Interpretable AI models for predicting distant metastasis development based on genetic data: kidney cancer example

Maria Boyko[1*], *Ekaterina* Antipushina[1,2,3], *Alexander* Bernstein[1,2], *Maxim* Sharaev[1,2], *Natalya* Apanovich[4], *Vsevolod* Matveev[5], *Vera* Alferova[6], and *Alexey* Matveev[5]

[1]BIMAI-Lab, Biomedically Informed Artificial Intelligence Laboratory, University of Sharjah, Sharjah, 27272, United Arab Emirates
[2]Applied AI Center, Skolkovo Institute of Science and Technology, Moscow, 121205, Russia
[3]Neuro Center, Skolkovo Institute of Science and Technology, Moscow, 121205, Russia
[4]Research Center for Medical Genetics (RCMG), Moscow, 115478, Russia
[5]National Medical Research Center of Oncology named after N.N. Blokhin, Moscow, 115522, Russia
[6]Pirogov Russian National Research Medical University, Moscow, 117513, Russia

Abstract. Kidney cancer has a high metastatic potential with up to 30% of patients developing distant metastasis after surgery. We assessed the value of AI models in predicting the metastatic potential of clear cell renal cell carcinoma (ccRCC), based on the genetic data. Tissue samples from patients with both metastatic and non-metastatic squamous cell carcinoma were analyzed, focusing on the expression and methylation levels of specific protein-coding (PC) and microRNA (miRNA) genes. Using quantitative PCR and data classification techniques, we found a correlation between metastasis and reduced expression of PC-genes CA9, NDUFA4L2, EGLN3, and BHLHE41, as well as increased methylation in miRNA genes MIR125B-1, MIR137, MIR375, MIR193A, and MIR34B. AI models were built for predicting distant metastases based on the expression values and methylation status of selected genes. One model is based on solving a regression problem and is non-interpretable, while another one is based on proposed decision rules and is interpretable. The quality of the models was assessed using sensitivity and specificity metrics, and cross-validation technology was used to ensure the reliability of the results.

1 Introduction

Advanced kidney cancer remains an incurable disease and represents a significant health challenge globally. ccRCC accounts for the majority of all RCC cases (75%) [1,2]. Despite medical advancements, the prognosis for patients with distant metastasis, remains grim, with

* Corresponding author: m.sharaev@skoltech.ru

survival rates dropping markedly over the last five years (survival rate is 12%) [1]. Recently, KEYNOTE-564 study demonstrated that adjuvant therapy with anti-PD1-inhibitor pembrolizumab for patients with ccRCC at high risk of recurrence after surgery provides statistically significant and clinically meaningful survival advantages. Although clinical parameters like T and N stage, grade and histological features help to identify the patients who are at a high risk of recurrence, multiple studies are looking for reliable molecular biomarkers for predicting metastatic potential of ccRCC.

miRNAs and gene methylation patterns, which have shown promise in elucidating the mechanisms underlying kidney cancer progression and treatment resistance [7-9]. However, the potential of these biomarkers in clinical settings has yet to be fully tapped. This study aims to bridge this gap by employing machine learning (ML) techniques to build predictive models based on the analysis of specific protein-coding and miRNA genes within kidney cancer samples. By integrating gene expression and methylation data, this research endeavors to enhance the accuracy of metastasis prediction, thus paving the way for personalized treatment approaches.

2 Materials and Methods

2.1 Samples collection

Samples of clear cell renal cell carcinoma tumors and normal tissue from the same organ obtained during surgical operations were collected and clinically characterized at the Research Institute of Clinical Oncology of N.N. Blokhin Federal State Budgetary Institution "Oncology Research Center." The study was conducted in accordance with the Declaration of Helsinki, and the research protocol was approved by the Ethics Committee (approval number 2017-4/2). After collection, the tissue was immediately frozen and stored at -70°C. A total of 80 paired tissue samples from patients with squamous cell carcinoma were examined. The sample included 31 patients with distant metastases and 49 with non-metastatic squamous cell carcinoma. The criterion for inclusion in the sample was the histologic diagnosis of renal cancer.

