Machine Learning Classification Models using RNA-seq Gene Expression Data for Early-and Late-Stage Kidney Renal Clear Cell Carcinoma

Lawrence S. Macalalad*, Geoffrey A. Solano, and Joey Mark S. Diaz

Department of Physical Sciences and Mathematics,
College of Arts and Sciences,
University of the Philippines Manila, Manila, Philippines
lsmacalalad@up.edu.ph, gasolano@up.edu.ph, jsdiaz3@up.edu.ph

Abstract. Kidney Renal Clear Cell Carcinoma (KIRC) is one of the most common cancers in the world. With limited ideal testing and screening methods, this disease has been prone to misclassification and poor prognosis leading to late-stage detection and metastasis. One of the few methods that have been developed to address this is the use of biomarkers through gene expression. This study used KIRC TCGA RNA-seq gene expression data to classify early-stage or late-stage KIRC. The expressed genes are the features filtered through Information Gain Feature selection. The identified features were then used to train six (6) Machine Learning (ML) classification models and were evaluated by five (5) performance measures. These models were the Naive Bayes (NB), Sequential Minimal Optimization (SMO), Stochastic Gradient Descent (SGD), Logistic Regression (LR), J48 Classification (J48), and Random Forest (RF) classifiers. The performance measures used were accuracy, sensitivity, specificity, F-value, and the auROC. The features used in training the ML classification models were also subject to biological functional analysis through Gene Ontology (GO) Cellular Component enrichment analysis, GO molecular function enrichment analysis, and the KEGG Pathway Enrichment Analysis. The Machine learning classifier performance evaluation revealed that RF was the best-performing model in early stage KIRC, with a good performance in accuracy and specificity, followed by SGD. In contrast, all models had poor performances with the Late stage KIRC, equating to an overall average performance. The GO biological functional analyses and KEGG pathway analysis also showed the dominance of cell cycle-associated genes. Possibilities of improving future ML Classification model development through modifications in feature selection were idealized from the KEGG results. Its results were able to identify the likelihood of having feature-associated noise due highly expressed metabolic genes by diverse cell types, leading to the poor late-stage KIRC performances of the tested models.

Keywords: RNA-seq, early- and late-stage diseases, Kidney Renal Clear Cell Carcinoma, machine learning, classification models

J. Caro et al. (eds.), Proceedings of the Workshop on Computation: Theory and Practice (WCTP 2024), Atlantis Highlights in Computer Sciences 23,
https://doi.org/10.2991/978-94-6463-684-0_17

1 Introduction

Renal cancer (RCC) is the sixth most common cancer in men and tenth most common cancer in women [1]. Among the different types of Renal Cancers, the Kidney Renal Clear Cell Carcinoma (KIRC) is the most common pathological type accounting for 75% [2]. KIRC is also known to have poor prognosis and high morbidity, which stems from failed early diagnoses [3][4]. Although screening and testing methods for RCC are still yet to be established, abdominal contrast enhanced computed tomography (CT) and ultrasound scanning (USS) are the current gold standards for tumor identification. Ideally, RCC and KIRC diagnosis and development tracking in a patient could also use urinalysis, with serum and urine biomarkers as well [5]. Due to these limitations, majority of the identified cases of KIRC cases have already metastasized or are in late stages with some even being misclassified [6][7]. One of the leading ways in identifying candidate biomarkers for any disease is through gene expression analysis. Gene expression can be analyzed using RNA-seq data, which also provides a preview of possible molecular and biological mechanisms through mRNA levels [8]. The Kidney injury molecule-1 (*KIM-1*) is currently one of the most utilized biomarker in serum and urine samples, which is an indicator of renal tubular injury apparent in RCC [9]. Other promising biomarkers include the Gonadotropin releasing hormone 1 (*GNRH1*) and Leukotriene B4 Receptor (*LTB4R*), which are both immune-related prognostic biomarkers [10]. Although several studies have been done using gene expression data to identify the presence of KIRC in RCC patients. There are only a few transcriptomic studies, let alone machine learning studies, done in identifying and classifying patients between the early- and late-stage KIRC progression which highlights the importance of doing a KIRC stage classification study based on gene expression or mRNA data. Similar studies have already been done in other cancers such as breast, colon, lung, and prostate cancer which shows its importance in the cancer diagnostics and progression research paradigm [11][12][13][14].

2 Cancer Classification Studies with RNA-seq Data

Pre-treatment processes such as initial cancer diagnosis, screening, and classification are important steps in treating cancer patients. This allows earlier detection and a more targeted and specific therapy and course of treatment, leading to better survival chances, low chances of relapse, and a better overall quality of life post-treatment [15]. The use of biomarkers has been an impactful factor in improving the pre-treatment processes allowing a better and more precise understanding of cancer progression [16]. With the rise of high throughput sequencing of RNA-Seq, researchers were able to screen high amounts of genetic and transcriptomic data, allowing for a wider view in biomarker selection and disease correlation [17][18].

