Identifying a novel microRNA biomarker for renal cell carcinoma using a machine learning approach

Hanika Pandya, Zara Thalji, Maura Barron

Academy of Science, Academies of Loudoun, Leesburg, Virginia, United States

ABSTRACT

In the United States, over 200,000 people live with RCC; the mortality of half these patients significantly increase due to the lack of adequate and rapid diagnosis (Campbell, 2014). MicroRNAs have been found to be strong biomarkers of tumors, easily detected in the bloodstream around the tumor (Moldovan *et al.*, 2014). Many regulated microRNAs affect the outcome of cancers. There is limited previous research in microRNAs' potential role in RCC. A total of 7223 isoform microRNA expressions of healthy and cancerous RCC samples were taken from the Gene Expression Omnibus. Of the microRNAs, the Best First Attribute Selector (BFAS) was used in the Weka interface to choose 53 microRNAs, which are most significant in predicting each patient's outcome. The dataset was separated between training and testing data using a 4-fold cross-validation. Moreover, algorithms were run with the selected features to determine the highest classification accuracy, precision, and recall. The BFAS with the J48 decision tree algorithm and the BFAS with a Hoeffding Tree algorithm each had an accuracy of 91.89%. According to these models, bta-miR-200c_st and ACA39_st may be significant as biomarkers of RCC.

Keywords: Biomarker, Hoeffding tree, J48 decision tree, MicroRNA, Renal cell carcinoma

INTRODUCTION

In the United States, approximately 200,000 people live with RCC annually (Cairns, 2011). The mortality rates of these patients are 19%, 26%, 47%, and 92% for stages 1, 2, 3, and 4, respectively, indicating the need for early detection (Cairns, 2011). Those with metastatic RCC (pT4) have a median survival of around 13 months and a 5-year survival rate of under 10% (Cairns, 2011). RCC is a malignant cancer of the kidneys with numerous subtypes: Clear cell, collecting duct, chromophobe, mucinous tubular, multicellular, papillary, and renal medullary. The cancer is created from many specialized cells located along the nephron, the basic structure of the kidney. Only 2–4% of RCC in patients have hereditary causes, but a mutation in one or more of the following tumor suppressor genes – von Hippel-Lindau, fumarate hydratase, folliculin, and succinate dehydrogenase genes – which stops the cells from proliferating excessively and becoming cancer, has been found to increase the likelihood of a person to develop RCC (Martinez *et al.*, 2000). The treatment options for patients are usually nephrectomy surgeries. If the cancer progresses too far, both kidneys may have to be removed, necessitating a kidney transplant (National Cancer Institute, 2019).

In the UK in 2014, almost half of cancer patients were diagnosed too late, reducing the likelihood of their treatment success, and increasing the mortality of treatable cancers (Campbell, 2014). Late diagnosis is caused by an array of factors: The expense of tumor detection machines and the inefficiency of doctor diagnosis to name a few. After diagnosis, treatment options usually require invasive surgeries and scans, intimidating patients even further (Arruebo *et al.*, 2011). Even with imaging, many early RCCs are not accurately distinguishable from other non-malignant renal lesions (American Cancer Society, 2019). Thus, a non-invasive cancer detection technology is needed. Research into biomarkers can be a promising avenue of developing early cancer diagnosis with a non-invasive analysis of patients' blood.

A common biomarker for cancers and other diseases is microRNA: Non-coding, tiny RNAs that target messenger RNAs to regulate gene expression. They are transcribed from DNA sequences and do not code for protein like other RNAs; instead, they engage with target RNAs to suppress protein expression, create overexpression of certain specific genes,

Doi: 10.30954/2319-5169.1.2021.2

Submission: 23-02-2021 **Acceptence:** 27-04-2021
Received: 8-03-2021 **Published:** 15-06-2021

***Corresponding Author:** Maura Barron, Academy of Science, Academies of Loudoun, Leesburg, Virginia, United States.
E-mail: maura.mcnamee@lcps.org

or act signaling molecules to mediate cell communications (O'brien *et al.*, 2018). The dysfunction of microRNAs manifested by altered microRNA concentrations is commonly present in either the tissue, blood, urine, or cerebrospinal fluid of disease-ridden patients (Moldovan *et al.*, 2014). Malignant changes happen when mutations occur in genes, in which the oncogenes are activated, and tumor suppressor genes are inactivated. In microRNAs, overexpression leads to reducing oncogene expression and vice versa (Paranjape *et al.*, 2009). Proto-oncogenes are genes that typically help cells grow. However, in a situation where further copies of proto-oncogene mutations are created, a gene can be permanently activated. This corrupted gene, which is always activated, causes a non-stop proliferation of cancer cells (American Cancer Society, 2019). Cancers are also believed to be created or able to grow by evading the process of apoptosis – programmed cell death that most cells undergo. Misregulation of microRNA can lead to the inactivation of tumor suppressor genes responsible for apoptosis and assists the proliferation of cells until the cancer is formed (American Cancer Society, 2019). Dysregulation of the microRNAs responsible for the inactivation of tumor suppressor genes and activation of oncogenes is an indicator for most types of cancer. However, it is not the only source of misregulation of genes that may cause cancer. Other issues include transcriptional misregulation, where the gene is unable to convert DNA to RNA correctly, leading to failed gene activity. This is a consideration while analyzing the role of microRNAs in cancers.

