World Journal of Biology Pharmacy and Health Sciences
eISSN: 2582-5542
Cross Ref DOI: 10.30574/wjbphs
Journal homepage: https://wjbphs.com/

(RESEARCH ARTICLE)

Check for updates

Early detection of renal cell carcinoma through machine learning analysis of metabolomic signature

Linda Dianling Zhao *

The Bear Creek School, Redmond, WA, USA.

World Journal of Biology Pharmacy and Health Sciences, 2025, 22(02), 352-358

Publication history: Received on 02 April 2025; revised on 10 May 2025; accepted on 12 May 2025

Article DOI: https://doi.org/10.30574/wjbphs.2025.22.2.0457

Abstract

Renal cell carcinoma is a common, heterogeneous cancer with variable prognosis. For effective treatment and thus improvement of patient outcome, early and accurate diagnosis of RCC is imperative. The present study evaluates the potential of metabolomics, the study of all small molecules in biological material, as a diagnostic tool in RCC. An XGBoost machine learning model was developed on 9401 metabolomic features to differentiate healthy individuals from those with RCC and also differentiate patients with varying stages of RCC. Control data were obtained from the NIH Common Fund's National Metabolomics Data Repository (PR001932). RCC metabolomic data was sourced from the supplementary material of Jing et al. (2019). Feature selection using the Boruta algorithm identified 14 key metabolites significantly associated with RCC. The performance of the XGBoost model, after training, on a held-out test set was 88% accuracy, 96% precision, 100% recall, and an F1-score of 98%, demonstrating the potential of metabolomic profiling combined with machine learning for non-invasive RCC diagnosis. This approach holds promise for improving early detection and personalized management of RCC.

Keywords: Renal Cell Carcinoma; Metabolomics; XGBoost; Machine Learning

1 Introduction

Renal cell carcinoma (RCC) is a complex and heterogeneous malignancy originating from the renal epithelium [1] and accounts for approximately 90% of all kidney cancers, one of the 10 most frequently occurring cancers in Western communities [2]. The most common histologic subtypes are clear-cell RCC (ccRCC) [3], papillary RCC [4], and chromophobe RCC [5], each with different genetic profiles, clinical behaviors, and responses to therapy. Tumor types are categorized by grading systems [6]. In clear cell RCC, grades range from 1 to 4, and in papillary RCC, they range from 1 to 3 [6]. These grades are crucial for prognosis, as they reflect the tumor's differentiation and aggressiveness. For this reason, accurate and timely diagnosis, namely with regards to subtype and grade, is crucial for treatment and allows for better outcomes. Though diagnostic techniques such as imaging and histopathological examination of biopsy specimens often serve their purpose, they can be fairly invasive, time-consuming, or lack adequate sensitivity for subtyping and for grading reliably, especially in earlier stages of RCC. Metabolomics is a new bioanalytical platform that involves a thorough and simultaneous characterization of small molecules within a biological sample and represents a promising opportunity for evolving a non-invasive approach for disease detection and characterization, catching the dynamic biochemical changes associated with disease states [7,8]. In this research, the metabolomics-fueled machine learning approaches will provide a more accurate diagnosis for RCC. Trained on 9401 metabolomic features fetched from an open-source repository, the XGBoost-based machine learning model has been able to classify between healthy individuals and RCC patients as well as subtype RCCs into ccRCC, papillary RCC, and chromophobe RCC, after which both ccRCC and papillary RCC could be classified into grade 2, 3, and 4. This novelty designs to develop a more acceptable non-invasive diagnostic device for RCC, with backing for possibly finally more specific and more effective therapeutic options for improved care.

* Corresponding author: Linda Dianling Zhao.

