Journal of Health Informatics in Africa

A Tree-based Machine Learning Approach for Precise Renal Cell Carcinoma Subtyping using RNA-seq Gene Expression Data

Oluwafemi Ogundare

Department of Medicine & Surgery, Faculty of Clinical Sciences, College of Medicine, University of Ibadan, Oyo State, Nigeria

Background and Purpose: Renal cell carcinoma (RCC) is a malignant neoplasm of the kidneys, characterized by distinct molecular and histological subtypes. Accurate subtyping is crucial for personalized treatment and improved patient outcomes. High-throughput sequencing has enabled precise gene expression profiling for cancer classification. This study compares tree-based and non-tree-based machine learning algorithms for differentiating between gene expression profiles of chromophobe, clear cell, and papillary RCCs.
Methods: RNA-seq data from a diverse cohort of patients diagnosed with these three cancer subtypes was used. Data preprocessing and normalization were performed, followed by feature selection using Analysis of Variance (ANOVA). Tree-based and non-tree-based algorithms were trained on the preprocessed data. The tree-based algorithms included decision tree, random forest, extra trees classifier, and bagging classifier. The non-tree-based algorithms included logistic regression, support vector machine, and naive bayes. Each algorithm was evaluated using sensitivity, specificity, F1 score, and AUC.
Results: Tree-based algorithms demonstrated superior performance across all evaluation metrics compared to non-tree-based algorithms. Specifically, the random forest classifier achieved the highest specificity and F1 score, the decision tree classifier achieved the highest sensitivity, while the bagging classifier achieved the highest AUC score. In contrast, non-tree-based algorithms showed comparatively lower performance in distinguishing between the cancer subtypes.
Conclusions: This study demonstrates the potential of machine learning, particularly tree-based models, for precise RCC subtyping. By leveraging tree-based models, we can effectively capture the complex, non-linear patterns in gene expression datasets. Future studies should aim to validate these findings across larger and more diverse datasets of RCC subtypes.

Keywords: Machine Learning, Renal Cancer, Gene Expression, Oncology.

1 Introduction

Kidney cancer is the seventh most common cause of cancer globally, and its prevalence is on the rise [1]. Renal cell carcinoma (RCC) accounts for more than 90% of all renal malignancies and is the most frequent malignant tumor of the kidney [2]. According to the International Agency for Research on Cancer, over 400,000 new cases of RCC are diagnosed each year, with more than 170,000 deaths globally [3]. The three most frequent histological subtypes of RCC are clear cell, papillary, and chromophobe, which account for more than 90% of all RCCs [4]. Accurate subtyping of RCC is crucial, as these subtypes greatly influence treatment and prognosis of these tumors [4] [5].

In recent years, RNA-seq gene expression platforms [6] [7] have emerged as the preferred method for simultaneous gene expression quantification when compared to DNA microarrays [8] [9]. Gene expression data from RNA-seq provide useful information on the differential activation of genes involved in cancer development [10]. Because cancer is a complex disease with several genetic changes, analysing gene expression data from tumor samples allows for the study of the molecular factors influencing disease

*Corresponding author address: Department of Medicine & Surgery, Faculty of Clinical Sciences, College of Medicine, University of Ibadan, 200005, Ibadan, Nigeria. Email: femiogundare001@gmail.com
 J Health Inform Afr. 2025;12(1):39-52. DOI: 10.12856/JHIA-2025-v12-i1-533

progression and patient outcomes [11]. By effectively extracting information from RNA-seq data, physicians can gain a more comprehensive molecular view of a patient's condition, potentially leading to more precise diagnostic and prognosis procedures [12].

While RNA-seq gene expression data has significantly improved cancer classification, it does have limitations, particularly due to its typically small sample size [13]. Moreover, these samples often include numerous genes that are uninformative, which can negatively impact the performance of classification algorithms [13] [14]. To tackle the challenge of high dimensionality, it is essential to perform gene selection by eliminating redundant and uninformative genes [15]. One strategy to address this is to first apply filtration and feature selection techniques before proceeding with model development [14] [16].

Researchers have employed various machine learning methods, including supervised and unsupervised learning, and deep learning, to classify cancers using gene expression profiles. In a study by Mohammed et al. [14], a novel deep learning architecture was developed by stacking the outputs of five one-dimensional convolutional neural networks (1D-CNNs) to a feedforward neural network. This model was used to classify five of the most commonly diagnosed cancers in women. The results of this study suggested that this model could potentially enhance early cancer detection and diagnosis in women, as well as inform the design of early treatment strategies to improve survival.

Divate et al. [17] used transcriptomic data from 37 different cancer types sourced from The Cancer Genome Atlas (TCGA) to develop a deep neural network for identifying cancer-specific gene expression signatures based on tissue of origin. Their model successfully identified 976 genes capable of classifying various cancer types with >97% accuracy. In a different study, Abdelwahab et al. [18] employed a hybrid approach combining mutual information and recursive feature elimination methods with a support vector classifier model. They also incorporated a random forest model as an embedded feature selection technique. This strategy led to the identification of 12 candidate biomarkers strongly associated with different lung cancer types, especially lung adenocarcinoma.

Marostica et al. [19] used whole-slide histopathology images and demographic, genomic, and clinical data from multiple sources to develop convolutional neural networks for diagnosing renal cancers and linking quantitative pathology patterns with genomic profiles and prognoses of patients. Their deep learning models successfully detected histological subtypes of RCC, predicted survival outcomes for stage I clear cell RCC patients, and identified patterns in histopathology images indicative of copy-number alterations and tumor mutation burden. Their findings demonstrated that machine learning and deep learning techniques can effectively extract clinically relevant signals from histologic and genomic data, potentially aiding in patient diagnosis, prognosis, and identification of important genomic variations.