Tissue DNA extraction involved phenol-chloroform extraction, while RNA isolation utilized the RNeasy Mini Kit (Qiagen, USA), with RNA quality assessed via gel electrophoresis and concentration estimated spectrophotometrically (Nanodrop 1000 spectrophotometer by Thermo Fisher Scientific, USA). Gene expression analysis employed reverse transcription and real-time PCR using TaqMan® kits for CA9, NDUFA4L2, EGLN3, and BHLHE41 (Applied Biosystems, USA), with QuantStudio™ Design and Analysis Software v1.5.2 utilized for data analysis. Methylation levels of miRNA genes were assessed via quantitative methyl-specific PCR, with amplification conducted using qPCRmix-HS SYBR reagents in the Bio-Rad CFX96 Real-Time PCR Detection System (Bio-Rad, Hercules, CA, USA), and DNA conversion completeness confirmed by control locus ACTB and commercial DNA preparations.

2.2 Machine learning AI models

We focused on three primary machine learning algorithms: Logistic Regression, which models the probability of a given input belonging to a specific category; Support Vector Machine (SVM), which identifies the optimal boundary between classes; and Random Forest, an ensemble method that aggregates predictions from multiple decision trees to boost accuracy. Central to our approach was the application of GridSearchCV (a hyperparameter tuning method) from the sklearn library (algorithms from Python 3.10.9 were used), that

systematically navigates through various parameter combinations using cross-validation with an emphasis on maximizing their predictive capabilities in detecting kidney cancer metastasis. This approach ensures each model is rigorously tested and optimized across different subsets of the data, significantly improving the reliability of our predictions.

Our models' quality was evaluated using a refined set of metrics crucial for medical diagnostic accuracy: sensitivity (the ability to correctly identify positive cases), specificity (the ability to correctly identify negative cases), and ROC-AUC (which assesses the trade-off between true positive rate and false positive rate across different thresholds). These metrics were chosen for their significance in medical settings, where the balance between accurately identifying disease presence (sensitivity) and confirming its absence (specificity) is vital, and where the overall model performance (ROC-AUC) is crucial for making informed clinical decisions.

However, AI models based on previously described algorithms are non-interpretable and therefore cannot be directly used in clinical practice for making medical decisions. To enhance the interpretability of developed ML models, we applied a set of comparative rules as outlined in [10]. According to the guidelines, simultaneous occurrence of at least three markers from the protein-coding gene panel signifies a heightened risk of metastasis, especially when their expression levels fall below the threshold identified through ROC-analysis for each respective gene. Similarly, in the context of the miRNA gene panel, an elevation in methylation across four out of five genes indicates a considerable metastatic risk, as determined by the ROC thresholds established for each gene.

In scenarios where only protein-coding genes are analyzed, observing a decline in expression for three out of four designated genes correlates with an increased metastatic likelihood. Conversely, in analyses focusing solely on miRNA genes, elevated methylation levels in at least four out of five genes are associated with a heightened risk of metastasis. These determinations rely on specific ROC-analysis-derived threshold values for each gene. Furthermore, in instances where a combined gene panel is used, the identification of six or more abnormalities – whether as decreased expression levels in protein-coding genes or increased methylation levels in miRNA genes – serves as an indicator of significant metastatic potential in the tumor [10]. The predictive validity of these markers for metastasis was ascertained through evaluation of sensitivity and specificity metrics.

3 Results and Discussion

We analyzed 80 paired tissue samples from patients identified with either metastatic or non-metastatic renal cell carcinoma. The analytical strategy encompassed several machine learning algorithms, including Logistic Regression, Random Forest, and SVM, to classify the samples effectively. All three selected algorithms were applied to all group of genes (PC genes, miRNA genes, and the combined group of these genes).

The quality of the algorithms was compared using the F1-score metric calculated from the precision and recall metrics (the precision is the number of correctly identified positive results divided by the number of all positive results, including those not identified correctly, and the recall is the number of correctly identified positive results divided by the number of all samples that should have been identified as positive). The comparison results are shown in the Table 1, and the metrics were calculated for the best algorithms (shown in bold).