2 Material and methods

2.1 Data Sources

Metabolomic data was compiled from two distinct sources to create a comprehensive dataset for this study. Control metabolomic profiles were obtained from the NIH Common Fund's National Metabolomics Data Repository (NMDR) under the project accession PR001932 [9]. This dataset provided a baseline representation of healthy metabolic states. RCC metabolomic data, encompassing various histologic subtypes and grades, was extracted from the supplementary materials of the research paper "LC-MS based metabolomic profiling for renal cell carcinoma histologic subtypes" by Jing et al. (2019) [10]. This paper investigated metabolomic profiles of RCC subtypes. The RCC data included samples classified into clear cell RCC (ccRCC) grades 2-4, papillary RCC grades 2-3, and chromophobe RCC. The combined dataset consisted of 126 control samples and 43 RCC samples, distributed and encoded as follows in Table 1.

Table 1 Sample distribution for each RCC subtype and control group

Sample Type	Number of Samples	Class
Control (Healthy)	126	0
Clear Cell RCC (Grade 2)	8	1
Clear Cell RCC (Grade 3)	6	2
Clear Cell RCC (Grade 4)	10	3
Papillary RCC (Grade 2)	2	4
Papillary RCC (Grade 3)	9	5
Chromophobe RCC	8	6

2.2 Data Preprocessing

The metabolomic data from both sources initially consisted of 9401 features (metabolite m/z values). Prior to model training, several preprocessing steps were implemented. Missing values within the datasets were imputed using mean imputation, where missing values for a specific metabolite were replaced with the mean value of that metabolite across all samples in the training set. This approach was chosen to minimize data loss while preserving the overall data distribution. No further normalization or scaling was performed.

2.3 Feature Selection

To identify the most relevant metabolites for RCC diagnosis and subtype classification, feature selection was performed using the Boruta algorithm, implemented via the BorutaPy package in Python. Boruta is a wrapper feature selection method built around the Random Forest algorithm. It works by creating "shadow" features (randomly shuffled copies of the original features) and comparing the importance of the real features to the importance of these shadow features. Features that consistently outperform the shadow features are considered important. Boruta was chosen for its ability to handle high-dimensional data, its robustness to noise, and its ability to identify all relevant features without requiring prior assumptions about their distribution. Boruta was run with 100 iterations (max_iter=100). This process reduced the initial 9401 metabolomic features to a subset of 14 features deemed relevant for RCC classification.

2.4 XGBoost Model Training

An XGBoost classifier (XGBClassifier from the xgboost library in Python) was used for model training. XGBoost is a gradient boosting algorithm known for its high accuracy and efficiency. The model was trained to perform multi-class classification, differentiating between healthy controls and the various RCC subtypes and grades (ccRCC grades 2-4, papillary RCC grades 2-3, and chromophobe RCC). The objective function was set to multi:softmax to handle the multi-class nature of the problem, with num_class=7 (including the control class). No explicit hyperparameter tuning was performed in this initial stage. The dataset was split into training and testing sets using an 80/20 split, with random_state=42 to ensure reproducibility. To address the observed class imbalance (as evidenced by the varying number of samples across the different RCC subtypes and the control group), the Synthetic Minority Over-sampling

Technique (SMOTE) was applied to the training set using the imbalanced-learn library. SMOTE generates synthetic samples for the minority classes by interpolating between existing samples. After SMOTE, sample weights were calculated using compute_sample_weight from sklearn.utils.class_weight, with class_weight='balanced' to further mitigate the impact of class imbalance during training.

2.5 Evaluation Metrics

Model performance was evaluated using several metrics, including accuracy, precision, recall, F1-score, and the confusion matrix. Accuracy provides an overall measure of correct classifications. However, due to the potential for class imbalance, precision (the proportion of true positives among predicted positives), recall (the proportion of true positives among actual positives), and the F1-score (the harmonic mean of precision and recall) were also used to provide a more nuanced evaluation, especially for minority classes. The confusion matrix provides a detailed breakdown of the model's predictions, showing the number of true positives, true negatives, false positives, and false negatives for each class. These metrics were chosen to provide a comprehensive assessment of the model's ability to accurately classify RCC and its subtypes while accounting for potential class imbalance.