In this study, we demonstrate that tree-based models comparatively outperform non-tree-based models in differentiating between the gene expression profiles of chromophobe, clear cell, and papillary RCCs. Findings from this study could inform future research aimed at the early detection and accurate classification of these cancer subtypes.

2 Materials and methods

For this study, RNA-seq gene expression data from Pan-Cancer Atlas (https://portal.gdc.cancer.gov/) was used [20]. The data was retrieved using R statistical software version 4.3.1 via the TCGAbiolinks package accessible on Bioconductor [21] [22]. The data consisted of 897 samples representing chromophobe, clear cell, and papillary RCCs. Subsequently, seven machine learning algorithms were trained to discriminate between these cancer subtypes based on their unique molecular signatures. The data download was performed using R software [21], while the machine learning methods were implemented using Python [23] and scikit-learn [24].

2.1 Datasets

We downloaded the dataset from the Pan-Cancer Atlas using the GDCquery function in the TCGAbiolinks package, available on Bioconductor [22]. The dataset included samples of chromophobe, clear cell, and papillary RCCs, with benign tumor cases excluded. The GDCquery function was configured with specific parameters to retrieve the desired data. We used the project codes TCGA-KICH, TCGA-KIRC, and TCGA-KIRP to obtain data for the three RCC subtypes. Data filtering was done by setting the data.category to

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

"Transcriptome Profiling" and the data.type to "Gene Expression Quantification." We restricted our query to The Cancer Genome Atlas files by using the barcode "TCGA*." Additionally, we specified "STAR – Counts" for the workflow.type and "Primary Tumor" for the sample.type. The resulting dataset was structured as a matrix, with samples represented in columns and genes in rows. In total, we obtained 897 samples and 60,660 genes. To address the issue of high dimensionality, we filtered out non-informative genes and performed feature selection to identify the most relevant genes for enhancing the performance of our machine learning models. Table 1 below shows a summary of the downloaded dataset, including the training and testing fractions for each tumor class.

Table 1. Number of samples across each cancer subtype

Cancer subtype	Number of samples	Training set (≈80%)	Test set (≈20%)
Chromophobe	66	53	13
Clear cell	541	427	114
Papillary	290	237	53
Total	897	717	180

2.2 Data Pre-processing

For the data preprocessing phase of our study, we employed the TCGAanalyze_Preprocessing function from the TCGAbiolinks package [22]. This function implements an array-array intensity correlation (AAIC) method to construct an N × N square symmetric matrix, where N represents the number of samples. Each element of this matrix corresponds to the Pearson correlation coefficient between a pair of samples. As illustrated in Figure 1, AAIC identifies samples with low correlation that could potentially be removed as outliers. In our analysis, no outliers were detected, suggesting a high degree of consistency across our dataset.

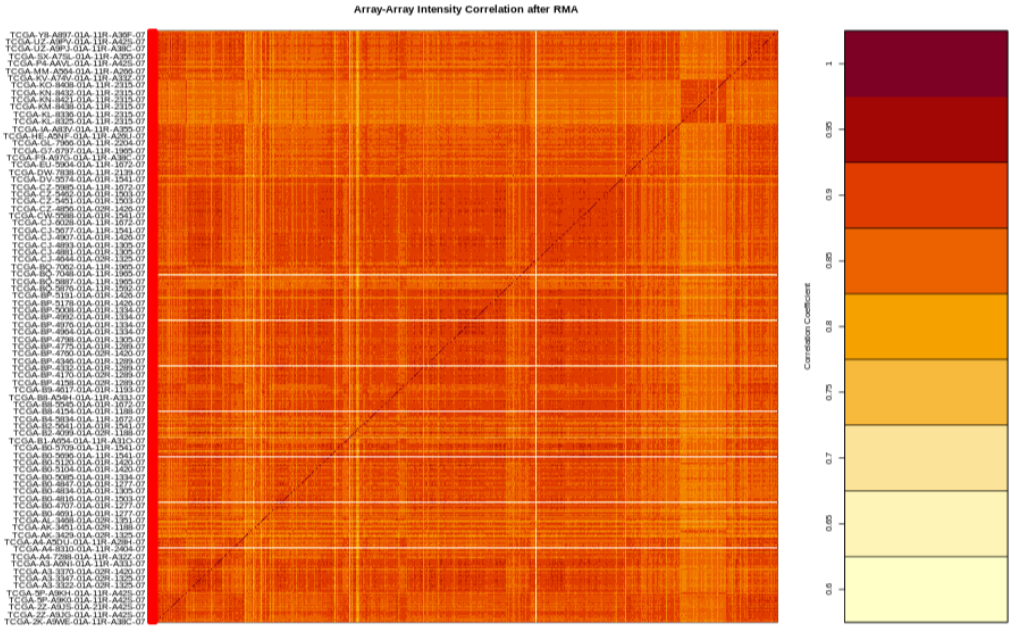


Figure 1. Array-array intensity correlation (AAIC) matrix defining the Pearson correlation coefficients among the samples.

Following this step, we conducted filtration of the expression matrix. Genes with a transcripts-per-million (TPM) value greater than 2000 were excluded from the matrix, as they were considered to be overly expressed and possibly indicative of housekeeping genes. Likewise, genes with a TPM value less than 100

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

were also removed from the matrix as they were regarded as underexpressed. The resulting matrix comprised 8,540 genes, indicating that a total of 52,120 genes were filtered out.