A statistical analysis of a dataset of cancer microRNAs was classified with random forests and support vector machine algorithms by Rehman *et al*. (2019). Since many microRNAs affect the subtype and outcome of cancer, ML was chosen as a suitable method to sort through the data (Strimbu and Travel, 2010). It concluded that ML could validate 12 experimentally determined microRNA biomarkers of breast cancer. Zakrzewska *et al.* (2019) used RWeka with decision tree-based models to reveal altered microRNA expression patterns in neuronal-glial tumors. A random forest algorithm identified novel diagnostic markers for soft-tissue sarcoma that distinguished between synovial sarcoma and malignant peripheral nerve sheath tumor using gene expression data from the Genotype-Tissue Expression project and the French Sarcoma Group (van IJzendoorn *et al*., 2019).

In RCC, miR-17-5p and miR-224 both target hypoxia-inducible factor 1α and Von Hippel-Landau (VHL) protein. These proteins are lost in approximately 70% of all RCCs, which may lead to an over-proliferation of nephrons and cancer (Al-Ali *et al.*, 2012). In addition, miR-106b, miR-1233, miR-1290, miR-210, miR-7-1, miR-320b, and miR-93 were all selected microRNAs that were highly differentially expressed in RCC by looking through microRNA profiles (Fedorko *et al.*, 2016). Bhowmick *et al.* (2018) utilized the Cancer Genome Atlas database, including 255 instances with renal clear cell carcinoma, to use ML to compile sets of possible microRNAs that often occur in various cancerous samples. Ten microRNAs were found to be most indicative of classification accuracy; they are possible biomarkers of RCC of clear cell subtype. Thirty-five discriminative microRNAs for the subtypes of RCC were found using data from the Cancer Atlas Genome project (Ali *et al.*, 2018).

Previous research did not compare healthy and cancerous samples or tissue and blood microRNAs to find discriminative biomarkers of any subtype of RCC. This investigation hopes to use healthy and cancerous samples from all subtypes of RCC, differentiating between tissue and blood microRNAs, to build classification models that reveal discriminative microRNAs; those that overlap between tissue and blood models will be evaluated as potential new biomarkers. Clear cell, papillary, and chromophore are the subtypes being investigated because they are the most common (Cairns, 2011).

The Waikato Environment for Knowledge Analysis (Weka) interface will be used because it facilitates both dataset and algorithm exploration (Witten *et al.*, 2016). A dataset specific to RCC was pulled from the Gene Expression Omnibus and the Cancer Genome Atlas. It has 18 data instances of microRNAs associated with chromophore RCC, 243 data instances of microRNAs for clear cell RCC, and 77 data instances of microRNAs for papillary RCC. Instances in these cases correspond to microRNAs, not the individual patient, so they will be reorganized to have microRNAs as attributes normalized to reads per million and with the instances characterized by patient id. All these databases have microRNA comparing cancerous tissue to normal tissue. However, they include other metal profiling such as high-grade versus low grade, metastasis, poor versus good outcome, and blood profiling. This can be used to describe the microRNA profiles of cancer patients (Luan *et al.*, 2016). Following the successful classification of possible microRNA biomarkers, the mirTarBase database will be used to determine the gene targets of the microRNA biomarkers.

MATERIALS AND METHODS

Isoform microRNA expression values were downloaded from the Gene Expression Omnibus, submitted by Wach *et al.* (2013), and last updated in 2017. It includes data from two of the major subtypes, clear cell and papillary. These two subtypes represent approximately 90% of all RCC cases. The entire data set had 7815 microRNAs as features and a total of 29,690 instances. The data were exported to excel and reformatted to classify the RCC subtype as either cancerous or healthy.

This process was repeated for each sample and added to the same data table. Missing values were substituted by the average expression for that attribute in the healthy or cancerous class. The dataset was converted into an AIFF file to be placed in Weka. Overfitting was avoided by cross-validation. This ensures that the software is not running the algorithm through the entire dataset until the model is complete. The algorithm is always unfamiliar with the testing data until it runs. Ten-fold is the standard because it creates enough splits to avoid familiarity. Four is the minimum number of folds that ensure some variance in a smaller dataset. Too many splits can lower the number of sample combinations in each fold and interfere with the algorithm's accuracy. Since the dataset had a limited number of samples, 4-fold cross-validation was used to achieve variance and accuracy.

Six preliminary algorithms were run through the dataset with no feature selection as a baseline for results in Weka. Three feature selection techniques were tested as wrappers with each algorithm, and the best feature selection method was selected as Best First Attribute Selector (BFAS). This information is explained in the results section. BFAS is an information gain feature selection technique in which an optimization algorithm searches the space of attribute subsets by greedy hill-climbing, building a solution by constantly looking at the next attribute that offers the greatest and most immediate benefit iteratively and incrementally – augmented with a backtracking facility. This evaluates the added attribute's benefit compared to the original one and removes it if it is detrimental. This feature selection technique was evaluated by CfsSubsetEval, which evaluates the worth of a subset of attributes by considering each feature's individual predictive ability and the degree of redundancy between them. With this feature selection technique, a panel of 52 microRNAs was generated for the model.