2.1 Cancer Classification studies using Machine Learning (ML)

Machine learning models have been utilized for cancer classification in several ways. It has been used for tumor localization, differentiation from other cancer subtypes, and disease progression or staging. A study done for tumor location prediction in prostate cancer was done by Hamzeh et. al. (2020), where they used gene expression data to identify the laterality of the tumor within the prostate [19]. They evaluated three different Machine Learning (ML) classifiers used, which were Naive Bayes, Random Forest, and SVM RBF which was also the best performer among the three. Bissanum et.al. (2021) also used a similar approach but applied it to triple negative breast cancer subtype classification. Their study used Support Vector Machines (SVM), K-nearest neighbor (KNN), Naïve Bayes (NB), Decision Tree (DT), Ensemble, Linear Discriminant, and Logistic Regression, and determined that the SVM algorithm produced the most accurate differentiation between the basal-like immune-activated (BLIA), basal-like immune-suppressed (BLIS), mesenchymal (MES), and luminal androgen receptor (LAR) subtypes [20]. Studies in evaluating classification models for cancer staging have also been performed, such as the study of Ubaldi et.al. (2021), where they developed machine learning models for lung cancer stage along with histology prediction using radiomics [21]. With the importance of evaluating and knowing the performance of ML classification models, this study explored six (6) classification models which are Naive Bayes (NB), Sequential Minimal Optimization (SMO), Stochastic Gradient Descent (SGD), Logistic Regression (LR), J48 Classification (J48), and Random Forest (RF).

2.2 Use of Gene expression as features from RNA-Seq Data

RNA-seq data or transcriptome data represent gene expression or mRNA levels. These data are analyzed to understand different molecular and biological processes occurring in a sample of interest. Widely used in cancer biomarker discovery studies, this data has been used in differentiating cancer types and stages as these have varying molecular mechanisms that are dictated by gene regulation [22]. With aims to standardize highly dimensional transcriptomic data, diverse pre-processing and normalization techniques have been used. One of the used normalization techniques by different studies and The Cancer Genome Atlas (TCGA) is log-transformation of mRNA expression z-scores [23][24]. This allows for a normalization based on a selected group of samples allowing a better understanding of gene upregulation and downregulation [25]. Despite this, some studies still use raw gene counts instead of log-transformed data, which may provide some differential gene expression bias [26].

3 Methodology Overview

This study explored different supervised machine learning classification models in classifying early- and late-stage KIRC. Publicly available RNA-Seq and clinical datasets from TCGA were initially collected and used in this study. These

datasets were then merged along with the inclusion of an additional data class—either early-stage or late-stage— for the analysis matrix. The data matrix created was then subjected to feature selection through Information Gain Attribute Evaluation. The selected features by the Information Gain Attribute Evaluation were then used to test six (6) classification models which are Naive Bayes (NB), Sequential Minimal Optimization (SMO), Stochastic Gradient Descent (SGD), Logistic Regression (LR), J48 Classification (J48), and Random Forest (RF). The identified features also underwent biological functional analysis through Gene Ontology (GO) Cellular Component enrichment analysis, GO molecular function enrichment analysis, and Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway Enrichment Analysis. The workflow of the methods is shown in Fig 1.

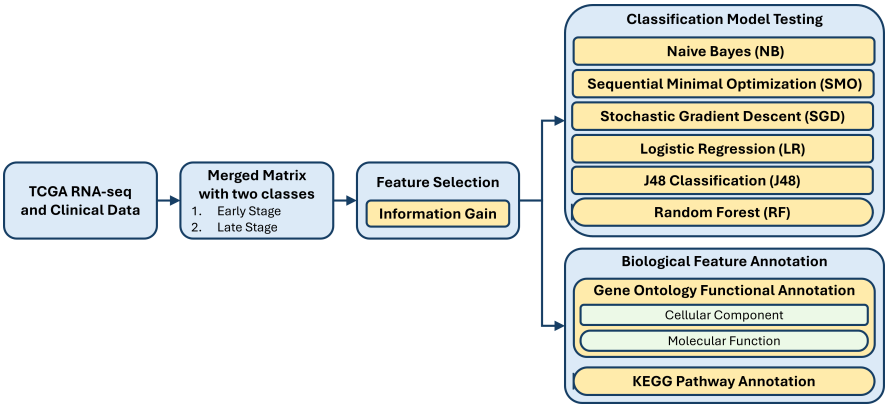

Fig. 1. Schematic diagram showing the overview of the methodology workflow

4 Datasets

TCGA Kidney Renal Clear Cell Carcinoma clinical and RNA-seq datasets (kirc_tcga) were accessed from the cBioPortal database and cohort builder (https://www.cbioportal.org/), which was developed by Memorial Sloan Kettering Cancer Center (MSK) [27]. These data are also notably linked to The Cancer Genome Atlas (TCGA) Provisional or Genome Data Analysis Center (GDAC) Firehose of Broad Institute (https://gdac.broadinstitute.org/), which is the source database of the "kirc_tcga" cohort [28]. A total of three datasets were initially retrieved from the database. These datasets were two (2) clinical datasets and an (1) RNA-seq data set. The two clinical datasets were the clinical-patient dataset and clinical-sample dataset. The clinical-patient dataset contained the cancer stages based on the American Joint Committee on Cancer (AJCC) Pathologic

Tumor Staging, while the RNA-seq data contained the gene expression values. To ensure a normalized and standard representation of mRNA expression levels, the gene expression data selected were log-transformed mRNA expression z-scores. These scores are calculated relative to all samples which is based on log RNA Seq V2 RSEM values. The two datasets were merged using the clinical-sample dataset as reference to have both cancer staging and gene expression data in one matrix since it contained both patient and sample IDs, which were separately used by the clinical-patient and RNA-seq data matrices. A class label of either early stage or late stage was also added in the matrix depending on the cancer stage as shown in table 1. The early stage class (N = 391) includes the samples classified as Stage I (N = 268) and Stage II (N=123), while the late stage class (N = 140) includes the Stage III (N = 57) and Stage IV (N = 83) samples. Samples with "NA" or blank values were also removed. In total, the created matrix had 20,354 features, which includes the sample ID, tumor staging, class, and gene expression values from 531 instances or samples, which was exported in a single CSV file. All merging and pre-processing of the data matrices were done in R version 4.3.1 through R Studio version 2024.04.1 Build 748.