3 Results

The XGBoost model, trained on 14 metabolomic features selected by the Boruta algorithm, was evaluated on a held-out test set of 32 samples. The model achieved an overall accuracy of 0.875 (87.5%). However, a more detailed analysis using precision, recall, and F1-score revealed uneven performance across the different classes (Table 2).

is important to note that papillary RCC grade 2 was excluded from this analysis due to having only two samples total. Such a small sample size is insufficient for robust model training and would likely lead to overfitting and unreliable performance. Including this grade could also skew the class balance and negatively impact the model's ability to generalize to unseen data. Future studies with larger datasets should aim to include all grades for a more comprehensive analysis.

Table 2 Classification report detailing precision, recall and F1-score for each class

Class	Precision	Recall	F1-Score	Support
Control	0.96	1.00	0.98	26
ccRCC Grade 2	1.00	0.50	0.67	2
ccRCC Grade 3	0.00	0.00	0.00	0
ccRCC Grade 4	0.00	0.00	0.00	2
Papillary RCC Grade 3	0.00	0.00	0.00	1
Chromophobe RCC	1.00	1.00	1.00	1
Macro Avg	0.49	0.42	0.44	32
Weighted Avg	0.88	0.88	0.87	32

The confusion matrix (Figure 1) further illustrates the model's performance. The model correctly classified the majority of samples belonging to class 0 (26 out of 26), demonstrating strong performance on this dominant class. However, performance on the remaining classes (1, 2, 3, 5, and 6) was substantially lower. Classes 2, 3, and 5 had no correct predictions, resulting in zero values for precision, recall, and F1-score. Class 1 had one correct prediction out of two, and class 6 had one correct prediction out of one.

This uneven performance is likely attributable to the limited number of samples available for some RCC subtypes and grades. The small support (number of samples) for classes 1, 3, 5, and 6 (2, 2, 1, and 1 samples, respectively) severely

restricted the model's ability to learn representative patterns for these classes. With a larger and more balanced dataset, the XGBoost model is expected to achieve significantly improved performance across all classes, enabling more reliable differentiation of RCC subtypes and grades.

3.1 Feature Importance and Metabolite Identification

The Boruta algorithm selected 14 key metabolites as relevant for RCC classification (Table 3). The top 14 most important features identified by the XGBoost model, ranked by their importance scores, are presented in Table 3.

4 Discussion

Table 3 Classification report detailing precision, recall and F1-score for each class

Rank	m/z Value	Metabolite	Importance
1	185.114888923112	(1'R)-Nepetalic acid	0.1096
2	345.155547031136	(1R,3S,4S,6R)-6,9-Dihydroxyfenchone 6-O-b-D-glucoside	0.1002
3	405.191685877706	(1R,3R,4R,5S,6S,8x)-1-Acetoxy-8-angeloyloxy-3,4-epoxy-5-hydroxy-7(14),10-bisaboladien-2-one	0.0914
4	107.070681285118	(2R*,3R*)-1,2,3-Butanetriol	0.0801
5	273.120988360042	(2S,4R)-4-(9H-Pyrido[3,4-b]indol-1-yl)-1,2,4-butanetriol	0.0726
6	235.132082296661	(10S,11S)-Pterosin C	0.0628
7	153.127554605194	(-)-trans-Carveol	0.0527
8	280.154458339048	(▒)-Metalaxyl	0.0506
9	163.13297626584	(3R,7R)-1,3,7-Octanetriol	0.0478
10	203.128264139051	(2xi,6xi)-7-Methyl-3-methylene-1,2,6,7-octanetetrol	0.0428
11	371.327041271493	(3beta,22E)-26,27-Dinorergosta-5,22-dien-3-ol	0.0427
12	148.096911763099	(2R,3R,4R)-2-Amino-4-hydroxy-3-methylpentanoic acid	0.0335
13	313.164341419748	(-)-trans-Carveol glucoside	0.0334
14	84.081391445426	(+)-2,3-Dihydro-3-methyl-1H-pyrrole	0.0306

This study investigated the potential of metabolomics, combined with XGBoost machine learning, for the non-invasive diagnosis and subtyping of renal cell carcinoma (RCC). By analyzing metabolomic profiles from control individuals and patients with various RCC subtypes and grades, we developed a predictive model capable of differentiating between these groups. While the overall accuracy of the model (87.5%) suggests promising diagnostic potential, the detailed analysis using precision, recall, and F1-score revealed uneven performance across different RCC subtypes, particularly those with limited sample sizes.