2.3 Feature Selection

ANOVA was further employed to reduce the number of genes. ANOVA evaluates the relationship between gene expression levels and cancer subtypes by comparing group means and calculating F-statistics. Genes with higher F-statistics are considered more relevant for distinguishing between the cancer subtypes, helping to identify the most informative genes in the expression dataset.

2.4 Data Splitting

The dataset was divided into an 80% training set and a 20% test set. We performed a 10-fold stratified cross-validation on the training set, where each fold served as a validation set while the remaining nine folds were used for training. This process was repeated ten times, with a different fold acting as the validation set in each iteration. The optimal hyperparameters for the models were determined based on this cross-validation. The test set was then used for the independent evaluation of the machine learning algorithms.

2.5 Machine Learning Algorithms

Seven machine learning algorithms were used: logistic regression, naive bayes, support vector machine, decision tree, random forest, extra trees classifier, and bagging classifier.

2.5.1 Logistic Regression.

Logistic regression is a statistical model used for binary classification tasks, where the goal is to predict one of two possible outcomes based on input features [25]. In our case, we want to classify samples into three categories: chromophobe, clear cell, and papillary RCCs. To adapt logistic regression for multi-class classification, we can use the "one-vs-all" (also known as "one-vs-rest") approach.

Let's denote the following:

- X as the input feature matrix with *n* samples (rows) and *m* genes (columns).
- Y as the output variable representing the class labels. In this case, Y will have three categories: chromophobe ($Y = 1$), clear cell ($Y = 2$), and papillary ($Y = 3$).

The probability that a sample belongs to a particular class is modeled using the sigmoid function ($\sigma(z)$):

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad \text{(i)}$$

Where Z is the linear combination of the input features and model parameters (θ):

$$z = \theta_0 + \theta_1 \cdot X_1 + \theta_2 \cdot X_2 + \cdots + \theta_m \cdot X_m \quad \text{(ii)}$$

And where, $X_1, X_2, \ldots, X_m$ are the gene expression values for a particular sample, and $\theta_0, \theta_1, \theta_2, \ldots, \theta_m$ are the model parameters to be learned from the data.

For the three-class classification, three separate logistic regression models are created, one for each class. The models are trained as follows:

Chromophobe RCC ($Y = 1$):

- Set Y^i to 1 if the sample is chromophobe RCC, and 0 otherwise.
- Train a logistic regression model with the sigmoid function using the above formulas.

Clear Cell RCC ($Y = 2$):

- Set Y^i to 1 if the sample is clear cell RCC, and 0 otherwise.
- Train a logistic regression model with the sigmoid function using the above formulas.

Papillary RCC ($Y = 3$):

- Set Y^i to 1 if the sample is papillary RCC, and 0 otherwise.
- Train a logistic regression model with the sigmoid function using the above formulas.

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

Once the models have been trained, prediction is made for a new sample by computing the probability for each class using the trained models:

$$P(Chromophobe \mid X) = \sigma\left(\theta_0^{(1)} + \theta_1^{(1)} \cdot X_1 + \theta_2^{(1)} \cdot X_2 + \cdots + \theta_m^{(1)} \cdot X_m\right) \quad \text{(iii)}$$

$$P(Clear\ Cell \mid X) = \sigma\left(\theta_0^{(2)} + \theta_1^{(2)} \cdot X_1 + \theta_2^{(2)} \cdot X_2 + \cdots + \theta_m^{(2)} \cdot X_m\right) \quad \text{(iv)}$$

$$P(Papillary \mid X) = \sigma\left(\theta_0^{(3)} + \theta_1^{(3)} \cdot X_1 + \theta_2^{(3)} \cdot X_2 + \cdots + \theta_m^{(3)} \cdot X_m\right) \quad \text{(v)}$$

Where $\theta^{(1)}, \theta^{(2)}, \theta^{(3)}$ are the learned model parameters for each class.

Finally, the predicted class for the new sample is the class with the highest probability:

$$\hat{Y} = \arg\max_Y P(Y \mid X) \quad \text{(vi)}$$

So, $\hat{Y}$ will be one of the three classes: chromophobe, clear cell, or papillary RCC, based on the highest computed probability.

2.5.2 Naive Bayes.

The naive bayes algorithm makes predictions using Bayes' theorem [26]. It assumes that the features of a given dataset are conditionally independent given the class. It begins by calculating the prior probabilities of each class (chromophobe RCC, clear cell RCC, and papillary RCCs) based on the frequency of each class in the dataset. This is done using the following formula:

$$P(C_k) = \frac{Number\ of\ samples\ in\ class\ C_k}{Total\ number\ of\ samples} \quad \text{(vii)}$$

For each class, it estimates the likelihood of observing the gene expressions for that class. This step can vary depending on the type of naive bayes algorithm (e.g., Gaussian, Multinomial, or Bernoulli), but it essentially computes the probability distribution of the features, that is, genes within each class. Naive bayes then assumes that the features are conditionally independent given the class. This simplifying assumption makes the calculations tractable and is expressed as:

$$P(X_1 = x_1, X_2 = x_2, \cdots, X_n = x_n \mid C_k) = P(X_1 = x_1 \mid C_k) * P(X_2 = x_2 \mid C_k) * \cdots * P(X_n = x_n \mid C_k)$$

(viii)