Table 1. Summary of number of instances in training data

Class Label	Clinical Tumor Stage	Total number of instances
Early Stage	Stage I	268
	Stage II	123
Late Stage	Stage III	57
	Stage IV	83

5 Feature Selection

Feature Selection was performed in WEKA version 3.8.6. WEKA or Waikato Environment for Knowledge Analysis is a Java-based open-source tool developed by the University of Waikato for performing machine learning techniques [29]. The feature selection method used to decrease dimensionality was Information Gain Attribute Evaluator with Ranker as the search method having a threshold of 0.05 [30][31]. This decreased the number of attributes from 20,351 to 766 genes. Some of the top selected genes are shown in table 2. Interestingly, most of the top-ranked genes are involved in post-translational modifications and ligand interactions, which are molecular phenomena of high occurrence in cancers [32]. It is worth noting that interacting or related genes that might be of importance in classification might be overlooked despite the efficient selection of information gain, due to its limitations on redundant features along with complex biological interactions that might not be fully captured by the selection method through gene expressions alone.

Table 2. Top 25 of the selected 766 most informative features after feature selection by the Information Gain attribute evaluator and Ranker search method.

Rank	Feature/Gene	Gene Name
1	*FAM160A1*	FHF complex subunit HOOK interacting protein 1A
2	*PLEKHA8P1*	Pleckstrin homology domain containing A8 pseudogene 1
3	*UBE2C*	Ubiquitin conjugating enzyme E2 C (UBE2C)
4	*SKA1*	Spindle and kinetochore associated complex subunit 1
5	*NR3C2*	Nuclear receptor subfamily 3 group C member 2
6	*PMM2*	Phosphomannomutase 2
7	*TNFSF4*	TNF superfamily member 4
8	*CDC20*	Cell division cycle 20
9	*C1QL1*	Complement C1q like 1
10	*PAIP2B*	Poly(A) binding protein interacting protein 2B
11	*KIF18B*	Kinesin family member 18B
12	*IL20RB*	Interleukin 20 receptor subunit beta
13	*CDCA3*	Cell division cycle associated 3
14	*UFSP2*	UFM1 specific peptidase 2
15	*IBA57*	Iron-sulfur cluster assembly factor IBA57
16	*DNTTIP1*	Deoxynucleotidyltransferase terminal interacting protein 1
17	*KIF20A*	Kinesin family member 20A
18	*CENPA*	Centromere protein A
19	*PAEP*	Progestagen associated endometrial protein
20	*GSDMB*	Gasdermin B
21	*BIRC5*	Baculoviral IAP repeat containing 5
22	*LIMCH1*	LIM and calponin homology domains 1
23	*CHRM3*	Cholinergic receptor muscarinic 3
24	*CEP55*	Centrosomal protein 55
25	*NUMBL*	NUMB like endocytic adaptor protein

6 Classification Model Training and Performance Evaluation

The training set containing the selected features and class was then used to classify early stage and late stage KIRC using six supervised learning classification models. The six classification models are Naive Bayes (NB), Sequential Minimal Optimization (SMO), Stochastic Gradient Descent (SGD), Logistic Regression (LR), J48 Classification (J48), and Random Forest (RF). All of which were performed in WEKA as well.

6.1 Naive Bayes (NB)

Naive Bayes is a widely used probabilistic classification algorithm. It is based on the well-known Bayes theorem. It also considered as an effective and robust theorem in different applications [33]. It has been used in different studies that

are using gene expression data for different diseases. One of these is the study done by Yang et.al. (2020), where they used it in evaluating a prediction model for the response in neoadjuvant chemotherapy of breast cancers [34].

6.2 Sequential Minimal Optimization (SMO)

The Sequential Minimal Optimization (SMO) algorithm was developed from a decomposition algorithm. It also solves the uncertainty problem in multi-sample classification, along with its relatively faster operation time which could be of good use in multi-sample and high-dimensional clinical and gene expression data [35]. Despite not being as widely used in gene expression data, advanced variations have been evaluated in other previous studies. The study of Rani Jawhar (2020) is one of the studies that utilized advanced SMO techniques by applying it for lung tumor classification [36].

6.3 Stochastic Gradient Descent (SGD)

Stochastic Gradient Descent (SGD) is an algorithm used to solve extensive machine learning problems and uses as many training samples as possible while performing only a small number of calculations per sample. This algorithm has been utilized in various tasks including natural language processing (NLP), visual data processing, and other deep learning applications [37]. It has also been utilized in studying single-cell RNA sequences, such as the study by Rajesh et.al. (2023) for colorectal cancer tumor identification [38].

6.4 Logistic Regression (LR)

Logistic regression (LR) is an algorithm used in estimating the association of predictive features with a binary classification. Being an extension of linear regression, this algorithm has been widely used in different types of biological studies [39]. It has been used in different disease classifications through gene expressions including Hepatocellular carcinoma survival predictions, Kidney transplantation rejection, and Alzheimer's disease prediction [40][41][42].

6.5 J48 Classification (J48)

J48 Classification (J48) is a decision tree classification method that uses the C4.5 algorithm, which is a widely used technique in classifying numeric and categorical data. J48 uses attribute values to divide data into ranges and ignores missing values [43]. It has been used in different gene expression patterns and classification studies such as cancer stage discrimination, diabetes high-risk population estimation, and biomarkers for early disease diagnosis [44][45][46][47].