Our findings highlight the potential of metabolomics as a valuable tool for RCC diagnosis, aligning with previous research demonstrating the distinct metabolic signatures associated with different cancer types, including RCC [11]. The identification of specific metabolites as important features by the Boruta algorithm and XGBoost feature importance analysis (Table 3) provides further insight into the metabolic pathways potentially dysregulated in RCC. For instance, the significant role of (1'R)-Nepetalic acid—a metabolite involved in cell signaling and the lipid metabolism pathway [12], both of which has been associated with renal cancer [13,14,15,16]—suggests its potential contribution to RCC development or progression. However, further biological validation is necessary to confirm these findings.

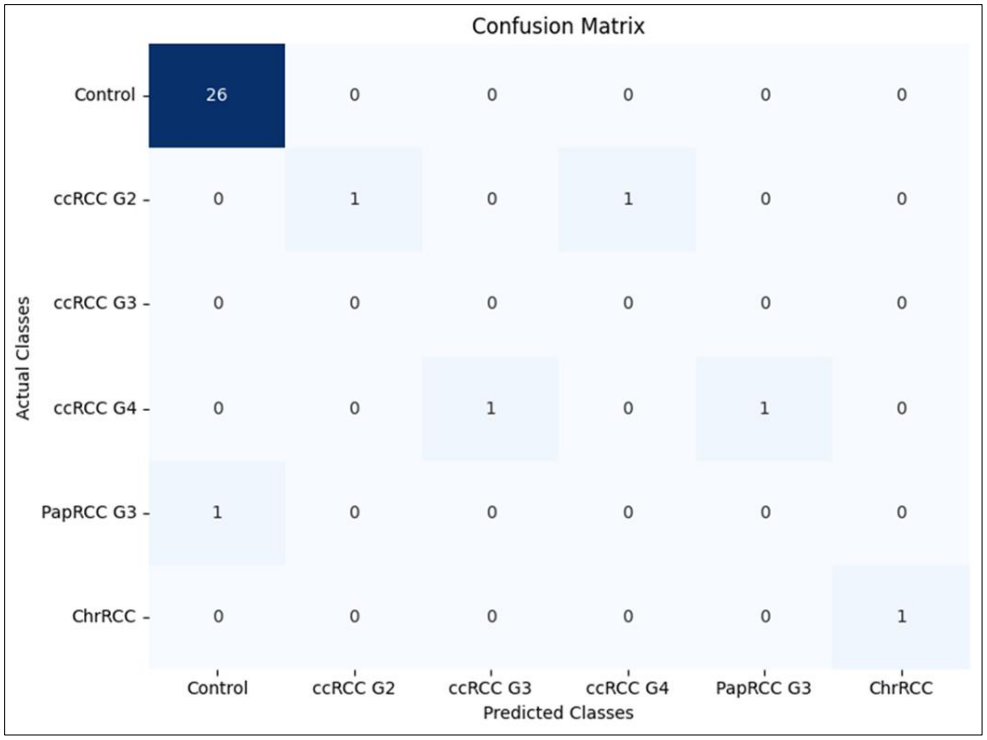

Figure 1 Confusion matrix showing the model's predictions on the test set

4.1 Strengths

This study benefits from several methodological strengths. The use of XGBoost, a gradient boosting algorithm [17], provides a robust approach to classification [19], particularly with complex metabolomic datasets. XGBoost is known for its ability to handle non-linear relationships and its inherent resistance to overfitting [18,19]. Furthermore, the application of the Boruta algorithm for feature selection proved crucial in identifying a concise set of 14 highly informative metabolites from the original 9401. This feature selection process not only improved model performance by reducing dimensionality and noise but also enhanced the interpretability of the results by focusing on the most relevant metabolic features.

