and the Youden Index was employed to determine an optimal classification threshold rather than using a default 0.5. This threshold, established during the training/validation phase, was then applied to the test set to ensure a superior balance between sensitivity and specificity. Model performance was evaluated using the area under the Receiver Operating Characteristic curve (AUC-ROC), the Concordance Index (C-index), recall, precision, the area under the Precision-Recall curve (AUC-PR), and specificity.

3.4. XAI methods and high-importance genes

In this study, three XAI methods were employed. (i) SHAP (SHapley Additive exPlanations) to quantify the local contribution of each gene to individual predictions; (ii) Permutation Feature Importance (PFI) to assess global relevance by measuring performance changes after gene shuffling; (iii) Global Surrogate Decision Trees to approximate the complex models' logic, providing a visual and rule-based interpretation of the most critical decision paths. High-importance genes were defined as those identified at least three times by any XAI method across the five ML models, prioritizing features with consistent predictive value across algorithms and interpretability approaches.

4. Results and Discussion

The experimental results showed that CatBoost consistently outperformed other algorithms for the KIRC subtype, achieving an AUC-ROC and C-index of 0.82 (Table 1), highlighting its effectiveness in handling transcriptomic feature distributions. For KIRP, performance was lower due to fewer positive samples and class imbalance; however, Random Forest and CatBoost still showed promising results (Table 1).

The XAI analysis provided biological insights consistent with recent literature. Twelve high-importance genes were identified for KIRC and ten for KIRP. For KIRC, the model successfully identified nine well-established biomarkers, including IL20RB [Zhang et al. 2022] and SLFN11 [Zhou et al. 2025], which are linked to inflammatory signaling, DNA damage response, and sensitivity to various chemotherapeutic agents. These findings serve as a critical internal validation of our methodology, confirming the model's capacity to capture biologically relevant signals. Furthermore, the identification of three novel candidates (C1DP1 [Kolenda et al. 2023], MUM1L1 [Zhang et al. 2023],

Table 1. Performance of supervised models on the KIRC and KIRP datasets.

Model	ROC-AUC	PR-AUC	Accuracy	Precision	Recall	Specificity	C-index	Threshold
				KIRC				
RF	0.798	0.702	0.747	0.605	0.657	0.791	0.798	0.438
GB	0.815	0.679	0.775	0.657	0.657	0.833	0.815	0.462
MLP	0.798	0.651	0.766	0.656	0.600	0.847	0.798	0.459
LR	0.809	**0.729**	0.738	0.574	**0.771**	0.722	0.809	0.436
CatBoost	**0.826**	**0.729**	**0.785**	**0.687**	0.628	**0.861**	**0.826**	0.420
				KIRP				
RF	0.854	0.529	**0.896**	**0.666**	**0.666**	**0.938**	0.854	0.285
GB	0.753	0.503	0.879	0.625	0.555	**0.938**	0.753	0.466
MLP	0.780	0.494	0.706	0.277	0.555	0.734	0.780	0.501
LR	0.870	0.533	0.844	0.500	0.555	0.897	0.870	0.515
CatBoost	**0.879**	**0.702**	0.810	0.416	0.555	0.857	**0.879**	0.414

and RP11-169K16.8 (pseudogene) [Zhou et al. 2019]), associated with cell cycle regulation, DNA damage response, immune evasion, epigenetic modulation, and transcriptional regulation. It offers fresh insights into the molecular drivers of KIRC survival (Figure 1a).

For KIRP, ten high-importance genes were identified, with top predictors including TOP2A [Wang et al. 2024], PDIA4 [Kang et al. 2023], and ACAA1 [Feng and Shen 2020] (Figure 1b). TOP2A is a well-documented marker of cell proliferation and a target for several anti-cancer drugs. Its high importance in the survival models reinforces the hypothesis that cell cycle acceleration is a primary driver of poor outcomes in papillary carcinoma. The convergence of these genes across different interpretability techniques provides a higher degree of confidence in their biological relevance.