6.6 Random Forest (RF)

Random Forest classification utilizes an ensemble learning approach for classification and regression problems. Its multiple-model and classifier integration is known to greatly improve ML accuracy [48]. This approach has been applied in a wide variety of omics studies, including transcriptomics. It has been used in mixed-type data of different cancers, risk predictions, and other types of predictive and profiling studies [49][50][51].

6.7 Performance Evaluation

The classification models were all tested through 10-fold cross validation. The performance metrics used in determining their performance were accuracy (true positive rate), sensitivity (precision), specificity (recall), F-value, and the area under the receiver operating characteristic curve (auROC) [52]. The performance of the six classification models with respect to class and its average can be seen in Table 3 . In early stage classification, SGD was the most sensitive (77.90%) and had the best group variation (F-value = 0.805) which are average and good performances respectively. Despite this good performance ($score > 80.00\%$), Random Forest was considered as the best performing model for the early stage class. Its good performance in accuracy (87.40%) and specificity (87.40%), were reflected by having the best auROC (0.798). The good performance in the early stage class provides a positive note especially with regards to clinical decision making. The good specificity and accuracy performances of the models in early stage KIRC provides confidence identifying true non-early stage KIRC, which would reduce the risk of having the cancer develop into an advanced stage of KIRC due to inappropriate treatment. In late stage, all models performed poorly and average at best. SGD was the best late stage classification model in three metrics: Accuracy (62.60%), Specificity (62.60%), and F-value (0.663) despite these still being considered poor performances ($score < 70.00\%$). For sensitivity, Random forest was the best late stage classifier (72.50%), along with the best auROC. Averaging the early and late stage classifications, SGD was considered the best model in almost all metrics except for the auROC: Accuracy (75.30%), Sensitivity (75.00%), Specificity (75.30%), and F-value (0.75) which in totality is still an average or moderate performance ($70.00\% < score > 80.00\%$). The leading performance of RF and SGD can also be observed in the ROC curves in Fig 2. The ROC curves also reflected the overall better performance of all models in the steeper ROC curve in early stage (Fig 2.A.) compared to the ROC curve of the late stage (Fig 2.B.).

7 Biological Functional Analysis of Selected RNA-seq Features

The feature selection done by Information gain, is based on a computational approach of parent and child node entropy. To see the value of the identified features

Table 3. Machine Learning (ML) Classification Model Performance in Early- and Late-Stage KIRC

Class	ML Model	Accuracy	Sensitivity	Specificity	F-value	auROC
Early Stage	Naive Bayes	78.20%	76.50%	78.20%	0.773	0.719
	SMO	81.50%	75.90%	81.50%	0.786	0.704
	SGD	83.40%	**77.90%**	83.40%	**0.805**	0.730
	Logistic	69.20%	69.00%	69.20%	0.691	0.634
	J48	75.10%	72.40%	75.10%	0.737	0.652
	RF	**87.40%**	74.30%	**87.40%**	0.803	**0.798**
Late Stage	Naive Bayes	62.10%	64.30%	62.10%	0.632	0.739
	SMO	59.20%	67.00%	59.20%	0.629	0.704
	SGD	**62.60%**	70.50%	**62.60%**	**0.663**	0.730
	Logistic	51.00%	51.20%	51.00%	0.511	0.627
	J49	54.90%	58.20%	54.90%	0.565	0.652
	RF	52.40%	**72.50%**	52.40%	0.608	**0.798**
Average	Naive Bayes	71.90%	71.80%	71.90%	0.718	0.727
	SMO	72.90%	72.50%	72.90%	0.725	0.704
	SGD	**75.30%**	**75.00%**	**75.30%**	**0.750**	0.730
	Logistic	62.10%	62.10%	62.10%	0.621	0.631
	J50	67.20%	66.90%	67.20%	0.670	0.652
	RF	73.80%	73.60%	73.80%	0.728	**0.798**

or genes, Cellular Component and Molecular Functional Enrichment Analysis along with Pathway Enrichment analysis were performed to have an idea on its biological roles and involvements. Both analyses were performed in ShinyGO 0.80 (http://bioinformatics.sdstate.edu/go/). ShinyGO is a web-based analysis tool developed by South Dakota State University that is based on R/Bioconductor packages for gene functional annotation and enrichment which is linked to various databases including the Gene Ontology (GO) Database and Kyoto Encyclopedia of Genes and Genomes (KEGG) [53].

7.1 Cellular Component and Molecular Functional Enrichment Analysis

The results of the GO Cellular Component Enrichment Analysis showed that the selected genes mostly belonged to the nuclear lumen and nucleoplasm as shown in Fig 3A. These nuclear components are involved in nuclear biogenesis, which is a process that occurs in cell division, a repeated phenomena in cancer tumor formation [54][55]. This is also the same process where chromosomes are involved, specifically in the mitotic cycle. Overall, the cellular components selected by the feature selection are mostly cell cycle-involved components, which may be due to the rapid cell proliferation that occurs in KIRC and cancer in general. A similar occurrence was observed with the findings from the molecular functional enrichment in Fig 3B. Most molecular functions were identified to be binding in nucleic acids, with varying specificity. Nucleic acid bindings, which are mostly