Using BFAS, the top algorithms with the highest original accuracies were considered most suitable for the data: Unpruned J48 decision tree and Hoeffding tree. It is important to note that the name of the J48 decision tree varies; Weka labeled it as such (it is otherwise referred to as a C4.5 decision tree classifier). The Hoeffding tree algorithm is an incremental type of decision tree in which a new node is built from an attribute based on its value the Hoeffding bound equation (Bifet and Gavalda, 2013). The unpruned J48 algorithm is a standard decision tree that creates a node based on its mathematical benefit to the classification accuracy (Bhargava *et al.*, 2013). The AUC-ROC, the area underneath the curve of receiver operating characteristics, is a performance metric created by taking the integral of a probability curve plotted against a model's false-positive rate and measuring how well a model can distinguish between classes. The AUC-PRC, or the area underneath the recall curve, is a performance metric created from taking the integral of a curve of precision values plotted against recall value and represented the models' tradeoff between precision and recall. A great model has an AUC-ROC and AUC-PRC value closer to 1, and a weak model has an AUC-ROC and AUC-PRC closer to 0. Both methods are used as performance metrics as they distinguish between classes by measuring separability. By depicting this distinction, it provides insight into how accurately and effectively the model works. Zheng *et al.* (2020) used an AUC-ROC on a microRNA data sample to demonstrate and compare the performance of the two models. The information gained from these metrics provided insight into the overall performance of the models. The models were compared to find the accuracy, precision, recall, AUC-ROC, and AUC-PRC in cancer prognosis.

RESULTS AND DISCUSSION

Equations for precision, recall, and accuracy:

$$\mathit{Formula\ for\ Precison\ \%} = \frac{\mathit{True\ Positives}}{\mathit{True\ Positives} + \mathit{False\ Positives}} \times 100$$

$$\mathit{Sample\ Calculation\ for\ Healthy} = \frac{16}{18} \times 100 \cong 88.9\%$$

$$\mathit{Formula\ for\ Recall\ \%} = \frac{\mathit{True\ Positives}}{\mathit{True\ Positives} + \mathit{False\ Negatives}} \times 100$$

$$\mathit{Sample\ Calculation\ for\ Healthy} = \frac{16}{17} \times 100 \cong 94.1\%$$

$$\mathit{Formula\ for\ Accuracy\ \%} = \frac{\mathit{True\ Positives} + \mathit{True\ Negatives}}{\mathit{Total\ Samples}} \times 100$$

$$\mathit{Sample\ Calculation\ for\ Accuracy} = \frac{16+18}{16+18+3} \times 100 = 91.89\%$$

****AUC−ROC and AUC−PRC were generated from Weka*

It is important to note that there was only one accuracy value for each model because accuracy is a measure of the model successfully predicting both healthy and RCC samples. Table 7, which showed the performance metrics for healthy versus cancerous classification with a Random Forest Algorithm, also had an accuracy of 91.892%, but the model was not deemed as an appropriate one due to its large size and tendency to overfit with the small sample size.

The unpruned J48 tree is less convoluted than the Hoeffding tree. It includes fewer nodes or pathways taken. An unpruned J48 decision tree was chosen over a pruned J48 decision tree because pruning is used to reduce the size of a tree and may lead to overfitting; since the unpruned J48 decision tree is relatively small, with two nodes as shown in Figure 1, there

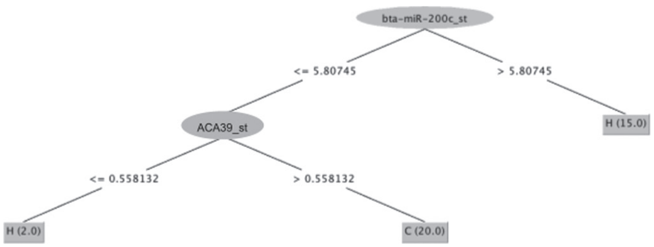


Figure 1: J-48 decision tree visualization: "H" indicates a healthy classification and "C" indicates a cancerous classification

Table 1: Small section of the final dataset used – labeled as healthy or cancerous samples

#ID_REF	14q-0_st	14qI-1_st	14qI-2_st
H	0.6634	0.621946	0.195252
C	0.541661	0.591711	0.348382

Table 2: Performance metrics of healthy versus RCC with JRip algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	68.4	76.5	0.809	0.792	72.973
RCC	77.8	70	0.809	0.782	

Table 3: Performance metrics by healthy versus RCC with SMO algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	91.7	64.7	0.799	0.755	81.081
RCC	76	95	0.799	0.749	

SMO: Sequential minimal optimization

Table 4: Performance metrics by healthy versus RCC with Ibk algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	78.6	64.7	0.741	0.662	75.678
RCC	73.9	85	0.741	0.716	

Table 5: Performance metrics by healthy versus RCC with J48 algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	88.2	88.2	0.915	0.848	89.189
RCC	90	90	0.915	0.904	

Table 6: Performance metrics by healthy versus RCC with Hoeffding Tree algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	100	52.9	0.781	0.772	78.378
RCC	71.4	100	0.753	0.704	

was no reason to try cutting the tree's size down and risk overfitting a model of a small sample size that is already prone to overfitting. The J48 is an efficient algorithm because it can complete the task with few nodes.

