



OPEN

## Panels of mRNAs and miRNAs for decoding molecular mechanisms of Renal Cell Carcinoma (RCC) subtypes utilizing Artificial Intelligence approaches

Seyed Mahdi Hosseiniyan Khatibi<sup>1,2,3</sup>, Mohammadreza Ardalan<sup>2</sup>, Mohammad Teshnehlab<sup>4</sup>, Sepideh Zununi Vahed<sup>2</sup>✉ & Saeed Pirmoradi<sup>1</sup>✉

Renal Cell Carcinoma (RCC) encompasses three histological subtypes, including clear cell RCC (KIRC), papillary RCC (KIRP), and chromophobe RCC (KICH) each of which has different clinical courses, genetic/epigenetic drivers, and therapeutic responses. This study aimed to identify the significant mRNAs and microRNA panels involved in the pathogenesis of RCC subtypes. The mRNA and microRNA transcripts profile were obtained from The Cancer Genome Atlas (TCGA), which were included 611 ccRCC patients, 321 pRCC patients, and 89 chRCC patients for mRNA data and 616 patients in the ccRCC subtype, 326 patients in the pRCC subtype, and 91 patients in the chRCC for miRNA data, respectively. To identify mRNAs and miRNAs, feature selection based on filter and graph algorithms was applied. Then, a deep model was used to classify the subtypes of the RCC. Finally, an association rule mining algorithm was used to disclose features with significant roles to trigger molecular mechanisms to cause RCC subtypes. Panels of 77 mRNAs and 73 miRNAs could discriminate the KIRC, KIRP, and KICH subtypes from each other with 92% (F1-score  $\geq 0.9$ , AUC  $\geq 0.89$ ) and 95% accuracy (F1-score  $\geq 0.93$ , AUC  $\geq 0.95$ ), respectively. The Association Rule Mining analysis could identify miR-28 (repeat count = 2642) and CSN7A (repeat count = 5794) along with the miR-125a (repeat count = 2591) and NMD3 (repeat count = 2306) with the highest repeat counts, in the KIRC and KIRP rules, respectively. This study found new panels of mRNAs and miRNAs to distinguish among RCC subtypes, which were able to provide new insights into the underlying responsible mechanisms for the initiation and progression of KIRC and KIRP. The proposed mRNA and miRNA panels have a high potential to be as biomarkers of RCC subtypes and should be examined in future clinical studies.

Renal Cell Carcinoma (RCC) possesses the 15th rank among frequent cancers, according to the GLOBOCAN report<sup>1</sup>. RCC comprises 3% of all malignant neoplastic cases in adults and approximately 90% of malignant kidney tumors<sup>2</sup>. Based on the American Cancer Society estimation, 1 in 46 men and 1 in 80 women will be diagnosed with RCC during their life. Also, the 5-year survival rate of patients with RCC is 73.7%<sup>3</sup>. Smoking, obesity, and hypertension are important risk factors that can affect RCC occurrence<sup>4</sup>. The incident and mortality rate of RCC highlight the importance of screening programs to develop reliable biomarkers for early detection of its subtypes<sup>5</sup>.

The histological subtypes of renal cancer include; clear-cell RCC (ccRCC or KIRC, 60–80% of all patients), papillary RCC (pRCC or KIRP, 10–15%), chromophobe RCC (chRCC or KICH, 5–10%), and other rare subtypes (< 1%)<sup>6</sup>. KIRC is identified by mutations in the *VHL* (von Hippel–Lindau) gene and the loss of chromosome 3p. The clinical severity of KIRC is higher compared with KIRP and KICH subtypes (5-year survival rate of 55–60%, 80–90%, and 90%, respectively)<sup>7</sup> due to the lack of effective biomarkers for earlier detection. In the early stages, KIRC is frequently asymptomatic and 25–30% of patients are often diagnosed at metastasis status;

<sup>1</sup>Clinical Research Development Unit of Tabriz Valiasr Hospital, Tabriz University of Medical Sciences, Tabriz, Iran. <sup>2</sup>Kidney Research Center, Tabriz University of Medical Sciences, Daneshgah Street, Tabriz Postal Code 51665118, Iran. <sup>3</sup>Rahat Breath and Sleep Research Center, Tabriz University of Medical Science, Tabriz, Iran. <sup>4</sup>Department of Electric and Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran. ✉email: sepide.zununi@gmail.com; said.pirmoradi@gmail.com