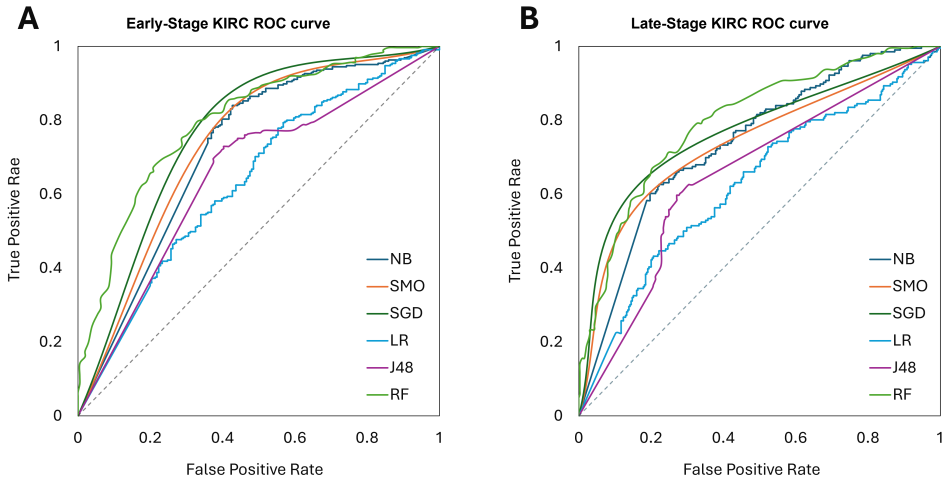

Fig. 2. ROC curve of tested ML classification models in (A) Early-stage and (B) Late-stage KIRC

explored in RNAs, can affect and change the development of a cell [56]. In a downstream view, these mechanisms cause signal transduction dysregulations and post-translational modifications that lead to pathway aberrations and errors in the cell cycle possibly leading to mutations and cancer development [57].

8 Biological Functional Analysis of Selected RNA-seq Features

8.1 Pathway Enrichment Analysis

Pathways are biomolecular processes which dictate the development of biological phenomena and diseases. The KEGG Pathway Enrichment analysis reflect the pathway involvements of genes, which are also the selected features in this case. The analysis showed a diverse array of pathways, which could be identified as immune-associated pathways, metabolic pathways, and the cell cycle. Interestingly, majority of the pathways of high confidence are involved in the cell cycle, which ranges from cell division to cell death — of either senescence or apoptosis— could be seen as the red to purple colored lines in Fig 4. It is also worth noting that despite, the high number of genes in Metabolic pathways, the confidence is still considered of poor status. This as a factor in the features used may have affected the overall poor to average performance of the ML classification models especially in the late stages of KIRC. This is because metabolism is a process that exists in all cells (but could be in different types), leading to an abundance of expressed genes and may have provided unwanted noise and dimensionality in the dataset. Although the presence of metabolic genes could

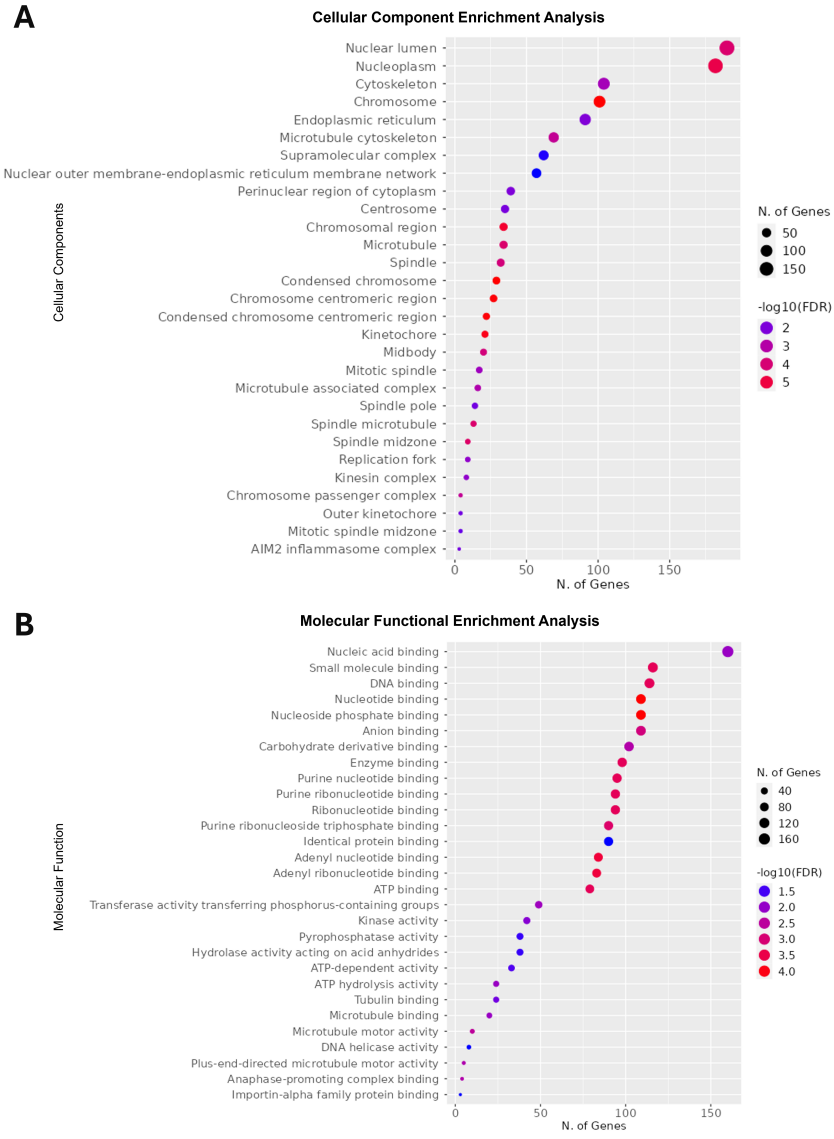

Fig. 3. Gene Ontology (GO) Functional Annotation by (A) Cellular component and (B) Molecular function of expressed genes selected by the Information Gain feature selection. Circle sizes indicate the number of genes (larger: more genes, smaller: less genes), while -log10(FDR) colors indicate the normalized false discover rate (Red: Higher confidence, Blue: Lower confidence). Only the GO terms satisfying the 0.05 FDR cut-off were included in this figure.

not be eliminated from the dataset, possible further dimension reduction methods, such as through use of Principal Components, could improve the the data dimensionality as well the selection of features.