Table 7: Performance metrics by healthy versus RCC with Random Forest Algorithm

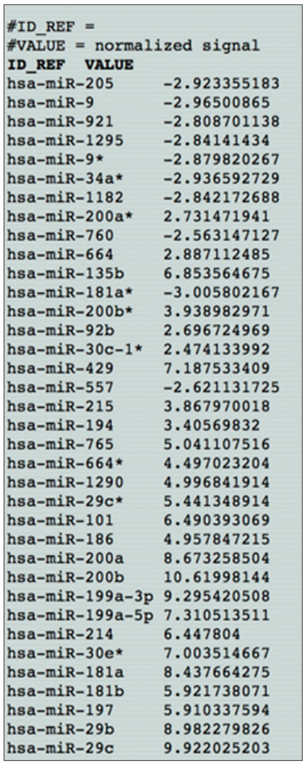
Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	84.6	64.7	0.731	0.780	91.892
RCC	90.0	90.0	0.731	0.715	

Table 8: Performance metrics by healthy versus RCC with Best First Attribute Selection and J48 algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	93.8	88.2	0.943	0.899	91.892
RCC	90.5	95	0.943	0.929	

Table 9: Performance metrics by healthy versus RCC with Best First Attribute Selection and Hoeffding Tree algorithm

Class	Precision (%)	Recall (%)	AUC-ROC	AUC-PRC	Accuracy (%)
Healthy	88.9	94.1	0.896	0.848	91.89
RCC	94.7	90	0.894	0.908	



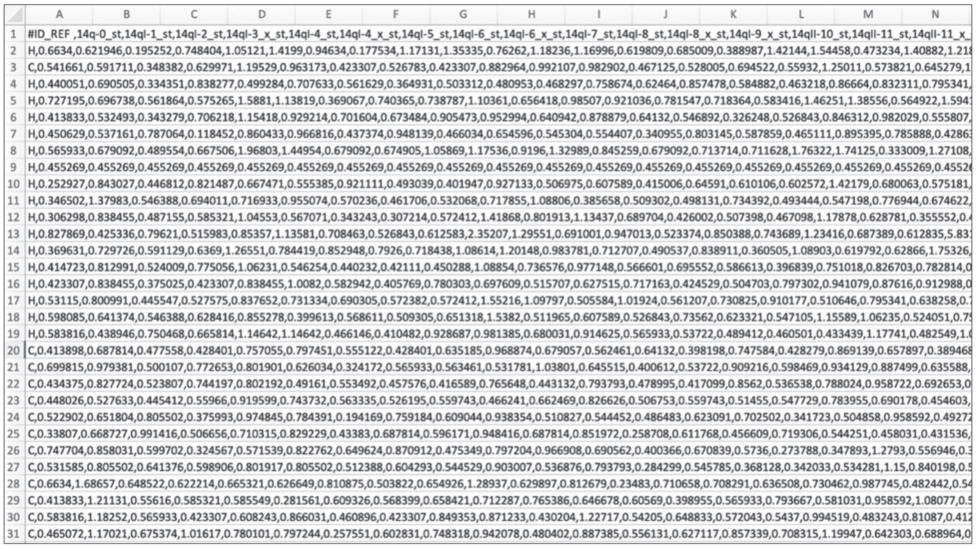
```
#ID_REF =
#VALUE = normalized signal
ID_REF   VALUE
hsa-miR-205      -2.923355183
hsa-miR-9        -2.96500865
hsa-miR-921      -2.808701138
hsa-miR-1295     -2.84141434
hsa-miR-9*       -2.879820267
hsa-miR-34a*     -2.936592729
hsa-miR-1182     -2.842172688
hsa-miR-200a*    2.731471941
hsa-miR-760      -2.563147127
hsa-miR-664      2.887112485
hsa-miR-135b     6.853564675
hsa-miR-181a*    -3.005802167
hsa-miR-200b*    3.938982971
hsa-miR-92b      2.696724969
hsa-miR-30c-1*   2.474133992
hsa-miR-429      7.187533409
hsa-miR-557      -2.621131725
hsa-miR-215      3.867970018
hsa-miR-194      3.40569832
hsa-miR-765      5.041107516
hsa-miR-664*     4.497023204
hsa-miR-1290     4.996841914
hsa-miR-29c*     5.441348914
hsa-miR-101      6.490393069
hsa-miR-186      4.957847215
hsa-miR-200a     8.673258504
hsa-miR-200b     10.61998144
hsa-miR-199a-3p 9.295420508
hsa-miR-199a-5p 7.310513511
hsa-miR-214      6.447804
hsa-miR-30e*     7.003514667
hsa-miR-181a     8.437664275
hsa-miR-181b     5.921738071
hsa-miR-197      5.910337594
hsa-miR-29b      8.982279826
hsa-miR-29c      9.922025203
```