Genomic Data	Number of patients with Renal Cell Carcinoma (RCC)		
	pRCC (KIRP)	ccRCC (KIRC)	chRCC (KICH)
mRNA	321	611	89
miRNA	326	616	91

**Table 1.** Subtypes information of RCC in detail.

hence, presenting a high mortality rate<sup>8</sup>. KIRP is characterized by the loss of chromosome 9p and trisomy of chromosomes. In KIRP, a few subgroups of patients have a satisfactory treatment outcome, while a wide range of patients needs to develop promising treatment strategies<sup>9,10</sup>. Small incidentally detected kidney masses have a major diagnostic dilemma since a section of them can be benign and managed conservatively<sup>11</sup>. A lower-risk RCC, KICH, is identified by the loss of chromosomes and can pose a slight risk to a patient if cautiously treated with routine surveillance rather than surgery<sup>11</sup>. KICH has enormous potential to be diagnosed earlier compared to other RCC subtypes due to its variable long-term outcome<sup>12</sup>. Due to these distinct clinical and biological behaviors, differentiating and accurate detection of RCC subtypes via non-invasive and precise biomarkers confidently help physicians perform an appropriate therapeutic decision. This could be different from total and partial nephrectomy or even close follow-ups.

It has been shown that candidate biomarkers are not reliable for the diagnosis and prognosis of different RCC subtypes<sup>13–16</sup> and none of them can be employed in clinical application<sup>17</sup>. Furthermore, computed tomography (CT) and abdominal ultrasound have problems including high costs and low sensitivity in the diagnosis of small tumors<sup>4,18</sup>. Thus, the development of non-invasive and accurate screening for different types of RCC is a necessity. Comprehensive understanding of the molecular mechanisms of RCC subtypes is the main challenge in RCC research to identify novel and reliable molecular biomarkers.

This study was designed to introduce panels of mRNAs and miRNAs for discriminating different subtypes of RCC and detect remarkable molecules that have significant pathological roles in RCC subtypes through machine learning and artificial intelligence approaches. Firstly, we implemented reading and pre-processing of the RNA-sequencing data. Next, to identify candidate features (mRNAs and miRNAs), we applied feature selection based on filter and graph algorithms. Then, to evaluate selected candidate features, we employed a deep learning model to classify the subtypes of RCC. Finally, an association rule mining algorithm was used for detecting remarkable features which play significant roles in firing molecular mechanisms to cause RCC subtypes.

## Material and method

**Material.** In this study, the renal cell carcinoma data, including mRNA, miRNA, and clinical data were downloaded from the GDC portal (<https://portal.gdc.cancer.org>) provided by The Cancer Genome Atlas (TCGA) dataset<sup>19</sup>. Downloaded data were utilized to study the molecular mechanism of RCC subtypes, including ccRCC (KIRC), pRCC (KIRP), and chRCC (KICH). In the TCGA renal cell carcinoma project, 60,482 mRNAs expression and 1,881 miRNAs expression were reported for each patient. mRNA expression and clinical data were provided for 611 ccRCC patients, 321 pRCC patients, and 89 chRCC patients. Moreover, miRNA expression and clinical data were reported for 616 patients in the ccRCC subtype, 326 patients in the pRCC subtype, and 91 patients in the chRCC subtype. Information on each subtype is presented in Table 1 in more detail. This study was conducted according to the principles of the Declaration of Helsinki (2013).

**Method.** The proposed method contained four main steps: reading and preprocessing, feature selection, classification, and filtering, as shown in Fig. 1. In the 1st and 2nd steps, miRNA, mRNA, and clinical data were preprocessed using the necessary preprocessing methods. In the 3rd step, irrelevant and redundant features (mRNAs and miRNAs) were removed from the data using the proposed algorithms. In the 4th step, the deep classifier model was applied to distinguish defined groups based on the candidate features obtained in the previous step. In the 5th step, the association rule mining algorithm was employed for identifying the predominant features among candidate features playing a crucial role in each group.