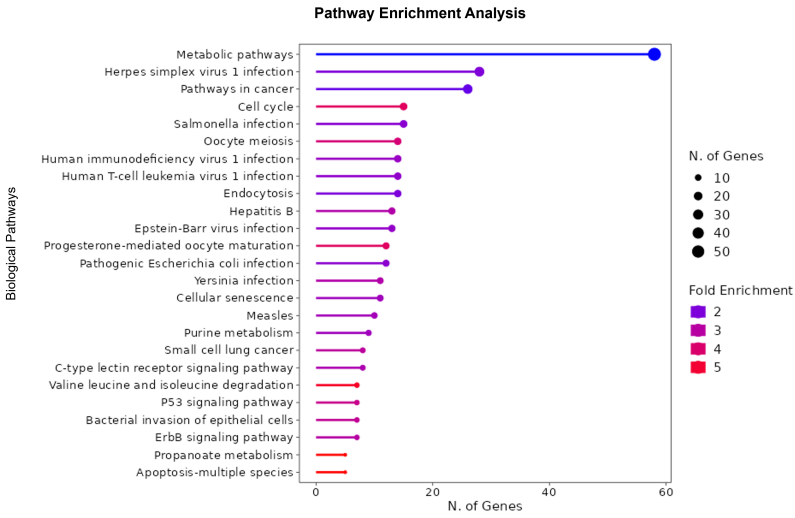

Fig. 4. KEGG Pathway Enrichment Analysis of expressed genes selected by the Information Gain feature selection. All pathways identified within the analysis are included in this figure. Circle sizes indicate the number of genes (larger: more genes, smaller: less genes), while Fold Enrichment colors indicate the confidence based on overrepresentations in specific pathways (Red: Higher confidence, Blue: Lower confidence).

9 Conclusion and Future Works

KIRC is a poorly identified and diagnosed disease which leads to a high mortality rate. This study explored and evaluated the performance of six machine learning classification models using RNA-seq expression data from the TCGA in classifying Early- and Late-Stage KIRC. It was found that in early stage KIRC, the RF was considered as the best performing model with a good performance in accuracy and specificity, while SGD was the most sensitive. Despite this good performance in early stage KIRC, all models had poor performances with the Late stage KIRC, which led to an overall average performance. Biological functional annotations through GO Cellular Component and Molecular Functional Enrichment Analysis also showed the heavy influence of cell cycle processes in the selected features of the Information Gain Attribute Evaluation. Interestingly, the abundance of genes from metabolic processes in the KEGG Pathway Enrichment analysis showed a possibility of having feature or dimensionality noise due

to highly expressed metabolic genes that are "by-products" of every cellular process.

With these findings, the it is recommended to explore the possibility of integrating pathway analysis data in the feature selection process which could also improve the overall performance of different ML classification models. Another possibility is to integrate multiple classification schemes along with different dimension reduction methods in classifying KIRC and other cancers with normalized gene expression data.

Acknowledgements

The authors would like to thank the Department of Physical Sciences and Mathematics, College of Arts and Sciences of the University of the Philippines Manila for their unwavering support and dedication.

Data Availability Statement

The Datasets used are available in the cBioPortal for Cancer Genomics: Kidney Renal Clear Cell Carcinoma (TCGA, Firehose Legacy) (https://www.cbioportal.org/study/summary?id=kirc_tcga), and the GDAC Firehouse: KIRC dataset (https://gdac.broadinstitute.org/runs/stddata__2016_01_28/data/KIRC/20160128/).

References

1. Bahadoram, S., Davoodi, M., Hassanzadeh, S., Bahadoram, M., Barahman, M., Mafakher, L. (2022). Renal cell carcinoma: an overview of the epidemiology, diagnosis, and treatment. G Ital Nefrol, 39(3), 2022.
2. Li, D., Liu, S., Xu, J., Chen, L., Xu, C., Chen, F., ... Wang, Y. (2021). Ferroptosis-related gene CHAC1 is a valid indicator for the poor prognosis of kidney renal clear cell carcinoma. Journal of Cellular and Molecular Medicine, 25(7), 3610-3621.
3. Xie, L., Li, H., Zhang, L., Ma, X., Dang, Y., Guo, J., ... Guo, X. (2020). Autophagy-related gene P4HB: a novel diagnosis and prognosis marker for kidney renal clear cell carcinoma. Aging (Albany NY), 12(2), 1828.
4. Cui, H., Shan, H., Miao, M. Z., Jiang, Z., Meng, Y., Chen, R., ... Liu, Y. (2020). Identification of the key genes and pathways involved in the tumorigenesis and prognosis of kidney renal clear cell carcinoma. Scientific reports, 10(1), 4271.
5. Usher-Smith, J., Simmons, R. K., Rossi, S. H., Stewart, G. D. (2020). Current evidence on screening for renal cancer. Nature Reviews Urology, 17(11), 637-642.
6. Rathmell, W. K., Rumble, R. B., Van Veldhuizen, P. J., Al-Ahmadie, H., Emamekhoo, H., Hauke, R. J., ... Singer, E. A. (2022). Management of metastatic clear cell renal cell carcinoma: ASCO guideline. Journal of Clinical Oncology, 40(25), 2957-2995.
7. Barthélémy, P., Rioux-Leclercq, N., Thibault, C., Saldana, C., Borchiellini, D., Chevreau, C., ... Négrier, S. (2021). Non-clear cell renal carcinomas: Review of new molecular insights and recent clinical data. Cancer Treatment Reviews, 97, 102191.