Figure 2: The raw data set – unedited

Precision is the measure of true positives overall samples classified as positive. In Table 8, the precision for the unpruned J48 decision tree with BFAS was 93.8% for healthy classification and 90.5% for cancerous classification so that the

Table 10: Performance metrics by healthy versus RCC with various feature selections

Feature selection type	Algorithm	Accuracy (%)	Confusion matrix
Gain ratio attribute evaluation	Hoeffding tree	83.78	a b [11 6 0 20]a=H b=c
Gain ratio attribute evaluation	J48 tree	89.18	a b [15 2 2 18]a=H b=c
Correlation attribute evaluation	Hoeffding tree	86.49	a b [12 5 0 20]a=H b=c
Correlation attribute evaluation	j48 tree	89.18	a b [15 2 2 18]a=H b=c



```
   A  B  C  D  E  F  G  H  I  J  K  L  M  N
1  #ID_REF ,14q-0_st,14ql-1_st,14ql-2_st,14ql-3_x_st,14ql-4_st,14ql-4_x_st,14ql-5_st,14ql-6_st,14ql-6_x_st,14ql-7_st,14ql-8_st,14ql-8_x_st,14ql-9_x_st,14qll-10_st,14qll-11_st,14qll-11_x_
2  H,0.6634,0.621946,0.195252,0.748404,1.05121,1.4199,0.94634,0.177534,1.17131,1.35335,0.76262,1.18236,1.16996,0.619809,0.685009,0.388987,1.42144,1.54458,0.473234,1.40882,1.218
3  C,0.541661,0.591711,0.348382,0.629971,1.19529,0.963173,0.423307,0.526783,0.423307,0.882964,0.992107,0.982902,0.467125,0.528005,0.694522,0.55932,1.25011,0.573821,0.645279,1
4  H,0.440051,0.690505,0.334351,0.838277,0.499284,0.707633,0.561629,0.364931,0.503312,0.480953,0.468297,0.758674,0.62464,0.857478,0.584882,0.463218,0.86664,0.832311,0.795341,
5  H,0.727195,0.696738,0.561864,0.575265,1.5881,1.13819,0.369067,0.740365,0.738787,1.10361,0.656418,0.98507,0.921036,0.781547,0.718364,0.583416,1.46251,1.38556,0.564922,1.594
6  H,0.413833,0.532493,0.343279,0.706218,1.15418,0.929214,0.701604,0.673484,0.905473,0.952994,0.640942,0.878879,0.64132,0.546892,0.326248,0.526843,0.846312,0.982029,0.555807,
7  H,0.450629,0.537161,0.787064,0.118452,0.860433,0.966816,0.437374,0.948139,0.466034,0.654596,0.545304,0.554407,0.340955,0.803145,0.587859,0.465111,0.895395,0.785888,0.4286
8  H,0.565933,0.679092,0.489554,0.667506,1.96803,1.44954,0.679092,0.674905,1.05869,1.17536,0.9196,1.32989,0.845259,0.679092,0.713714,0.711628,1.76322,1.74125,0.333009,1.27108,
9  H,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.455269,0.4552
10 H,0.252927,0.843027,0.446812,0.821487,0.667471,0.555385,0.921111,0.493039,0.401947,0.927133,0.506975,0.607589,0.415006,0.64591,0.610106,0.602572,1.42179,0.680063,0.575181,
11 H,0.346502,1.37983,0.546388,0.694011,0.716933,0.955074,0.570236,0.461706,0.532068,0.717855,1.08806,0.385658,0.509302,0.498131,0.734392,0.493444,0.547198,0.776944,0.674622,
12 H,0.306298,0.838455,0.487155,0.585321,1.04553,0.567071,0.343243,0.307214,0.572412,1.41868,0.801913,1.13437,0.689704,0.426002,0.507398,0.467098,1.17878,0.628781,0.355552,0.4
13 H,0.827869,0.425336,0.79621,0.515983,0.85357,1.13581,0.708463,0.526843,0.612583,2.35207,1.29551,0.691001,0.947013,0.523374,0.850388,0.743689,1.23416,0.687389,0.612835,5.83
14 H,0.369631,0.729726,0.591129,0.6369,1.26551,0.784419,0.852948,0.7926,0.718438,1.08614,1.20148,0.983781,0.712707,0.490537,0.838911,0.360505,1.08903,0.619792,0.62866,1.75326,