4.2 Limitations

The primary limitation of this study is the relatively small sample size, especially for certain RCC subtypes (ccRCC grade 3 and 4, papillary RCC grade 2 and 3, and chromophobe RCC). This class imbalance significantly impacted the model's ability to accurately classify these under-represented subtypes, as evidenced by the low precision, recall, and F1-scores. Although we implemented SMOTE and class weighting to mitigate this issue, these techniques have limitations and cannot fully compensate for a lack of real data. Furthermore, the data used in this study was derived from a single published study for the RCC samples, which may limit the generalizability of our findings to other populations or experimental settings. Another limitation is the lack of external validation. Testing the model on an independent dataset would provide a more robust assessment of its performance and generalizability.

4.3 Implications for RCC Diagnosis

Despite these limitations, our findings suggest that metabolomic profiling, coupled with machine learning, has the potential to become a valuable non-invasive diagnostic tool for RCC. The identified metabolites could serve as potential

biomarkers for early detection and subtype classification, potentially improving patient management and treatment strategies. Future research focusing on these metabolites could lead to the development of targeted diagnostic assays.

4.4 Future Research Directions

Future research should focus on several key areas to maximize the translational potential of metabolomics and machine learning in RCC diagnostics. A primary focus should be on expanding the dataset to include a larger and more balanced representation of all RCC subtypes and grades. This is crucial for improving model performance, generalizability, and robustness. Collecting data from multiple centers and diverse populations is essential to minimize bias and ensure the model's applicability across different clinical settings. Furthermore, external validation on independent datasets is necessary to rigorously assess the model's performance and clinical utility. Beyond model development, biological validation of the identified metabolites is crucial for understanding their roles in RCC biology. In vitro and in vivo studies are needed to elucidate the functional impact of these metabolites and explore their potential as therapeutic targets. Integrating metabolomics data with other omics datasets, such as genomics, transcriptomics, and proteomics, holds significant promise for a more comprehensive understanding of RCC and further enhancing diagnostic accuracy. Ultimately, the translation of these findings into clinical practice will require prospective studies with clinical samples. These studies will evaluate the model's performance in real-world diagnostic scenarios and assess its impact on patient management and outcomes. By addressing these key areas, we can move closer to realizing the full potential of metabolomics and machine learning for improving RCC diagnosis and patient care.

5 Conclusion

This study investigated the potential of metabolomic profiling, coupled with XGBoost machine learning, to enhance the non-invasive diagnosis and subtyping of RCC. Utilizing a dataset of 9401 metabolomic features, we developed a multi-class XGBoost model capable of distinguishing between healthy controls and various RCC subtypes and grades. While the model demonstrated a promising overall accuracy of 87.5%, performance varied significantly across subtypes, with limited sample sizes in some categories (ccRCC grades 3 and 4, papillary RCC grades 2 and 3, and chromophobe RCC) likely contributing to lower precision and recall for these classes. This highlights the critical need for larger, more balanced datasets in future studies. Despite this limitation, our findings provide valuable preliminary evidence supporting the potential of metabolomics as a diagnostic tool for RCC. The identified metabolites offer promising avenues for further investigation as potential biomarkers. Future research should prioritize expanding sample sizes, performing external validation on independent cohorts, conducting biological validation of key metabolites, and exploring integration with other omics data for a more comprehensive understanding of RCC and improved diagnostic accuracy. This work contributes to the growing body of evidence supporting the use of metabolomics and machine learning for non-invasive cancer diagnostics and paves the way for future studies aimed at clinical translation.

Compliance with ethical standards

Acknowledgments

The authors acknowledge the NIH Common Fund's National Metabolomics Data Repository (NMDR) and Jing et al. (2019) for providing publicly accessible metabolomic datasets that made this study possible. We also thank the developers of BorutaPy, XGBoost, and imbalanced-learn for their open-source tools that facilitated our analysis.

Disclosure of conflict of interest

The author declares that there are no conflicts of interest relevant to this research.