Table 1. F1-scores of predicting the metastases development on 10-fold cross-validation

Gene group	Algorithm	F1-score
both genes group	SVM	0.78 ± 0.17
	LogisticRegression	**0.79 ± 0.15**
	RandomForest	0.76 ± 0.18
PC genes	SVM	0.70 ± 0.12
	LogisticRegression	**0.73 ± 0.16**
	RandomForest	0.67 ± 0.17
miRNA genes	**SVM**	**0.74 ± 0.147**
	LogisticRegression	0.70 ± 0.15
	RandomForest	0.72 ± 0.15

Other metrics (Sensitivity, Specificity, and ROC-AUC) were calculated for the best algorithms, these results are presented in Table 2.

Table 2. Sensitivity, Specificity, and ROC-AUC for the best algorithm

Method	Gene group	Sensitivity	Specificity	Roc-AUC
Logistic Regression	both genes group	0.78 ± 0.17	0.87 ± 0.19	0.88 ± 0.13
Logistic Regression	PC-genes	0.74 ± 0.19	0.78 ± 0.30	0.76 ± 0.22
SVM	miRNA genes	0.74 ± 0.30	0.812+-0.19	0.81 ± 0.11

Following the established practical application guidelines [10], the detection of three or more concurrent markers from the protein-coding gene panel was indicative of increased metastatic potential. This was particularly the case when gene expression levels dropped below ROC-analysis-defined cutoffs. Similarly, an uptick in methylation across four of the five miRNA genes hinted at a strong likelihood of metastasis, aligning with predefined ROC thresholds. When applying these markers in real-world scenarios, the reduced expression of three out of four protein-coding genes, or the increased methylation in four out of five miRNA genes, underscored a significant metastasis risk. This assessment was underpinned by ROC-derived threshold values for each gene. Additionally, the application of a combined gene panel highlighted that the presence of six or more irregularities, be it diminished expression in protein-coding genes or heightened methylation in miRNA genes, flagged a tumor's increased likelihood of metastatic progression.

For these rules, their Sensitivity and Specificity were calculated, shown in Table 3, and cross-validation technology was used to ensure the reliability of the results.

Table 3. Sensitivity and Specificity metrics calculated for different genes panels

Rules	Sensitivity	Specificity
Utilizing both genes group	0.76	0.86
Utilizing PC genes	0.73	0.79
Utilizing miRNA genes	0.68	0.79

The practical utility of these gene panels in forecasting metastasis advancement with using the described rules was further validated by Sensitivity и Specificity metrics, detailed in Table 3, and cross-validation technology was used to ensure the reliability of the results. These results underscoring their significance in identifying patients at heightened risk for metastatic disease.

4 Conclusion

We identified crucial biomarkers that indicate the likelihood of metastasis. Crucially, we developed interpretable machine learning models that adhere to established medical guidelines, thereby ensuring that the analysis remains grounded in clinically relevant standards. Our methodology involved constructing a comprehensive gene panel, which includes protein-coding genes, as well as the methylation status of miRNA genes. Among the models developed, the one that utilized both protein-coding and miRNA genes stood out for its superior predictive performance. This approach not only leverages the individual strengths of each gene type but also underscores the importance of integrating diverse biological markers to enhance the accuracy and reliability of metastasis prediction. Consequently, this model represents a significant advancement in our ability to identify patients at increased risk of ccRCC metastasis. Another important result is that the transition from non-interpretable models to interpretable ones does not reduce the values of sensitivity and specificity indicators.

Acknowledgments

This research was partially supported by the Russian Science Foundation grant No. 21-71-10136 (Construction of interpretable AI models).

Authors' contribution

Selection of AI models, computational experiments and analyzing results, M.B.; collection, analysis and characterization of clinical samples, publication of articles theme, A.M.; computational experiments, manuscript preparation, E.A.; concept choice, design of the research, related to mathematical modeling, interpretation of results, preparation of the manuscript, A.B., M.Sh.; clinical interpretation of results, V.A.; literature review, data analysis, N.A., V.M.