**Reading and preprocessing.** miRNA and mRNA data were interpreted as a matrix composed of rows and columns, with approximately equal numbers of rows/columns; 1033/1881 and 1021/60,482, respectively. Next, redundant features were removed from the data, including 2256 mRNAs from 60,482 mRNAs and 336 miRNAs from 1881 miRNAs. Then, the hold-out cross-validation method was utilized to split data into three parts; training, validation, and test data by considering 70%, 10%, and 20% contributions, respectively. Finally, the min–max was used to scale mRNAs and miRNAs values in the [0 1] range. We applied z-score [Eq. (1)] for normalization and min–max [Eq. (2)] in feature selection and classification steps based on the need of the proposed algorithms in these steps, respectively.

$$y = \frac{x - \mu}{\sigma} \quad (1)$$

$$y = \frac{1}{Max - Min}(x - Min) \quad (2)$$

In Eqs. (1) and (2),  $y$  and  $x$  are the normalized and raw feature values, respectively. The  $\mu$ ,  $\sigma$ ,  $Min$ , and  $Max$  are defined as mean value, standard deviation, minimum value, and maximum value of  $x$ , sequentially.

**Feature selection.** In feature selection, we applied a new graph-based method for feature selection in mRNA and miRNA data separately<sup>20</sup>. Therefore, a suitable subset of candidate features was identified using the proposed algorithms. In miRNA data, a graph-based algorithm extracted the suitable subset of candidate features among 1545 features. However, in mRNA data, we first utilized the filter-based method for primary feature selection due to the existence of 58,226 features that could cause a high computational cost in the graph-based algorithm. Thus, the filter-based approach reduced the mRNA data dimension from 58,226 to 1000, and a graph-based algorithm extracted the suitable subset of candidate features among 1000 features. In the following, the proposed algorithms are specified in more detail.

**Filter method.** Filter-based methods interpret the feature selection process by calculating importance measures for each feature separately. These algorithms do not apply any evaluation tool such as a classifier and are called classifier-independent techniques. Filter-based methods have low computational and time costs; however, they do not consider features interactions in the feature selection process. In this sub-step, we applied the filter-based method for primary feature selection in mRNA data. Filter-based method help to remove some irrelevant features and reduce the dimensionality for use in the graph-based process. Many filter methods have been introduced and utilized in various data with different domains, but a few are suitable for high dimension and low sample size (HDLSS) data<sup>21</sup>. In this regard, we used the *AMGM* measure to evaluate the importance of features. The *AMGM* filter has illustrated its powerful potential in HDLSS data<sup>22</sup>. The *AMGM* measure is calculated by Eq. (3).

$$AMGM:R_i = \frac{AM_i}{GM_i} \in [1, +\infty) \quad (3)$$

In Eq. (3), the  $AM_i$  and  $GM_i$  are arithmetic mean, and the geometric mean of the  $i$ th feature are shown in Eqs. (4) and (5), respectively.  $R_i$  presents a dispersion of the  $i$ th feature among all samples. The higher  $R_i$  mentions a high dispersion and more relevant feature concerning defined phenotype. When  $R_i$  is close to one, which means the  $i$ th feature has low relevancy with a defined phenotype.

$$AM_i = \bar{x}_i = \frac{1}{n} \sum_{i=1}^n x_{ij} \quad (4)$$

$$GM_i = \left( \prod_{i=1}^n x_{ij} \right)^{\frac{1}{n}} \quad (5)$$

If the  $i$ th feature contains zero among reported values, then  $GM_i = 0$  based on Eq. (5) and *AMGM* measure will be inefficient. The modified version of *AMGM* is defined to avoid this problem based on Eq. (6). In the revised version, the exponential function was applied to features in the numerator and denominator of the *AMGM* formula.