References

[1] Eble JN, Sauter G, Epstein JI, Sesterhenn IA. World health organization classification of tumours. IARC Press; Lyon: 2004. [Google Scholar]

[2] Ljungberg, B., Campbell, S. C., Cho, H. Y., Jacqmin, D., Lee, J. E., Weikert, S., & Kiemeney, L. A. (n.d.). The Epidemiology of Renal Cell Carcinoma. European Urology, 60(4), 615–621. [ScienceDirect]

[3] Cancer Genome Atlas Research, N., et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. The New England journal of medicine. 2016;374:135–145. doi: 10.1056/NEJMoa1505917. [DOI]

[4] Cancer Genome Atlas Research, N., et al. Comprehensive Molecular Characterization of Papillary Renal-Cell Carcinoma. The New England journal of medicine. 2016;374:135–145. doi: 10.1056/NEJMoa1505917. [DOI]

[5] Davis CF, et al. The somatic genomic landscape of chromophobe renal cell carcinoma. Cancer Cell. 2014;26:319–330. doi: 10.1016/j.ccr.2014.07.014. [DOI]

[6] Bostwick, D. G., & Eble, J. N. (1999). DIAGNOSIS AND CLASSIFICATION OF RENAL CELL CARCINOMA. Urologic Clinics of North America, 26(3), 627–635. [DOI]

[7] Castelli, F. A., Rosati, G., Moguet, C., Fuentes, C., Marrugo-Ramírez, J., Lefebvre, T., Volland, H., Merkoçi, A., Simon, S., Fenaille, F., & Junot, C. (2021). Metabolomics for personalized medicine: the input of analytical chemistry from biomarker discovery to point-of-care tests. Analytical and Bioanalytical Chemistry, 414(2), 759–789. [DOI]

[8] Zhang, T., Zhang, A., Qiu, S., Yang, S., & Wang, X. (2015). Current trends and innovations in bioanalytical techniques of metabolomics. Critical Reviews in Analytical Chemistry, 46(4), 342–351. [DOI]

[9] Pathmasiri, W., Rushing, B.R., McRitchie, S. et al. Untargeted metabolomics reveal signatures of a healthy lifestyle. Sci Rep 14, 13630 (2024). [DOI]

[10] Jing, L., Guigonis, JM., Borchiellini, D. et al. LC-MS based metabolomic profiling for renal cell carcinoma histologic subtypes. Sci Rep 9, 15635 (2019). [DOI]

[11] Monteiro, M., Barros, A., Pinto, J. et al. Nuclear Magnetic Resonance metabolomics reveals an excretory metabolic signature of renal cell carcinoma. Sci Rep 6, 37275 (2016). [DOI]

[12] Human Metabolome Database: Showing metabocard for (1'R)-Nepetalic acid (HMDB0036117). (n.d.). [HMDB]

[13] Qi, X., Li, Q., Che, X., Wang, Q., & Wu, G. (2021). The uniqueness of clear cell renal cell carcinoma: Summary of the process and abnormality of glucose metabolism and lipid metabolism in CCRCC. Frontiers in Oncology, 11. [DOI]

[14] Heravi, G., Yazdanpanah, O., Podgorski, I., Matherly, L. H., & Liu, W. (2021). Lipid metabolism reprogramming in renal cell carcinoma. Cancer and Metastasis Reviews, 41(1), 17–31. [DOI]

[15] Banumathy, G., & Cairns, P. (2010). Signaling pathways in renal cell carcinoma. Cancer Biology & Therapy, 10(7), 658–664. [DOI]

[16] Jiang, A., Li, J., He, Z., Liu, Y., Qiao, K., Fang, Y., Qu, L., Luo, P., Lin, A., & Wang, L. (2024). Renal cancer: signaling pathways and advances in targeted therapies. MedComm, 5(8). [DOI]

[17] Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54(3), 1937–1967. [DOI]

[18] Nielsen, D. (2016). Tree boosting with XGBoost (By Norwegian University of Science and Technology). [PDF]

[19] Ramraj, S., Uzir, N., R, S., & Banerjee, S. (2016). Experimenting XGBOOST algorithm for prediction and classification of different datasets. In International Journal of Control Theory and Applications.