To predict a new sample, that is, a new set of gene expressions, naive bayes computes the posterior probability for each class by using Bayes' theorem:

$$P(C_k \mid X) = \frac{P(X \mid C_k) * P(C_k)}{P(X)} \quad \text{(ix)}$$

Where $P(X|C_k)$ is the likelihood of observing the gene expressions for class C_k, $P(C_k)$ is the prior probability of class C_k, and $P(X)$ is a normalization constant. Naive bayes then predicts the class that maximizes the posterior probability:

$$\text{Predicted Class} = \arg\max_{C_k} P(C_k \mid X) \quad \text{(x)}$$

2.5.3 Support Vector Machines (SVMs).

SVMs are a type of supervised machine learning algorithm used for both classification and regression tasks [27]. When applied to a dataset with three distinct classes, the SVM algorithm seeks to determine the best hyperplane that separates the classes while maximizing the distance between the closest points from different classes. It does this by solving the following optimization problem:

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

$$\min_{w,b} \frac{1}{2}\left\|w\right\|^2 + C\sum_{i=1}^{n} \max\left(0,\ 1 - y_i\left(w^* x_i - b\right)\right) \quad \text{(xi)}$$

Where w represents the weights, b is the bias term, C is a hyperparameter controlling the penalty for misclassification, x_i are the input features, and y_i is the target class label (1 for the positive class, -1 for the negative class). The decision boundary is determined by the hyperplane that separates the data points of different classes with the largest margin. New data points are classified based on which side of the hyperplane they fall on.

2.5.4 Decision Tree.

Decision tree is a machine learning algorithm whose output is determined by recursively partitioning data based on the values of different features, with the goal of maximizing the information gain or Gini impurity reduction at each step. This is achieved through a series of if-else conditions leading to leaf nodes, which represent the final class predictions [28].

2.5.5 Random Forest.

Random forest is a machine learning algorithm based on ensemble learning that builds multiple decision trees using bootstrapped samples of data, with the final output determined by aggregating the predictions of individual trees, often through a majority vote. The output can be represented as an average or a weighted sum of the individual decision tree predictions [29].

2.5.6 Extra Trees Classifier.

Extra trees classifier also constructs multiple randomized decision trees. However, it differs from random forest in that it splits nodes using randomly selected features and thresholds, rather than searching for the best split. The final prediction is determined by aggregating the outputs of individual trees, typically through majority voting [30].

2.5.7 Bagging Classifier.

Bagging classifier creates multiple subsets of the original dataset through bootstrap sampling, trains a separate classifier (like decision trees) on each subset, and combines their predictions. The final output is determined by aggregating individual classifier predictions through majority voting [31].

2.6 Class Imbalance

To address the class imbalance, class weighting was implemented based on the frequency of each class in the dataset. The weighting was inversely proportional to the class frequencies, such that the class with the least frequency was assigned the highest weight, and the class with the most frequency was assigned the lowest weight.

2.7 Performance Evaluation

Four statistical metrics, namely sensitivity, specificity, F1 score, and AUC, were selected for assessing and comparing the performance of the various models in this study.

Sensitivity, also known as recall or true positive rate, assesses how well the model captures all the actual cancer cases. It measures the ratio of true positives to the total number of actual cancer cases.

$$Sensitivity = \frac{True\ \ positives}{True\ \ positives + False\ \ negatives} \quad \text{(xii)}$$

Specificity, otherwise known as true negative rate, evaluates how well the model correctly identifies non-cancer cases among the cases it predicted as negative. It measures the ratio of true negatives to the total number of actual non-cancer cases.

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

$$Specificity = \frac{True\ negatives}{True\ negatives + False\ positives} \quad \text{(xiii)}$$

F1 score is the harmonic mean of precision and recall, providing a balance between the two metrics.

$$F1\ score = 2 * \frac{\text{Pr}ecision \times \text{Re}call}{\text{Pr}ecision + \text{Re}call} \quad \text{(xiv)}$$

AUC assesses the ability of the model to discriminate between cancer and non-cancer cases at different classification thresholds. It represents the area under the receiver operating characteristic curve, which is a graphical representation of the true positive rate against the false positive rate at various threshold settings.

3 Results

We evaluated the performance of the machine learning algorithms using a hold-out test set comprising 180 samples. Table 2 shows the overall performance of each of the algorithms.

Table 2. Overall performance of each of the algorithms.

Model	Sensitivity	Specificity	F1 Score	AUC
Logistic Regression	82.53	94.63	86.58	97.78
Naive Bayes	90.27	95.89	88.83	98.06
Support Vector Machine	85.63	96.69	89.64	98.23
Decision Tree	93.21	96.44	89.51	96.82
Random Forest	93.12	98.09	91.42	99.14
Extra Trees Classifier	92.20	97.38	89.84	98.49
Bagging Classifier	84.42	95.72	87.43	99.16

Overall, the tree-based ensemble models, particularly random forest, demonstrated superior performance across all metrics. Random forest achieved the highest specificity (98.09%) and F1 score (91.42%), indicating its robust ability to correctly identify negative cases and maintain a strong balance between precision and recall. Decision tree achieved the highest sensitivity (93.21%), suggesting its effectiveness in identifying positive cases.

Bagging classifier, despite not leading in sensitivity, specificity, or F1 score, achieved the highest AUC score (99.16%). Extra trees classifier also showed competitive performance, achieving high scores across all metrics.

In contrast, non-tree-based algorithms, including logistic regression and SVM, showed comparatively lower performance in distinguishing between cancer subtypes.

Table 3 shows the performance of each of the algorithms across the three cancer subtypes.

Table 3. Performance of each of the algorithms across the three cancer subtypes.