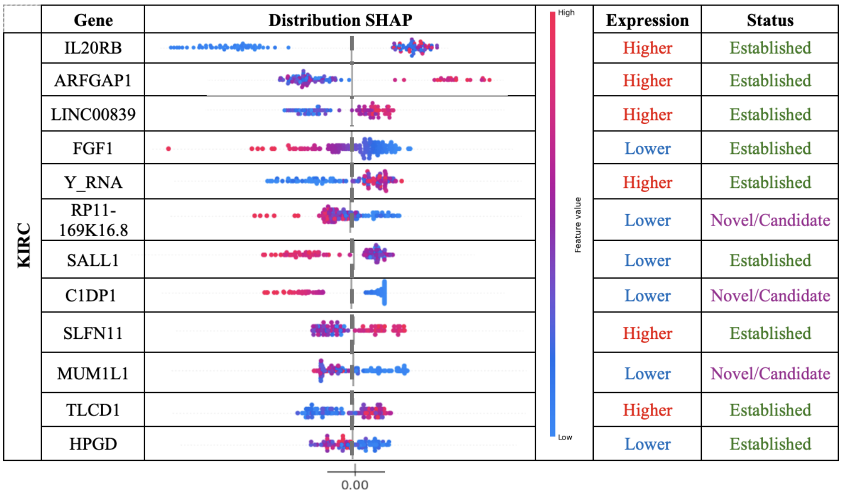
5. Conclusion

This study shows that combining high-performance ML models with XAI techniques is an effective strategy to identify prognostic biomarkers. The strong performance of CatBoost, along with the identification of established and novel biologically plausible genes, highlights the molecular heterogeneity of renal cancer and the potential of gene signatures to improve risk stratification. By leveraging XAI to interpret these biomarkers, researchers can prioritize molecular targets, supporting future therapeutic strategies and personalized treatment development.

Although the use of XAI helps mitigate the ”black-box” problem, the present study is an ongoing project and faces limitations as such. Key challenges include class imbalance, high-dimensional feature selection, and the risk of overfitting, which must be carefully controlled to ensure model reliability. Future work will address these issues and assess generalizability on external datasets. Additionally, integrating clinical and multi-omics data (e.g., proteomics and methylation) may enable more comprehensive models. Ultimately, combining experimental validation with advanced XAI approaches will be essential to translate these findings into robust clinical applications for precision oncology.

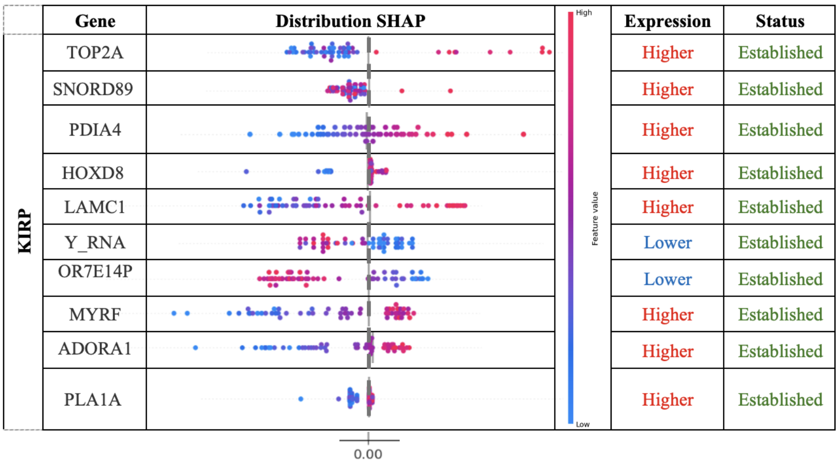
Acknowledgments

This study was partially funded by CAPES - Finance Code 001, FAPERGS [Proj. No. 21/2551-0002052-0 and No. 22/2551-0000390-7] and CNPq [Proj. No. 308075/2021-8].



	Gene	Distribution SHAP	Expression	Status
KIRC	IL20RB		Higher	Established
	ARFGAP1		Higher	Established
	LINC00839		Higher	Established
	FGF1		Lower	Established
	Y_RNA		Higher	Established
	RP11-169K16.8		Lower	Novel/Candidate
	SALL1		Lower	Established
	C1DP1		Lower	Novel/Candidate
	SLFN11		Higher	Established
	MUM1L1		Lower	Novel/Candidate
	TLCD1		Higher	Established
	HPGD		Lower	Established

(a) High importance genes for KIRC.



	Gene	Distribution SHAP	Expression	Status
KIRP	TOP2A		Higher	Established
	SNORD89		Higher	Established
	PDIA4		Higher	Established
	HOXD8		Higher	Established
	LAMC1		Higher	Established
	Y_RNA		Lower	Established
	OR7E14P		Lower	Established
	MYRF		Higher	Established
	ADORA1		Higher	Established
	PLA1A		Higher	Established

(b) High importance genes for KIRP.