15 H,0.414723,0.812991,0.524009,0.775056,1.06231,0.546254,0.440232,0.42111,0.450288,1.08854,0.736576,0.977148,0.566601,0.695552,0.586613,0.396839,0.751018,0.826703,0.782814,0
16 H,0.423307,0.838455,0.375025,0.423307,0.838455,1.0082,0.582942,0.405769,0.780303,0.697609,0.515707,0.627515,0.717163,0.424529,0.504703,0.797302,0.941079,0.87616,0.912988,0
17 H,0.53115,0.800991,0.445547,0.527575,0.837652,0.731334,0.690305,0.572382,0.572412,1.55216,1.09797,0.505584,1.01924,0.561207,0.730825,0.910177,0.510646,0.795341,0.638258,0.7
18 H,0.598085,0.641374,0.546388,0.628416,0.855278,0.399613,0.568611,0.509305,0.651318,1.5382,0.511965,0.607589,0.526843,0.73562,0.623321,0.547105,1.15589,1.06235,0.524051,0.75
19 H,0.583816,0.438946,0.750468,0.665814,1.14642,1.14642,0.466146,0.410482,0.928687,0.981385,0.680031,0.914625,0.565933,0.53722,0.489412,0.460501,0.433439,1.17741,0.482549,1.0
20 C,0.413898,0.687814,0.477558,0.428401,0.757055,0.797451,0.555122,0.428401,0.635185,0.968874,0.679057,0.562461,0.64132,0.398198,0.747584,0.428279,0.869139,0.657897,0.389468
21 C,0.699815,0.979381,0.500107,0.772653,0.801901,0.626034,0.324172,0.565933,0.563461,0.531781,1.03801,0.645515,0.400612,0.53722,0.909216,0.598469,0.934129,0.887499,0.635588,
22 C,0.434375,0.827724,0.523807,0.744197,0.802192,0.49161,0.553492,0.457576,0.416589,0.765648,0.443132,0.793793,0.478995,0.417099,0.8562,0.536538,0.788024,0.958722,0.692653,0
23 C,0.448026,0.527633,0.445412,0.55966,0.919599,0.743732,0.563335,0.526195,0.559743,0.466241,0.662469,0.826626,0.506753,0.559743,0.51455,0.547729,0.783955,0.690178,0.454603,
24 C,0.522902,0.651804,0.805502,0.375993,0.974845,0.784391,0.194169,0.759184,0.609044,0.938354,0.510827,0.544452,0.486483,0.623091,0.702502,0.341723,0.504858,0.958592,0.4927
25 C,0.33807,0.668727,0.991416,0.506656,0.710315,0.829229,0.43383,0.687814,0.596171,0.948416,0.687814,0.851972,0.258708,0.611768,0.456609,0.719306,0.544251,0.458031,0.431536,
26 C,0.747704,0.858031,0.599702,0.324567,0.571539,0.822762,0.649624,0.870912,0.475349,0.797204,0.966908,0.690562,0.400366,0.670839,0.5736,0.273788,0.347893,1.2793,0.556946,0.3
27 C,0.531585,0.805502,0.641376,0.598906,0.801917,0.805502,0.512388,0.604293,0.544529,0.903007,0.536876,0.793793,0.284299,0.545785,0.368128,0.342033,0.534281,1.15,0.840198,0.5
28 C,0.6634,1.68657,0.648522,0.622214,0.665321,0.626649,0.810875,0.503822,0.654926,1.28937,0.629897,0.812679,0.23483,0.710658,0.708291,0.636508,0.730462,0.987745,0.482442,0.54
29 C,0.413833,1.21131,0.55616,0.585321,0.585549,0.281561,0.609326,0.568399,0.658421,0.712287,0.765386,0.646678,0.60569,0.398955,0.565933,0.793667,0.581031,0.958592,1.08077,0.5
30 C,0.583816,1.18252,0.565933,0.423307,0.608243,0.866031,0.460896,0.423307,0.849353,0.871233,0.430204,1.22717,0.54205,0.648833,0.572043,0.5437,0.994519,0.483243,0.81087,0.412
31 C,0.465072,1.17021,0.675374,1.01617,0.780101,0.797244,0.257551,0.602831,0.748318,0.942078,0.480402,0.887385,0.556131,0.627117,0.857339,0.708315,1.19947,0.642303,0.688964,0
```