Classifier	Performance Metrics								
	Sensitivity			Specificity			F1 Score		
	KICH	KIRC	KIRP	KICH	KIRC	KIRP	KICH	KIRC	KIRP
Logistic Regression	61.54	97.37	88.68	100.00	89.39	94.49	76.19	95.69	87.85
Naive Bayes	84.62	95.61	90.57	97.60	92.42	97.64	78.57	95.61	92.31
Support Vector Machine	61.54	99.12	96.23	100.00	92.42	97.64	76.19	97.41	95.33

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

Decision Tree	92.31	92.98	94.34	97.01	95.45	96.85	80.00	95.07	93.46
Random Forest	84.62	97.74	100.00	98.20	100.00	96.06	81.48	97.30	95.50
Extra Trees Classifier	84.62	93.86	98.11	97.60	98.48	96.06	78.57	96.40	94.55
Bagging Classifier	61.54	97.37	94.34	99.40	90.91	96.85	72.73	96.10	93.46

KICH, Chromophobe; KIRC, Clear Cell; KIRP, Papillary.

The confusion matrices for the tree-based models are shown in Figure 2, Figure 3, Figure 4, and Figure 5.

Figure 2. Confusion matrix for decision tree. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*

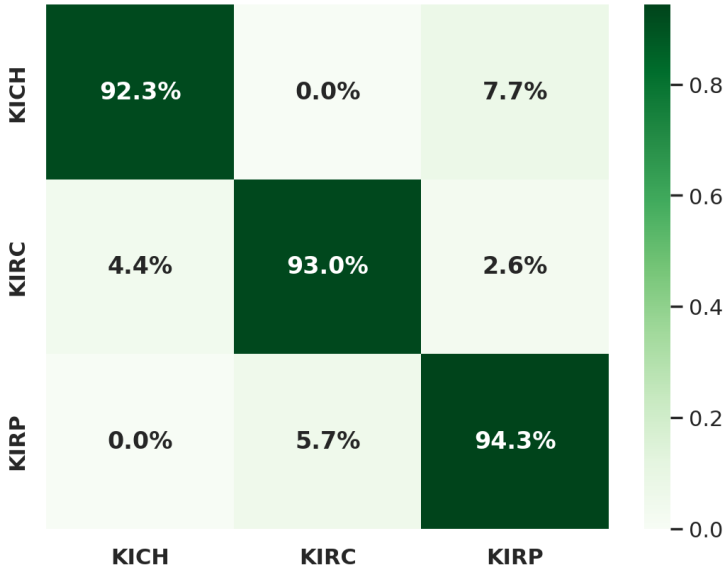
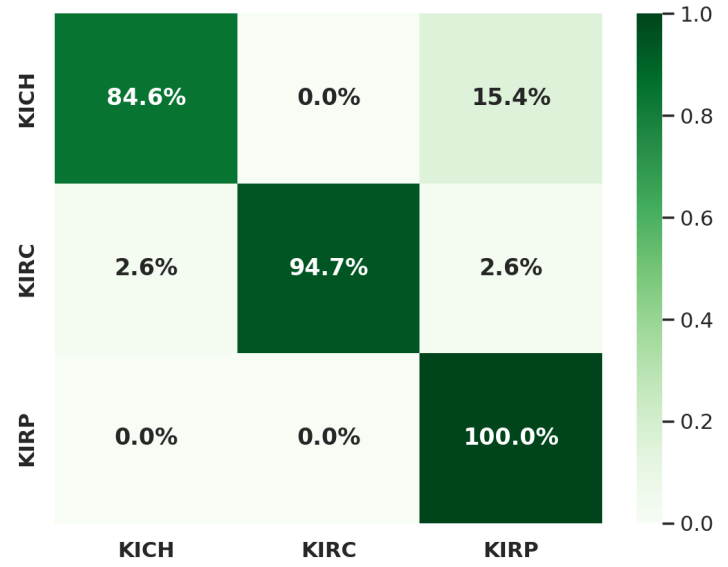


Figure 3. Confusion matrix for random forest. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*



 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

Figure 4. Confusion matrix for extra trees classifier. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*

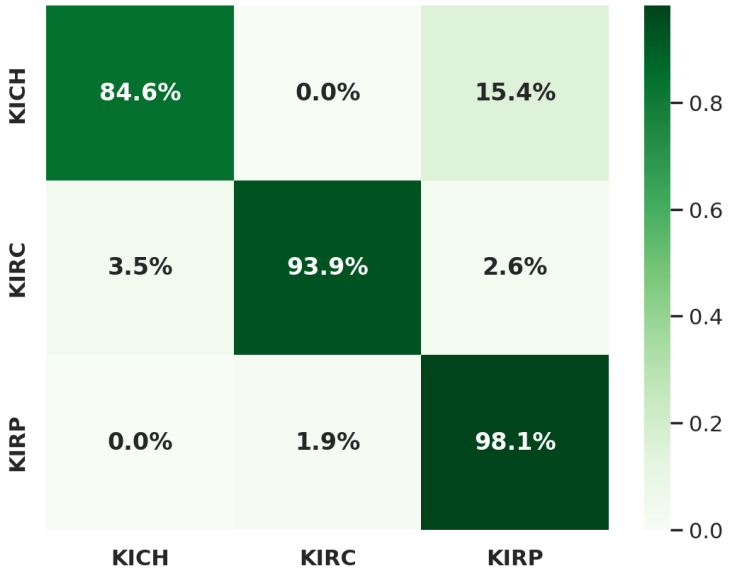
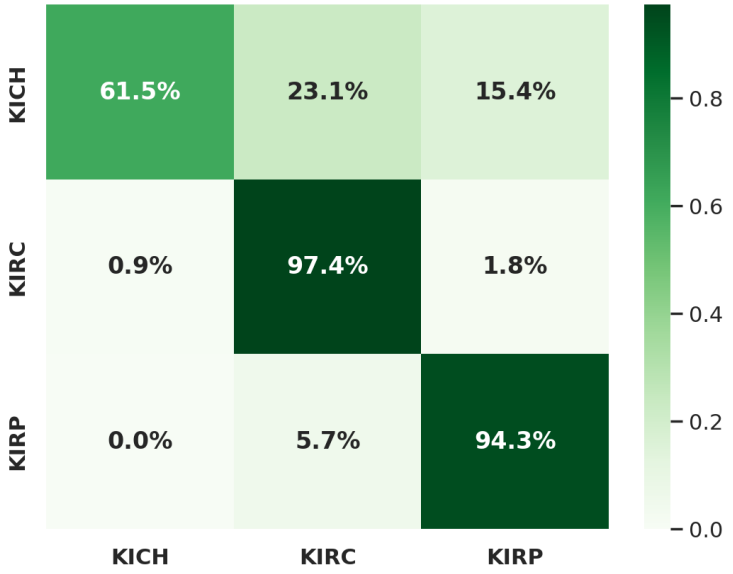