References

1. T. Makino, S. Kadomoto, K. Izumi, A. Mizokami, Epidemiology and Prevention of Renal Cell Carcinoma. Cancers, **14(16)**, 4059 (2022). https://doi.org/10.3390/cancers14164059
2. B. Ljungberg (Chair), L. Albiges, J. Bedke, A. Bex (Vice-chair), U. Capitanio, R. H. Giles (Patient Advocate), M. Hora, T. Klatte, L. Marconi, T. Powles, A. Volpe. Guidelines associates: Y. Abu-Ghanem, R. Campi, S. Dabestani, S. Fernández-Pello Montes, F. Hofmann, T. Kuusk, R. Tahbaz. EAU Guidelines. Edn. presented at the EAU Annual Congress Milan, 2023, 100. ISBN 978-94-92671-19-6.
3. A. D. Kaprin, V. V. Starinsky, A. O. Shakhzadova Malignant neoplasms in Russia in 2020 (morbidity and mortality). P. A. Herzen MNIOI – branch of FGBU "NMRC Radiology" of the Ministry of Health of the Russian Federation, 2021, 252.
4. C.G. Ma, W.H. Xu, Y. Xu, J. Wang, W. R. Liu, D. L. Cao, H. K. Wang, G. H. Shi, Y. P. Zhu, Y. Y. Qu, H. L. Zhang, D. W. Ye. Identification and validation of novel metastasis-related signatures of clear cell renal cell carcinoma using gene expression databases. Am J Transl Res., **12(8),** 4108-4126 (2020). PMID: 32913492.
5. B. Wan, Y. Yang, Z. Zhang. Identification of differentially methylated genes associated with clear cell renal cell carcinoma and their prognostic values. J Environ Public Health, 8405945 (2023). https://doi.org/10.1155/2023/8405945 PMID: 36793506.

6. T. Zhong, Z. Jiang, X. Wang, H. Wang, M. Song, W. Chen, S. Yang, Key genes associated with prognosis and metastasis of clear cell renal cell carcinoma. PeerJ, 2022, 10: e12493. https://doi.org/10.7717/peerj.12493. PMID: 35036081.
7. G. Outeiro-Pinho, D. Barros-Silva, E. Aznar, A. I. Sousa, M. Vieira-Coimbra, J. Oliveira, C. S. Gonçalves, B. M. Costa, K. Junker, R. Henrique, C. Jerónimo, MicroRNA-30a-5pme: a novel diagnostic and prognostic biomarker for clear cell renal cell carcinoma in tissue and urine samples. J Exp Clin Cancer Res., **39(1)**, 98 (2020). https://doi.org/10.1186/s13046-020-01600-3. Erratum in: J Exp Clin Cancer Res., **41(1)**, 247 (2022). PMID: 32487203; PMCID: PMC7323611.
8. S. Grammatikaki, H. Katifelis, A. A. Farooqi, K. Stravodimos, M. V. Karamouzis, K. Souliotis, D. Varvaras, M. Gazouli, An overview of epigenetics in clear cell renal cell carcinoma. In Vivo, **37(1)**, 1-10 (2023) https://doi.org/10.21873/invivo.13049. PMID: 36593023.
9. N. Patil, M. L. Abba, C. Zhou, S. Chang, T. Gaiser, J. H. Leupold, H. Allgayer, Changes in methylation across structural and microRNA genes relevant for progression and metastasis in colorectal cancer. Cancers, **13**, 5951 (2021). https://doi.org/10.3390/cancers13235951
10. N. Apanovich, A. Matveev, N. Ivanova, A. Burdennyy, P. Apanovich, I. Pronina, E. Filippova, T. Kazubskaya, V. Loginov, E. Braga, A. Alimov, Prediction of distant metastases in patients with kidney cancer based on gene expression and methylation analysis. Diagnostics, **13**. 2289 (2023). https://doi.org/10.3390/diagnostics13132289