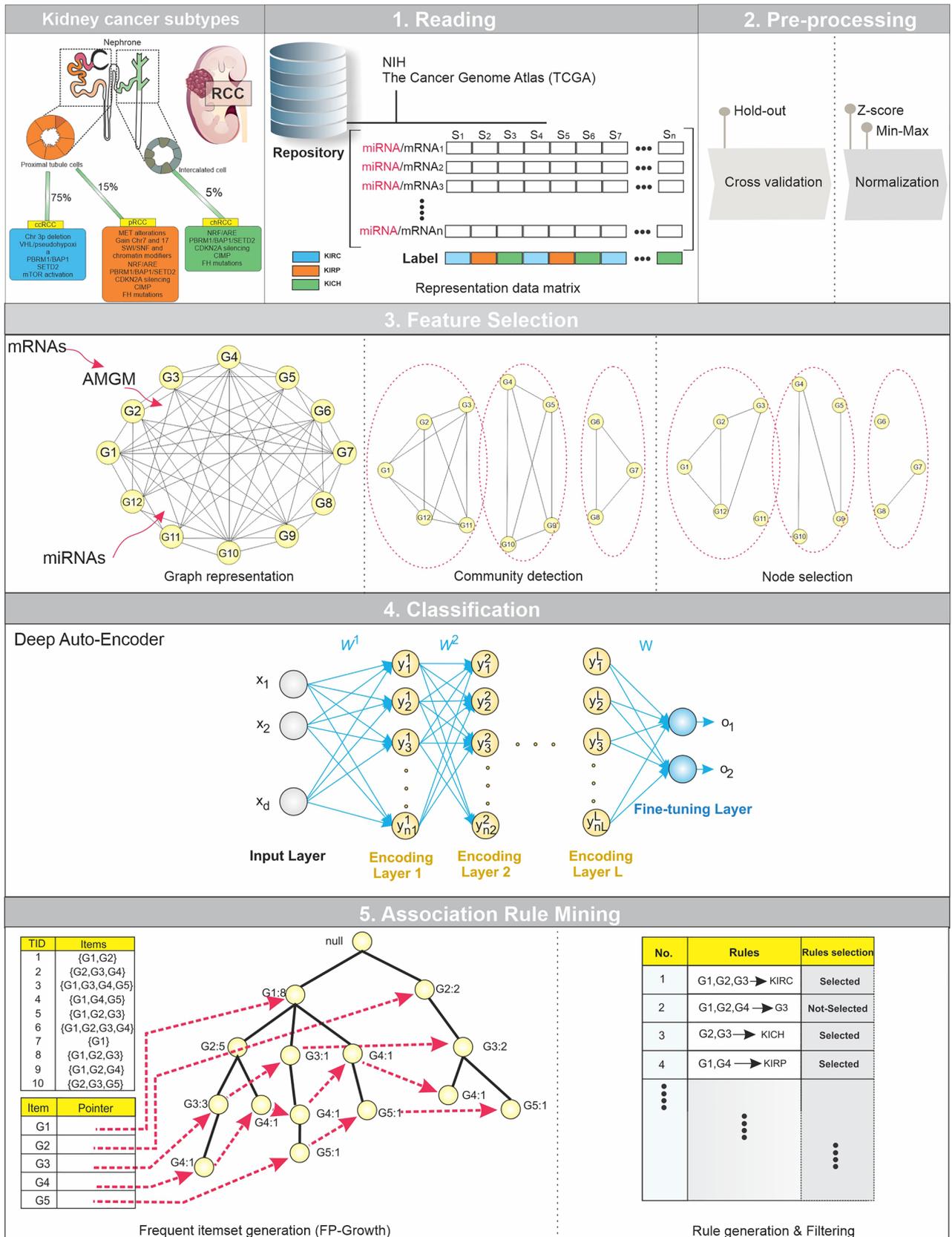
$$AMGM:R_i = \frac{\frac{1}{n} \sum_{i=1}^n \exp(x_{ij})}{\left( \prod_{i=1}^n \exp(x_{ij}) \right)^{\frac{1}{n}}} = \frac{1}{n \times \exp(\bar{x}_i)} \sum_{i=1}^n \exp(x_{ij}) \quad (6)$$

**Graph method.** Recently, graph-based methods are used for feature selection<sup>20,23</sup>. These methods display the search space as a graph using a graph representation of features. Then the principles of graph theory are employed for selecting the most relevant attributes. The graph-based method contains three sub-steps, including graph representation of feature space, community detection/graph clustering, and selecting the significant nodes in each cluster.

**Graph representation.** In the first phase, the feature set is mapped to the graph space. The graph is defined by  $G = \langle F, E \rangle$ , in which  $F = \{F_1, F_2, \dots, F_n\}$  and  $E = \{(F_i, F_j) : F_i, F_j \in F\}$  is a set of nodes and edges, respectively. Each feature represents a node, and relevance between two features deposes an edge in the graph structure. In this work, the  $W_{ij}$  is used to measure the relationship between two features  $F_i$  and  $F_j$ , as shown in Eq. (7).

$$W_{ij} = \begin{cases} \beta \times \text{Relevancy} - (1 - \beta) \times \text{Redundancy} & \text{if } i \neq j \\ 1 & \text{otherwise} \end{cases} \quad (7)$$

$$\text{Relevancy} = \frac{AMGM(F_i) + AMGM(F_j)}{2} \quad (8)$$



◀ **Figure 1.** The overview of the proposed method: Five main steps, including reading, preprocessing, feature selection, classification, and association rule mining were applied to miRNA and mRNA expression data. (1) In the reading step, each dataset was downloaded from the TCGA repository. (2) The preprocessing step includes two sub-steps, cross-validation, and normalization. (3) The feature selection step contains two sub-steps, the filter method based on AMG value for mRNA data and the graph-based method, in which candidate miRNAs/mRNAs with more relevance to RCC subtypes were selected. (4) A deep classifier model was utilized to evaluate the discrimination power of selected miRNAs/mRNAs. (5) The Association Rule Mining method discovers the hidden relationship between selected miRNAs/mRNAs and RCC subtypes in the first level and the complex relationship among selected miRNAs/mRNAs in the second level.