Figure 5. Confusion matrix for bagging classifier. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*



The receiver operating characteristic (ROC) curves for the tree-based models are shown in Figure 6, Figure 7, Figure 8, and Figure 9.

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

Figure 6. Multi-class ROC curves for decision tree. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*

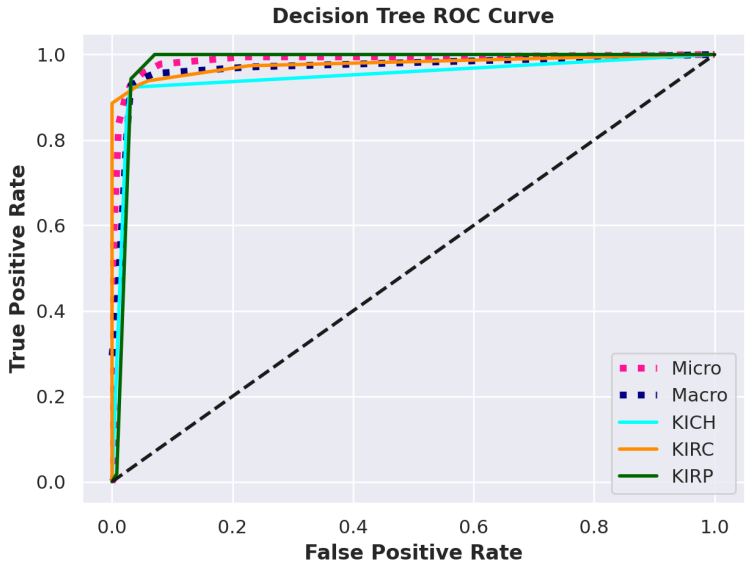
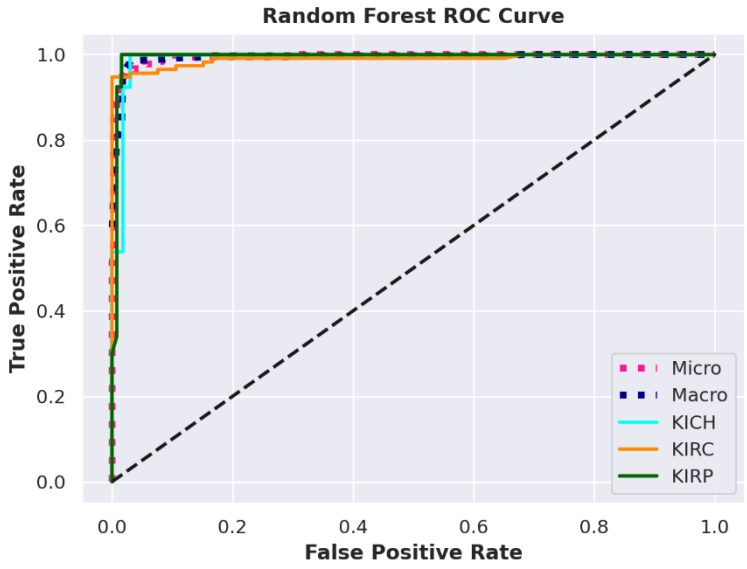


Figure 7. Multi-class ROC curves for random forest. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*



 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

Figure 8. Multi-class ROC curves for extra trees classifier. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*