Figure 3: Small subsection of the data classified as healthy versus cancerous

model can be evaluated as moderately precise on the testing data. The recall is the measure of true positives overall actual positives. While the recall of 88.2% for healthy classification is not considered vital, the 95% recall for cancerous classification is considered very strong. The correct diagnosis of cancer patients is weighted more heavily than the correct diagnosis of healthy patients when evaluating the model's performance as desired.

Again, the ability for the model to diagnose cancerous patients is slightly more important than the ability to diagnose healthy patients, so the 0.929 AUC-PRC value for cancerous classification in Table 8 is indicative of a strong model.

Table 1 represents a subsection of the final edited data set as a reference point. Tables 2 through 6 contain various combinations of performance metrics of the algorithms. Table 2 has a relatively low performance metric as well as Table 4 and Table 6. In comparison, Table 3 and Table 5 were able to establish an accuracy of above 80%, but this was still not considered strong enough. Finally, while the accuracy value for Table 9 is extremely strong in respect to this research, it is still weaker than Table 7 and Table 8, thus it was not selected. Below is Table 10, a representation of the confusion matrix of each of the following algorithms with either gain ratio attribute evaluation or correlation attribute evaluation. This information provides insight into the false positive, false negative, true positive, and true negative presence in each model.

As seen above, Figure 2 represents the raw unedited data taken directly from the source with no changes made. In comparison, Figure 3 is that very source that was entered into excel. As the file size for both are quite large, only a small subsection was added.

The lower AUC-PRC value for healthy classification indicated that the Hoeffding tree was not as good of a model as the J48 model. Overall, significant evidence suggests that both models are strong enough to classify between healthy and cancerous samples in the RCC dataset. Specifically, both trees used the same two microRNAs as the critical features in their trees.

In addition, ACA39_st, abridged labeling of microRNA, was a significant feature used in both models that may be a novel biomarker that has not been recorded in literature. This is a variant of a microRNA coding linked to acetyl CoA, which is a protein primarily responsible for lipid metabolism. A properly functioning kidney must be able to metabolize lipid binding proteins well. In addition, improper lipid metabolism, and subsequent lipid accumulation, in the kidney has been a hallmark of the clear cell subtype of RCC (Sanchez and Simon, 2018). Thus, there is some reasoning to believe a microRNA associated with acetyl CoA and the microRNA ACA39_st would play RCC. These results show a possible significant novel biomarker for RCC. A novel microRNA biomarker for RCC can open the door for cheap and efficient early detection of cancer, increasing the chance of survival. The role of acetyl

CoA (ACA39_st) is not clear; there is a need for future research that investigates the link between ACA39_st and RCC as a potential biomarker.

CONCLUSION

Zakrzewska *et al*. (2019) determined that the feature selected to give a model's highest predictive ability may predict the patient's disease or cancer. Thus, the microRNAs chosen as the nodes in the J48 model, bta-miR-200c_st and ACA39_st suspected to be highly predictive of RCC diagnosis. While the Hoeffding tree cannot be visualized, it uses all the 52 microRNAs in the preselected panel of features to build the tree, including bta-miR-200c_st and ACA39_st. The role of miR-200c in RCC has been supported in the past. Nakada et al. (2008) found significant downregulation of microRNAs miR141 and miR-200c in RCC profiling.

ACKNOWLEDGMENTS

We would like to acknowledge the help of Dr. Julie Sohl and Mr. Duke Writer with our deepest gratitude. They provided foundational support across multiple aspects of this research. Thank you to Mr. Sundaram Thirukkurungudi for aiding us in the revisions of this paper. Furthermore, we would like to thank the Academies of Loudoun for providing us with the facilities and opportunity to conduct our research.

FINANCIAL SUPPORT AND SPONSORSHIP

No additional financial support other than instructional funding from the Academies of Loudoun was received.

CONFLICTS OF INTEREST

The authors and the people listed in the acknowledgments did not have any conflicts of interest with this research.

REFERENCES

Al-Ali, B.M., Ress, A.L., Gerger, A. and Pichler, M. 2012. MicroRNAs in renal cell carcinoma: Implications for pathogenesis, diagnosis, prognosis and therapy. *Anticancer Res.*, 32, 3727-3732.

Anon. 2014. Oncogenes and Tumor Suppressor Genes. American Cancer Society. Available from: https://www.cancer.org/cancer/cancer-causes/genetics/genes-and-cancer/oncogenes-tumor-suppressor-genes.html. [Last accessed on 2021 Jan 28].

Anon. 2018. Key Statistics About Kidney Cancer. American Cancer Society. Available from: https://www.cancer.org/cancer/kidney-cancer/about/key-statistics.html. [Last accessed on 2021 Jan 28].

Anon. 2019. Renal Cell Cancer Treatment (PDQ®)-Patient Version. National Cancer Institute. Available from: https://www.cancer.gov/types/kidney/patient/kidney-treatment-pdq. [Last accessed on 2021 Jan 28].

Arruebo, M., Vilaboa, N., Sáez-Gutierrez, B., Lambea, J., Tres, A., Valladares, M. and González-Fernández, Á. 2011. Assessment of the evolution of cancer treatment therapies. *Cancers*, 3, 3279-3330.

Bhargava, N., Sharma, G., Bhargava, R. and Mathuria, M. 2013. Decision tree analysis on j48 algorithm for data mining. *Proc. Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, 3, 6.

Bhowmick, S.S., Saha, I., Bhattacharjee, D., Genovese, L.M. and Geraci, F. 2018. Genome-wide analysis of NGS data to compile cancer-specific panels of miRNA biomarkers. *PLoS One*, 13, e0200353.

Bifet, A. and Gavalda, R. 2009, August. Adaptive learning from evolving data streams. In: International Symposium on Intelligent Data Analysis. Springer, Berlin, Heidelberg. pp. 249-260.

Cairns, P. 2011. Renal cell carcinoma. *Cancer Biomark*., 9, 461-473.

Campbell, D. 2014. Almost Half of Cancer Patients Diagnosed Too Late, The Guardian. Available from: https://www.theguardian.com/society/2014/sep/22/cancer-late-diagnosis-half-patients. [Last accessed on 2021 Jan 29].