$$\text{Redundancy} = \left| \frac{\langle F_i, F_j \rangle}{\|F_i\| \times \|F_j\|} \right| = \cos(\theta_{F_i, F_j}) \quad (9)$$

$W_{ij}$  contains two parts, including relevancy and redundancy. In relevancy measure [Eq. (8)],  $AMGM(F_i)$  and  $AMGM(F_j)$  are the  $AMGM$  value of features  $F_i$  and  $F_j$ , respectively. Unlike the Bakhshandeh et al. that used symmetric uncertainty measure<sup>20</sup>, we applied the  $AMGM$  to measure relevancy; due to its high potential in HDLSS data. In redundancy measure, the similarity of two features is calculated based on the cosine similarity by Eq. (9), in which  $\langle \cdot, \cdot \rangle$  and  $\|\cdot\|$  are the inner product and Euclidean norm, respectively. Also,  $\beta$  ( $0 < \beta < 1$ ) is the user-defined parameter that controls the potency of each part.

$\widehat{W}_{ij}$  is the normalized relationship between two features in the range [0 1], calculated by the SoftMax scaling function, as shown in Eq. (10). We employed  $\widehat{W}_{ij}$  as the weight of edge in the graph structure. In the SoftMax scaling,  $\overline{W}$ ,  $\sigma$  are the mean and standard deviation of weights, respectively.

$$\widehat{W}_{ij} = \frac{1}{1 + \exp\left(-\frac{W_{ij} - \overline{W}}{\sigma}\right)} \quad (10)$$

**Community detection.** In the second phase, we applied a community detection algorithm to cluster graph space. The community detection algorithm performs clustering by finding groups of nodes that have high-density connectors internally. Nodes in the same community have high similarity properties. Community detection assists in better understanding sophisticated networks, such as genomic data networks. In recent years, researchers have introduced various community detection algorithms. Louvain algorithm is one of the fastest among them. Louvain algorithm performs the clustering process by maximizing the modularity objective function, in which the quality of partitions is compared, by the community detection process<sup>24</sup>. In addition, the Louvain algorithm is simple in the implementation phase.

In a graph with  $n$  nodes, the Louvain algorithm starts firstly with  $n$  communities by allocating each node to one community. The algorithm works based on a random selection of a node and transferring from its community to another one, then the gain modularity is calculated. This process is repeated until no improvement happens, and the algorithm will finish.

**Node selection.** In the third phase, we need to select the significant nodes in each community/cluster. In this regard, the Maximum Independent set (MIS) concept based on graph theory is utilized for node selection in subgraphs. A subset of the vertex set of a graph is independent if and only if it includes no pair of adjacent vertices. Identifying the independent set with the maximum size is an NP-hard optimization problem. It is doubtful that there exists an efficient algorithm for finding an MIS of a graph. In this study, we applied the proposed algorithm of R. Boppana et al. based on<sup>25</sup> to get the MIS in each community/cluster. Also, we defined the adjacency matrix to employ in the R. Boppana algorithm. It is a Boolean graph matrix that is calculated by Eqs. (11) and (12). Where  $\gamma$  is the user-defined parameter ( $0 < \gamma < 1$ ).

$$\text{Adjacency Matrix} = [a_{ij}]_{n \times n} \quad (11)$$

$$a_{ij} = \begin{cases} 1 & \widehat{W}_{ij} > \gamma \text{ and } i \neq j \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

**Classification.** In this step, we applied a classifier to evaluate the candidate features selected in the previous step. High accuracy (or any user-defined measure) of classification can mark the success of the feature selection method in choosing the relevant attributes. Otherwise feature selection method cannot identify relevant features.