Figure 1. Distribution of SHAP values for high-importance genes .Each point represents a sample. The position on the X-axis indicates the gene's contribution to the predictive model. Points to the right are associated with a worse prognosis. The colors reflect the gene expression levels.

References

Albertson, D., Barry, M., Liu, T., Mahlow, J., and Sirohi, D. (2025). Genotype phenotype correlation of renal tumors in the cancer genome atlas database. *International Journal of Surgical Pathology*, 33(2):289–301.

Feng, H. and Shen, W. (2020). ACAA1 is a predictive factor of survival and is correlated with T cell infiltration in non-small cell lung cancer. *Frontiers in Oncology*, 10:564796.

Kang, L., Wang, D., Shen, T., Liu, X., Dai, B., Zhou, D., Shen, H., Gong, J., Li, G., Hu, Y., et al. (2023). PDIA4 confers resistance to ferroptosis via induction of ATF4/SLC7A11 in renal cell carcinoma. *Cell death & disease*, 14(3):193.

Kolenda, T., Poter, P., Guglas, K., Kozłowska-Masłoń, J., Braska, A., Kazimierczak, U., and Teresiak, A. (2023). Biological role and diagnostic utility of ribosomal protein

L23a pseudogene 53 in cutaneous melanoma. *Reports of Practical Oncology and Radiotherapy*, 28(2):255–270.

Peduzzi, G., Felici, A., Pellungrini, R., and Campa, D. (2025). Explainable machine learning identifies a polygenic risk score as a key predictor of pancreatic cancer risk in the uk biobank. *Digestive and Liver Disease*, 57(4):915–922.

Seo, H., Park, J.-H., Lee, J., and Chung, B. C. (2025). Explainable ai based feature selection in cancer RNA-seq. *ICT Express*.

Shah, H. H. and Lodhi, S. K. (2025). AI in personalized medicine: Tailoring treatment plans based on individual patient data. *Global Trends in Science and Technology*, 1(1):15–29.

Sharma, V., Davies, A., and Ainsworth, J. (2021). Clinical risk prediction models: the canary in the coalmine for artificial intelligence in healthcare? *BMJ Health & Care Informatics*, 28(1):e100421.

Su, L., Hounye, A. H., Pan, Q., Miao, K., Wang, J., Hou, M., and Xiong, L. (2024). Explainable cancer factors discovery: Shapley additive explanation for machine learning models demonstrates the best practices in the case of pancreatic cancer. *Pancreatology*, 24(3):404–423.

Tabibu, S., Vinod, P., and Jawahar, C. (2019). Pan-renal cell carcinoma classification and survival prediction from histopathology images using deep learning. *Scientific reports*, 9(1):10509.

Tran, D., Nguyen, H., Pham, V.-D., Nguyen, P., Nguyen Luu, H., Minh Phan, L., Blair DeStefano, C., Jim Yeung, S.-C., and Nguyen, T. (2025). A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. *Briefings in Bioinformatics*, 26(2):bbaf150.

Wang, Q., Liu, J., Li, R., Wang, S., Xu, Y., Wang, Y., Zhang, H., Zhou, Y., Zhang, X., Chen, X., et al. (2024). Assessing the role of programmed cell death signatures and related gene TOP2A in progression and prognostic prediction of clear cell renal cell carcinoma. *Cancer Cell International*, 24(1):164.

Zhang, H., Liu, Y., Wang, B., and Wang, C. (2022). Interleukin 20 receptor subunit beta (il20rb) predicts poor prognosis and regulates immune cell infiltration in clear cell renal cell carcinoma. *BMC Genomic Data*, 23(1):58.

Zhang, L., Wu, X., Fan, X., and Ai, H. (2023). MUM1L1 as a tumor suppressor and potential biomarker in ovarian cancer: evidence from bioinformatics analysis and basic experiments. *Combinatorial Chemistry & High Throughput Screening*, 26(14):2487–2501.

Zhou, K., Li, Y., Wang, W., Chen, Y., Qian, B., Liang, Y., Li, H., Xu, R., and Zhuang, L. (2025). SLFN11: a pan-cancer biomarker for DNA-targeted drugs sensitivity and therapeutic strategy guidance. *Frontiers in Oncology*, 15:1582738.

Zhou, Y., Zheng, X., Xu, B., Hu, W., Huang, T., and Jiang, J. (2019). The identification and analysis of mRNA–lncRNA–miRNA cliques from the integrative network of ovarian cancer. *Frontiers in genetics*, 10:751.