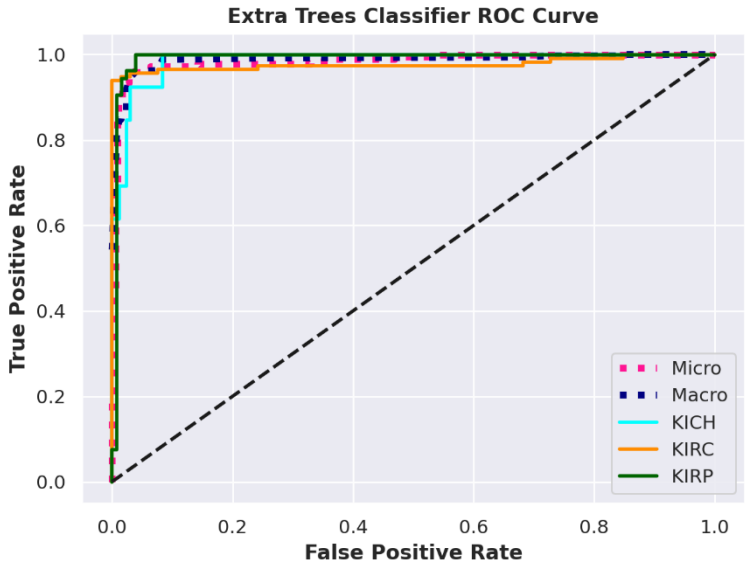
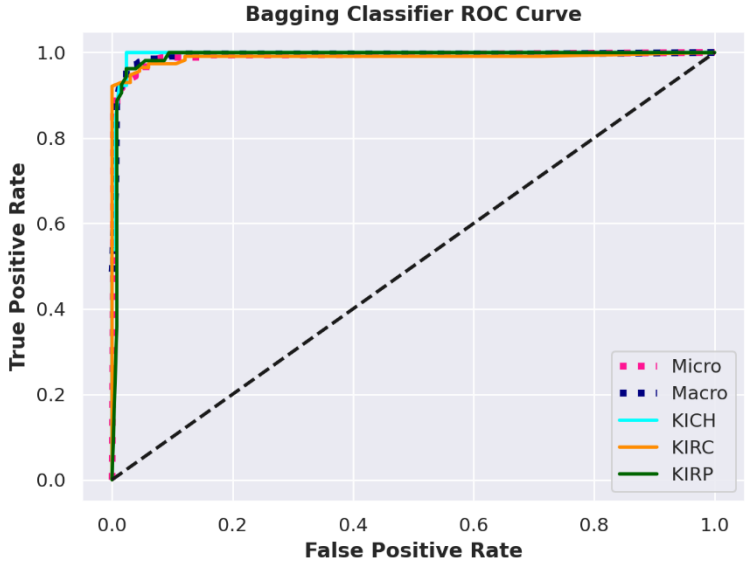


Figure 9. Multi-class ROC curves for bagging classifier. *KICH, Chromophobe; KIRC, Clear cell; KIRP, Papillary.*



4 Discussion

This work extends previous efforts in cancer genomics, such as TCGA project, which has provided comprehensive molecular characterization of various cancer types, including renal cell carcinoma [32]. Our approach builds upon previous work in cancer subtyping using gene expression data, such as the study by Ramaswamy et al. [33], which utilized SVMs for tumor classification. However, our research demonstrates that tree-based algorithms outperform non-tree-based algorithms in this context. Our tree-based method offers the advantage of handling complex, non-linear relationships in high-dimensional gene expression data.

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

The superior performance of tree-based models, particularly random forest, can be attributed to their ability to capture complex, non-linear relationships in gene expression datasets [34] [35]. Cancer subtype classification often involves complex interactions between multiple genes [11] [36], which may not be adequately captured by models such as logistic regression or SVM [37]. The ensemble nature of random forest, combining multiple decision trees, allows it to model these complex relationships more effectively, resulting in its high specificity and F1 scores.

The high sensitivity achieved by the decision tree model suggests that simpler tree-based structures can effectively identify positive cases [38]. However, the slightly lower specificity compared to random forest indicates a trade-off between sensitivity and specificity. In clinical settings, the choice between these models may depend on whether it is more important to minimize false positives or false negatives for the particular cancer subtype being studied.

The high AUC score of the bagging classifier suggests its ability to discriminate between cancer subtypes across various classification thresholds [39]. This further supports the effectiveness of tree-based models in RCC subtyping.

The relatively lower performance of non-tree-based algorithms like logistic regression and SVM underscores the importance of model selection in genomic studies. While these models are often favored for their interpretability and efficiency [40], our results suggest that they may not adequately capture the complexity of gene expression data in cancer subtype classification [11] [36].

These findings hold significant implications for both research and clinical practice. In research settings, our results emphasize the importance of considering tree-based and ensemble methods when analyzing gene expression data, particularly for cancer subtype classification. Clinically, the high performance of these models suggests potential for improving diagnostic accuracy and treatment planning in RCC.

However, it is important to note that model selection should not be based solely on performance metrics. Factors such as interpretability, computational efficiency, and the specific requirements of the clinical application should also be considered. For instance, while random forest showed the best overall performance, its “black box” nature [41] may make it challenging to interpret in clinical settings where understanding the decision-making process is important [42].

Our study was limited by the small sample size and the significant class imbalance in the RCC samples, particularly with chromophobe RCC compared to the other cancer subtypes. These limitations could potentially affect the reliability of our findings. Also, the study focused solely on RNA-seq gene expression profiles and did not incorporate other data types, such as histological images, mutation profiles, or copy number alterations, which could have impacted the outcome.

Future research should focus on validating these findings across larger and more diverse datasets of RCC subtypes and taking further steps to incorporate additional genomic and clinical data to enhance the clinical utility of machine learning-based RCC subtyping models.

5 Conclusion

In conclusion, we demonstrated the superiority of tree-based models in differentiating between the gene expression profiles of RCC subtypes, namely chromophobe, clear cell, and papillary RCCs. Future research should explore the integration of other data types to improve clinical utility of these models.

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

Acknowledgements

The author received no specific funding or support for this research.

Statement on conflicts of interest

The author has no competing interests to declare.