Employing this method, we constructed a self-organizing deep auto-encoder model to classify data based on candidate features. A self-organizing deep auto-encoder is a specific type of deep auto-encoder that can determine its structure automatically, including the number of neurons and layers<sup>26</sup>. The description details of the self-organizing deep auto-encoder are available in Supplementary Methods. First, the training process of the deep model and the model selection were performed by training and validation data, respectively. Next, the performance of the classification was estimated by employing test data. The accuracy,  $F_1$ -score, and AUC-ROC were applied to evaluate the classifier performance, as shown in Eqs. (13)–(16).

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + TN} \times 100 \quad (13)$$

$$F_1 - \text{score} = 2 \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (15)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$

where TP, TN, FP, and FN are True Positive, True Negative, False Positive, and False Negative, respectively. AUC-ROC is the area under the Receiver Operating Characteristic (ROC) curve.

**Association rule mining.** Association Rule Mining can extract efficient associations among data items. An association rule is defined in  $A \rightarrow C$  form, in which A and C are Antecedent and Consequent, respectively. If we consider A as the feature(s) and C as the feature(s)/user-defined phenotype, association rule mining can show interesting dependency between feature(s)-feature(s) and feature(s)-phenotype. Therefore, it is reasonable to apply the association rule mining-based method to identify important features among candidate features. In this regard, we applied the association rule mining algorithm to candidate features, which was obtained in the feature selection step. In the following, some principal concepts of association rule mining are introduced. Support, Confidence, and Lift are three important measures in association rule mining. Let  $I = \{i_1, i_2, \dots, i_d, y\}$  be a set of items,  $D = \{d_1, d_2, \dots, d_n\}$  be a dataset of n instances,  $F = \{f_1, f_2, \dots, f_m\}$  be the features space with m features, and  $Y = \{0, 1\}$  be the user-defined phenotype. The  $d_i$  can be presented as a tuple  $(X_i, y_i)$ , where  $X_i \in f_1 \times f_2 \times \dots \times f_m$  and  $y_i \in Y$ . Also,  $A \rightarrow C$  is an association rule, where  $A \subset I$ ,  $C \subset I$ , and  $A \cap B = \varnothing$ . The support of rule  $A \rightarrow C$  is the probability of instances containing both A and C, as shown in Eq. (17). Support evaluates the rule's usefulness.

$$\text{Support}(A \rightarrow C) = \frac{\text{support}(A \cup C)}{n} \quad (17)$$

In Eq. (17),  $\text{Support}(A) = |\{d_i | A \subseteq X_i, d_i \in D\}|$  is the number of instances that includes the  $i$  itemset. The confidence of rule  $A \rightarrow C$  is the probability, which shows the frequency of cases with C among all samples containing A, as shown in Eq. (18). Confidence estimates the rule's certainty.

$$\text{Confidence}(A \rightarrow C) = P(C|A) = \frac{\text{support}(A \cup C)}{\text{support}(A)} \quad (18)$$

The Lift of rule  $A \rightarrow C$  determines the dependency between the occurrence of itemset A and C. When the Lift value is more (less) than one, the occurrence of A is positively (negatively) correlated with the occurrence of C. If the Lift value is equal to one, then A and C are independent. The Lift value is shown in Eq. (19).