Chang, I., Mitsui, Y., Fukuhara, S., Gill, A., Wong, D.K., Yamamura, S., Shahryari, V., Tabatabai, Z.L., Dahiya, R., Shin, D.M. and Tanaka, Y. 2015. Loss of miR-200c up-regulates CYP1B1 and confers docetaxel resistance in renal cell carcinoma. *Oncotarget*, 6, 7774.

Fedorko, M., Pacik, D., Wasserbauer, R., Juracek, J., Varga, G., Ghazal, M. and Nussir, M.I. 2016. MicroRNAs in the pathogenesis of renal cell carcinoma and their diagnostic and prognostic utility as cancer biomarkers. *Int. J. Biol. Markers*, 31(1), 26-37.

Luan, J., Wang, J., Su, Q., Chen, X., Jiang, G. and Xu, X. 2016. Meta-analysis of the differentially expressed microRNA profiles in nasopharyngeal carcinoma. *Oncotarget*, 7, 10513.

Martinez, A., Fullwood, P., Kondo, K., Kishida, T., Yao, M., Maher, E.R. and Latif, F. 2000. Role of chromosome 3p12-p21 tumour suppressor genes in clear cell renal cell carcinoma: Analysis of VHL dependent and VHL independent pathways of tumorigenesis. *Mol. Pathol.,* 53, 137.

Morii, K., Yamasaki, S., Doi, S., Irifuku, T., Sasaki, K., Doi, T., Nakashima, A., Arihiro, K. and Masaki, T. 2019. microRNA-200c regulates KLOTHO expression in human kidney cells under oxidative stress. *PLoS One*, 14, e0218468.

Moldovan, L., Batte, K.E., Trgovcich, J., Wisler, J., Marsh, C.B. and Piper, M. 2014. Methodological challenges in utilizing mi RNA s as circulating biomarkers. *J. Cell. Mol. Med*., 18, 371-390.

Muhamed Ali, A., Zhuang, H., Ibrahim, A., Rehman, O., Huang, M. and Wu, A. 2018. A machine learning approach for the classification of kidney cancer subtypes using miRNA genome data. *Appl. Sci*., 8, 2422.

Nakada, C., Matsuura, K., Tsukamoto, Y., Tanigawa, M., Yoshimoto, T., Narimatsu, T., Nguyen, L.T., Hijiya, N., Uchida, T., Sato, F. and Mimata, H. 2008. Genome-wide microRNA expression profiling in renal cell carcinoma: Significant down-regulation of miR-141 and miR-200c. *J. Pathol*., 216, 418-427.

O'Brien, J., Hayder, H., Zayed, Y. and Peng, C., 2018. Overview of microRNA biogenesis, mechanisms of actions, and circulation. *Front. Endocrinol*., 9, 402.

Paranjape, T., Slack, F.J. and Weidhaas, J.B. 2009. MicroRNAs:

Tools for cancer diagnostics. *Gut*, 58, 1546-1554.

Rehman, O., Zhuang, H., Muhamed Ali, A., Ibrahim, A. and Li, Z. 2019. Validation of miRNAs as breast cancer biomarkers with a machine learning approach. *Cancers*, 11, 431.

Sanchez, D.J. and Simon, M.C. 2018. Genetic and metabolic hallmarks of clear cell renal cell carcinoma. *Biochim. Biophys. Acta Rev. Cancer*, 1870, 23-31.

Strimbu, K. and Tavel, J.A. 2010. What are biomarkers? *Curr. Opin. HIV AIDS*, 5, 463.

van IJzendoorn, D.G., Szuhai, K., Briaire-de Bruijn, I.H., Kostine, M., Kuijjer, M.L. and Bovée, J.V. 2019. Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. *PLoS Comput. Boil.*, 15, e1006826.

Wach, S., Nolte, E., Theil, A., Stöhr, C., Rau, T.T., Hartmann, A., Ekici, A., Keck, B., Taubert, H. and Wullich, B. 2013. MicroRNA profiles classify papillary renal cell carcinoma subtypes. *Br. J. Cancer*, 109, 714-722.

Witten, I.H., Frank, E., Hall, M.A. and Pal, C.J. 2005. Practical Machine Learning Tools and Techniques. Vol. 578. Morgan Kaufmann, United States. p. 1.

Zakrzewska, M., Gruszka, R., Stawiski, K., Fendler, W., Kordacka, J., Grajkowska, W., Daszkiewicz, P., Liberski, P.P. and Zakrzewski, K. 2019. Expression-based decision tree model reveals distinct microRNA expression pattern in pediatric neuronal and mixed neuronal-glial tumors. *BMC Cancer*, 19, 1-11.

Zheng, X., Fu, X., Wang, K. and Wang, M. 2020. Deep neural networks for human microRNA precursor detection. *BMC Bioinform.*, 21, 1-7.

How to cite this article: Pandya, H., Thalji, Z., Barron, M. 2021. Identifying a novel microRNA biomarker for renal cell carcinoma using a machine learning approach. *Int. J. Bioinform. Biol. Sci*. 9(1), 9-15.