8. Niu, L., Guo, W., Song, X., Song, X., Xie, L. (2021). Tumor-educated leukocytes mRNA as a diagnostic biomarker for non-small cell lung cancer. Thoracic cancer, 12(6), 737-745.
9. Alderson, H. V., Ritchie, J. P., Pagano, S., Middleton, R. J., Pruijm, M., Vuilleumier, N., Kalra, P. A. (2016). The associations of blood kidney injury Molecule-1 and neutrophil Gelatinase–associated Lipocalin with progression from CKD to ESRD. Clinical Journal of the American Society of Nephrology, 11(12), 2141-2149.
10. Wu, H. H., Yan, X., Chen, Z., Du, G. W., Bai, X. J., Tuoheti, K., Liu, T. Z. (2021). GNRH1 and LTB4R might be novel immune-related prognostic biomarkers in clear cell renal cell carcinoma (ccRCC). Cancer cell international, 21, 1-14.
11. Wu, J., Hicks, C. (2021). Breast cancer type classification using machine learning. Journal of personalized medicine, 11(2), 61.
12. Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., ... Lv, X. (2022). Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Computers in biology and medicine, 145, 105409.
13. Yuan, F., Lu, L., Zou, Q. (2020). Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1866(8), 165822.
14. Hamzeh, O., Alkhateeb, A., Zheng, J., Kandalam, S., Rueda, L. (2020). Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. BMC bioinformatics, 21, 1-10.
15. Correa, A. F., Jegede, O. A., Haas, N. B., Flaherty, K. T., Pins, M. R., Adeniran, A., ... Uzzo, R. G. (2021). Predicting disease recurrence, early progression, and overall survival following surgical resection for high-risk localized and locally advanced renal cell carcinoma. European urology, 80(1), 20-31.
16. Garufi, A., D'Orazi, V., Pistritto, G., Cirone, M., D'Orazi, G. (2023). HIPK2 in angiogenesis: A promising biomarker in cancer progression and in angiogenic diseases. Cancers, 15(5), 1566.
17. Yang, X., Kui, L., Tang, M., Li, D., Wei, K., Chen, W., ... Dong, Y. (2020). High-throughput transcriptome profiling in drug and biomarker discovery. Frontiers in genetics, 11, 19.
18. Hong, M., Tao, S., Zhang, L., Diao, L. T., Huang, X., Huang, S., ... Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. Journal of hematology oncology, 13, 1-16.
19. Hamzeh, O., Alkhateeb, A., Zheng, J., Kandalam, S., Rueda, L. (2020). Prediction of tumor location in prostate cancer tissue using a machine learning system on gene expression data. BMC bioinformatics, 21, 1-10.
20. Bissanum, R., Chaichulee, S., Kamolphiwong, R., Navakanitworakul, R., Kanokwiroon, K. (2021). Molecular classification models for triple negative breast cancer subtype using machine learning. Journal of Personalized Medicine, 11(9), 881.
21. Ubaldi, L., Valenti, V., Borgese, R. F., Collura, G., Fantacci, M. E., Ferrera, G., ... Marrale, M. (2021). Strategies to develop radiomics and machine learning models for lung cancer stage and histology prediction using small data samples. Physica Medica, 90, 13-22.
22. Hong, M., Tao, S., Zhang, L., Diao, L. T., Huang, X., Huang, S., ... Zhang, H. (2020). RNA sequencing: new technologies and applications in cancer research. Journal of hematology oncology, 13, 1-16.
23. Castaldo, R., Pane, K., Nicolai, E., Salvatore, M., Franzese, M. (2020). The impact of normalization approaches to automatically detect radiogenomic phenotypes characterizing breast cancer receptors status. Cancers, 12(2), 518.