$$\text{Lift}(A \rightarrow C) = \frac{P(A \cup C)}{P(A)P(C)} \quad (19)$$

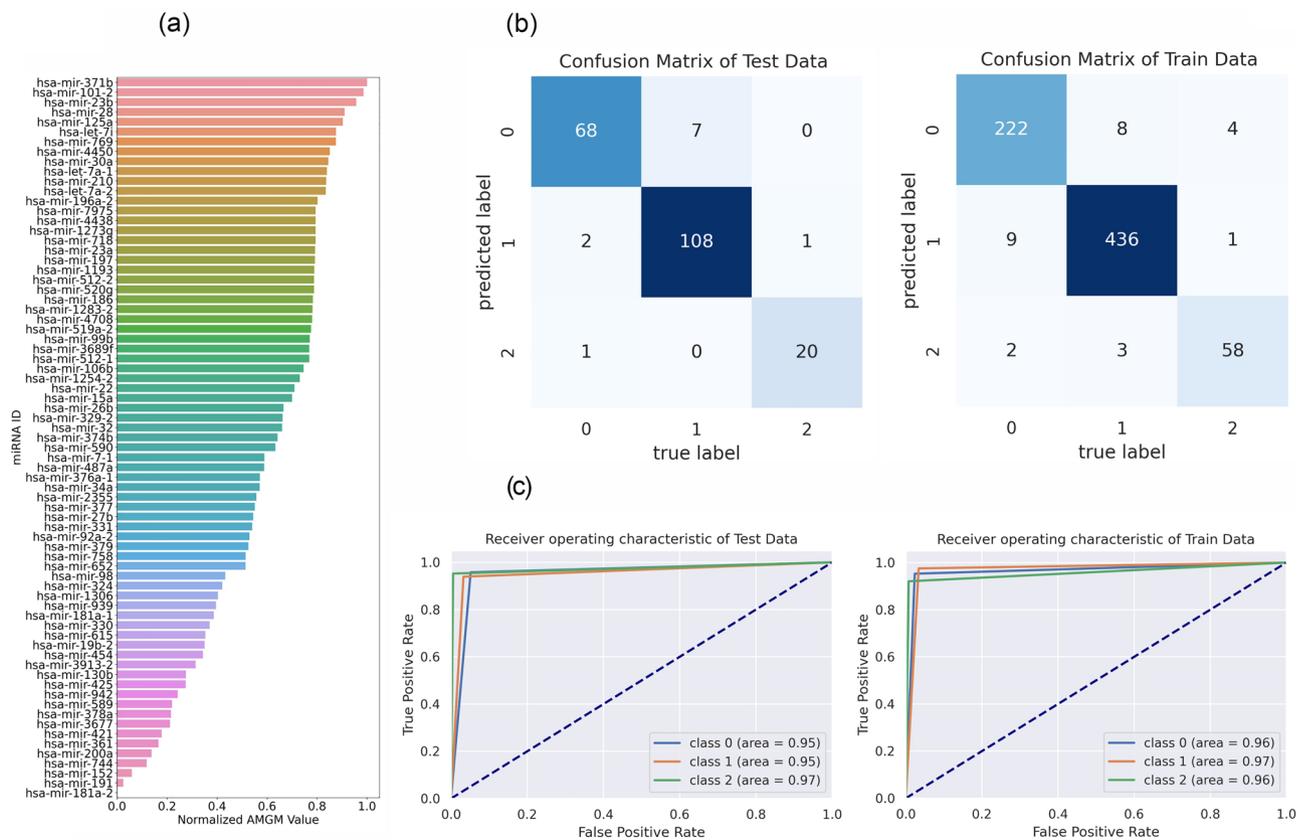
**FP-growth algorithm.** Association rule mining mainly contains two phases: Frequent itemset generation and Rule generation. In the Frequent itemset generation, the algorithm generates all itemsets iteratively, then itemset that its support count is more than the min\_support (user-defined threshold) reported as a frequent itemset. In the rule generation, association rules are made based on frequent itemsets. Agrawal et al. introduced the Apriori algorithm as one of the first association analysis algorithms, in which the property of frequent itemset is applied successfully by R. Agrawal<sup>27</sup>. In recent years many researchers have proposed many improved algorithms, such as FP-Growth<sup>28,29</sup>, Apriori-Hybrid<sup>30</sup>, Fuzzy association rule<sup>31</sup>, etc.

In this study, we utilized the FP-Growth algorithm for association analysis. In terms of computational, storage space, and time, the FP-Growth is one of the best algorithms for association analysis due to one-time searching itemset space. The pseudo-code of the FP-Growth algorithm is shown in Tables 1 and 2 in Supplementary method in more detail.

## Results

We executed the following steps to the miRNA and mRNA data (Fig. 1). In the feature selection step, we applied a graph-based method to the miRNA data. First, the graph was constructed with 1545 nodes and 1,194,285 edges, and the weights of the edges were calculated based on the AMGM and cosine similarity measures. Next, the Louvain algorithm identified 40 communities/clusters from each; and finally, 73 candidate miRNAs were selected among communities/clusters using the MIS algorithm. Moreover,  $\beta$  and  $\gamma$ , the parameters of the graph-based method were set to 0.6 and 0.3, respectively.

In mRNA data, first, we used the filter method based on the AMGM measure to remove some irrelevant features. In the primary feature selection, we selected 1000 top features with the highest AMGM value, then we



**Figure 2.** The performance of feature selection and classification steps in the miRNA data. **(a)** Bar plot of candidate miRNAs (73) based on normalized AMG values. **(b)** Confusion matrix of training and test data, in which 0, 1, and 2 are pointed to KIRP, KIRC, and KICH groups, respectively. **(c)** Receiver Operating Characteristic (ROC) curve of training and test data. The Area under the Curve of Receiver Operating Characteristic (AUC-ROC) is reported for each curve.