References

[1] Coffey NJ, Simon MC. Metabolic alterations in hereditary and sporadic renal cell carcinoma. Nature Reviews Nephrology 2024 20:4 2024; 20: 233–250.
[2] Pandey J, Syed W. Renal Cancer. 2024.
[3] Global Cancer Observatory. https://gco.iarc.fr/en (21 September 2024, date last accessed).
[4] Lopez-Beltran A, Carrasco JC, Cheng L, Scarpelli M, Kirkali Z, Montironi R. 2009 update on the classification of renal epithelial tumors in adults. Int J Urol 2009; 16: 432–443.
[5] DeCastro GJ, McKiernan JM. Epidemiology, clinical staging, and presentation of renal cell carcinoma. Urol Clin North Am 2008; 35: 581–592.
[6] Statistical Analysis of Next Generation Sequencing Data. Statistical Analysis of Next Generation Sequencing Data 2014;
[7] Rai MF, Tycksen ED, Sandell LJ, Brophy RH. Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears. J Orthop Res 2018; 36: 484–497.
[8] Koch CM, Chiu SF, Akbarpour M *et al.* A Beginner's Guide to Analysis of RNA Sequencing Data. Am J Respir Cell Mol Biol 2018; 59: 145–157.
[9] Zhao S, Zhang B, Zhang Y *et al.* Bioinformatics for RNA-Seq Data Analysis. Bioinformatics - Updated Features and Applications 2016;
[10] López-García G, Jerez JM, Franco L, Veredas FJ. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. PLoS One 2020; 15.
[11] Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011; 144: 646–674.
[12] Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 2009; 10: 57–63.
[13] Yang S, Naiman DQ. Multiclass cancer classification based on gene expression comparison. Stat Appl Genet Mol Biol 2014; 13: 477–496.
[14] Mohammed M, Mwambi H, Mboya IB, Elbashir MK, Omolo B. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Scientific Reports 2021 11:1 2021; 11: 1–22.
[15] Raut SA, Sathe SR, Raut A. Bioinformatics: Trends in gene expression analysis. ICBBT 2010 - 2010 International Conference on Bioinformatics and Biomedical Technology 2010; 97–100.
[16] Haury AC, Gestraud P, Vert JP. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. PLoS One 2011; 6.
[17] Divate M, Tyagi A, Richard DJ, Prasad PA, Gowda H, Nagaraj SH. Deep Learning-Based Pan-Cancer Classification Model Reveals Tissue-of-Origin Specific Gene Expression Signatures. Cancers (Basel) 2022; 14.
[18] Abdelwahab O, Awad N, Elserafy M, Badr E. A feature selection-based framework to identify biomarkers for cancer diagnosis: A focus on lung adenocarcinoma. PLoS One 2022; 17.
[19] Marostica E, Barber R, Denize T *et al.* Development of a histopathology informatics pipeline for classification and prediction of clinical outcomes in subtypes of renal cell carcinoma. Clinical Cancer Research 2021; 27: 2868–2878.
[20] Weinstein JN, Collisson EA, Mills GB *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 2013; 45: 1113–1120.
[21] R Core Team (2020) R A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. - References - Scientific Research Publishing. https://scirp.org/reference/referencespapers?referenceid=3064798 (21 September 2024, date last accessed).
[22] Colaprico A, Silva TC, Olsen C *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res 2016; 44: e71.
[23] van Rossum G. Python reference manual. 1995 (21 September 2024, date last accessed).

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533

[24] Pedregosa F, Varoquaux G, Gramfort A *et al.* Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research 2012; 12: 2825–2830.
[25] Böhning D. Multinomial logistic regression algorithm. Ann Inst Stat Math 1992; 44: 197–200.
[26] Murphy KP. Naive Bayes classifiers.
[27] Schölkopf B. SVMs - A practical consequence of learning theory. IEEE Intelligent Systems and Their Applications 1998; 13: 18–21.
[28] Freund Y, Mason L. The Alternating Decision Tree Learning Algorithm. International Conference on Machine Learning 1999;
[29] Liu Y, Wang Y, Zhang J. New Machine Learning Algorithm: Random Forest. International Conference on Information Computing and Applications 2012; 7473 LNCS: 246–252.
[30] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006; 63: 3–42.
[31] Breiman L. Bagging predictors. Mach Learn 1996; 24: 123–140.
[32] Ricketts CJ, De Cubas AA, Fan H *et al.* The Cancer Genome Atlas Comprehensive Molecular Characterization of Renal Cell Carcinoma. Cell Rep 2018; 23: 313.
[33] Ramaswamy S, Tamayo P, Rifkin R *et al.* Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci U S A 2001; 98: 15149–54.
[34] Qi Y. Random Forest for Bioinformatics. Ensemble Machine Learning 2012; 307–323.
[35] Goldstein BA, Polley EC, Briggs FBS. Random forests for genetic association studies. Stat Appl Genet Mol Biol 2011; 10.
[36] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000; 100: 57–70.
[37] Praveen Kumar V, Sowmya I. A Review on Pros and Cons of Machine Learning Algorithms. www.jespublication.com 2021; 12.
[38] Mienye ID, Jere N. A Survey of Decision Trees: Concepts, Algorithms, and Applications. IEEE Access 2024; 12: 86716–86727.
[39] Ling CX, Huang J. AUC: a Statistically Consistent and more Discriminating Measure than Accuracy.
[40] Räz T. ML Interpretability: Simple Isn't Easy. Stud Hist Philos Sci 2022; 103: 159–167.
[41] Simon SM, Glaum P, Valdovinos FS. Interpreting random forest analysis of ecological models to move from prediction to explanation. Sci Rep 2023; 13.
[42] Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol 2022; 38: 204–213.

 J Health Inform Afr. 2025;12(1):39-53. DOI: 10.12856/JHIA-2025-v12-i1-533