24. Tihagam, R. D., Bhatnagar, S. (2023). A multi-platform normalization method for meta-analysis of gene expression data. Methods, 217, 43-48.
25. Smith, L. C., Venosa, A., Gow, A. J., Laskin, J. D., Laskin, D. L. (2020). Transcriptional profiling of lung macrophages during pulmonary injury induced by nitrogen mustard. Annals of the New York Academy of Sciences, 1480(1), 146-154.
26. Mahin, K. F., Robiuddin, M., Islam, M., Ashraf, S., Yeasmin, F., Shatabda, S. (2022). PanClassif: Improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. Genomics, 114(2), 110264.
27. Gui, Y., Liu, X., Wang, C., Yang, P. (2021). Overexpressing PTTG Family Genes Predict Poor Prognosis in KIRC.
28. Priyam, J., Saxena, U. (2023). Computational gene expression and network analysis of Myc reveal insights into its diagnostic and prognostic role in subtypes of renal cancer. Applied biochemistry and biotechnology, 195(7), 4251-4276.
29. Alpan, K., İlgi, G. S. (2020, October). Classification of diabetes dataset with data mining techniques by using WEKA approach. In 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT) (pp. 1-7). IEEE.
30. Rodrigues, V., Deusdado, S. (2020). Metalearning approach for leukemia informative genes prioritization. Journal of Integrative Bioinformatics, 17(1), 20190069.
31. Prasetiyowati, M. I., Maulidevi, N. U., Surendro, K. (2022). The accuracy of Random Forest performance can be improved by conducting a feature selection with a balancing strategy. PeerJ Computer Science, 8, e1041.
32. Pan, S., Chen, R. (2022). Pathological implication of protein post-translational modifications in cancer. Molecular aspects of medicine, 86, 101097.
33. Wickramasinghe, I., Kalutarage, H. (2021). Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 25(3), 2277-2293.
34. Yang, L., Fu, B., Li, Y., Liu, Y., Huang, W., Feng, S., ... Bu, H. (2020). Prediction model of the response to neoadjuvant chemotherapy in breast cancers by a Naive Bayes algorithm. Computer methods and programs in biomedicine, 192, 105458.
35. Zhao, C., Dai, L., Huang, Y. (2023). Fractional Order Sequential Minimal Optimization Classification Method. Fractal and Fractional, 7(8), 637.
36. Vijila Rani, K., Joseph Jawhar, S. (2020). Automatic segmentation and classification of lung tumour using advance sequential minimal optimisation techniques. IET Image Processing, 14(14), 3355-3365.
37. Tian, Y., Zhang, Y., Zhang, H. (2023). Recent advances in stochastic gradient descent in deep learning. Mathematics, 11(3), 682.
38. Rajesh, M., Senapati, B., Das, R., Martha, S. (2023). Identifying colorectal tumor for single cell RNA sequence using rectified linear unit with stochastic gradient descent. Procedia Computer Science, 218, 189-198.
39. Schober, P., Vetter, T. R. (2021). Logistic regression in medical research. Anesthesia Analgesia, 132(2), 365-366.
40. Ksiażek, W., Gandor, M., Pławiak, P. (2021). Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. Computers in Biology and Medicine, 134, 104431.
41. Park, S., Guo, K., Heilman, R. L., Poggio, E. D., Taber, D. J., Marsh, C. L., ... Friedewald, J. J. (2021). Combining blood gene expression and cellfree DNA to diagnose subclinical rejection in kidney transplant recipients. Clinical Journal of the American Society of Nephrology, 16(10), 1539-1551.
42. Lee, T., Lee, H. (2020). Prediction of Alzheimer's disease using blood gene expression data. Scientific reports, 10(1), 3485.

43. Hermawan, D. R., Fatihah, M. F. G., Kurniawati, L., Helen, A. (2021, October). Comparative study of J48 decision tree classification algorithm, random tree, and random forest on in-vehicle CouponRecommendation data. In 2021 International conference on artificial intelligence and big data analytics (pp. 1-6). IEEE.
44. Ramisa, A. J., Hossain, A., Islam, S. M. I., Swadesh, P. M., Islam, M. T., Rahman, M. A., Parvez, M. Z. (2021, December). Gene Expression Data Classification and Pattern Analysis Using Data Driven Approach. In 2021 International Conference on Machine Learning and Cybernetics (ICMLC) (pp. 1-9). IEEE.
45. Kumar, R., Bhanti, P., Marwal, A., Gaur, R. K. (2020). Gene expression-based supervised classification models for discriminating early-and late-stage prostate cancer. Proceedings of the National Academy of Sciences, India Section B: Biological Sciences, 90, 541-565.
46. Pei, D., Yang, T., Zhang, C. (2020). Estimation of diabetes in a high-risk adult Chinese population using J48 decision tree model. Diabetes, Metabolic Syndrome and Obesity, 4621-4630.
47. Sahu, A., Qazi, S., Raza, K., Singh, A., Verma, S. (2022). Machine learning-based approach for early diagnosis of breast cancer using biomarkers and gene expression profiles. In Computational intelligence in oncology: applications in diagnosis, prognosis and therapeutics of cancers (pp. 285-306). Singapore: Springer Singapore.
48. Dhivyaa, C. R., Sangeetha, K., Balamurugan, M., Amaran, S., Vetriselvi, T., Johnpaul, P. (2020). Skin lesion classification using decision trees and random forest algorithms. Journal of Ambient Intelligence and Humanized Computing, 1-13.
49. Quist, J., Taylor, L., Staaf, J., Grigoriadis, A. (2021). Random forest modelling of high-dimensional mixed-type data for breast cancer classification. Cancers, 13(5), 991.
50. Mendapara, K. (2024). Development and evaluation of a chronic kidney disease risk prediction model using random forest. Frontiers in Genetics, 15, 1409755.
51. Guo, L., Wang, Z., Du, Y., Mao, J., Zhang, J., Yu, Z., ... Li, Y. (2020). Random-forest algorithm based biomarkers in predicting prognosis in the patients with hepatocellular carcinoma. Cancer cell international, 20, 1-12.
52. Tharwat, A. (2021). Classification assessment methods. Applied computing and informatics, 17(1), 168-192.
53. Ge, S. X., Jung, D., Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics, 36(8), 2628-2629.
54. Manda, N. K., Golla, U., Sesham, K., Desai, P., Joshi, S., Patel, S., ... Rokana, N. (2023). Tuning between nuclear organization and functionality in health and disease. Cells, 12(5), 706.
55. Flores, L. F., Tader, B. R., Tolosa, E. J., Sigafoos, A. N., Marks, D. L., Fernandez-Zapico, M. E. (2021). Nuclear dynamics and chromatin structure: implications for pancreatic cancer. Cells, 10(10), 2624.
56. Qin, H., Ni, H., Liu, Y., Yuan, Y., Xi, T., Li, X., Zheng, L. (2020). RNA-binding proteins in tumor progression. Journal of hematology oncology, 13, 1-23.
57. Yu, F., Wei, J., Cui, X., Yu, C., Ni, W., Bungert, J., ... Qian, Z. (2021). Post-translational modification of RNA m6A demethylase ALKBH5 regulates ROS-induced DNA damage response. Nucleic acids research, 49(10), 5779-5797.