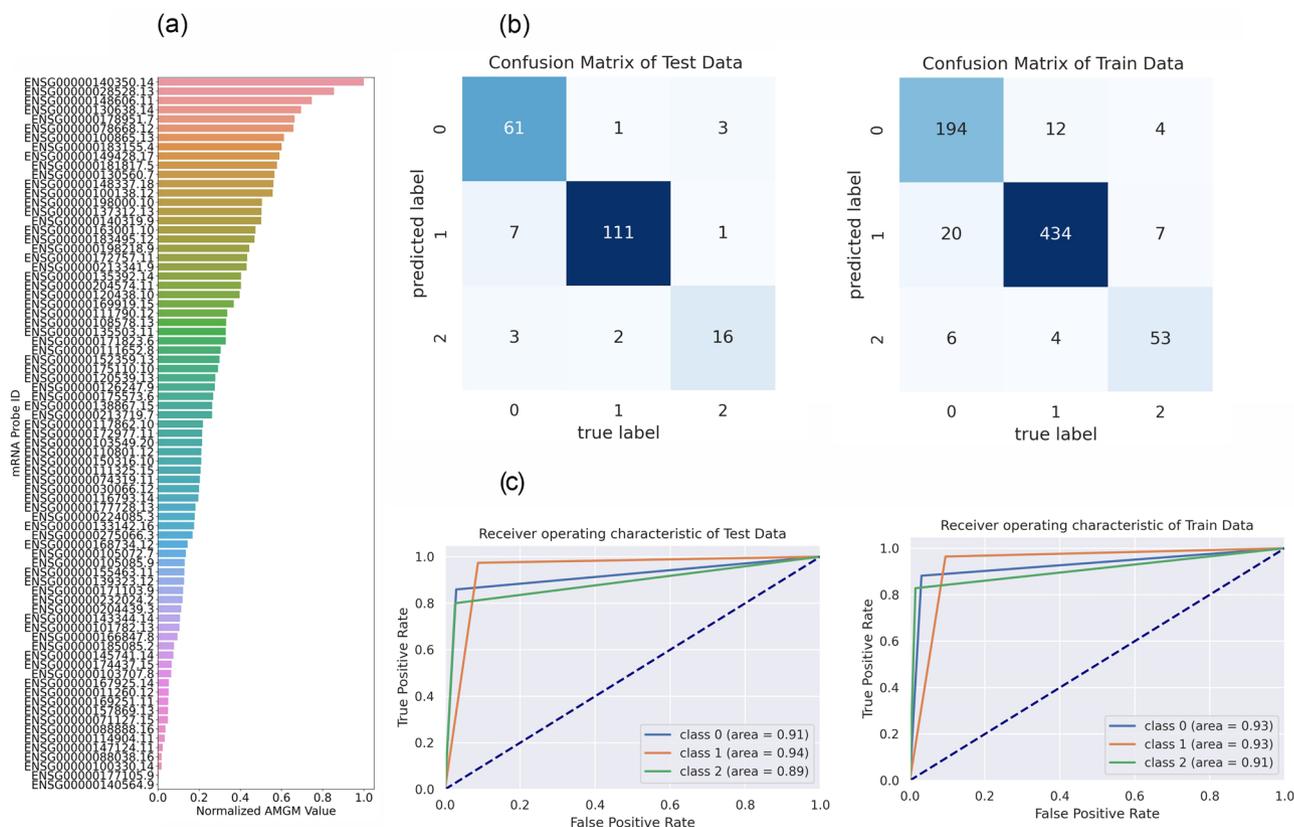
employed a graph-based method for the mRNA data. The graph was composed of 1000 nodes and 504,510 edges. Next, community detection of the mRNA network graph was performed by the Louvain algorithm that identified 73 communities/clusters. Finally, 77 candidate mRNAs were selected from communities/clusters using the MIS algorithm where  $\beta$  and  $\gamma$  parameters of the graph-based method, were set to 0.5 and 0.3, respectively. The list of 73 candidate miRNAs and 77 candidate mRNAs are reported in Supplementary Tables 1 and 2, respectively. Also, these candidate mRNAs and miRNAs are illustrated in Figs. 2a and 3a based on their sorted AMG measure.

We applied the classifier to evaluate the discrimination power of candidate features (mRNAs/miRNAs) among RCC subtypes. In this regard, the self-organizing deep auto-encoder was utilized for performing classification tasks. The accuracy,  $F_1$ -score, and AUC of mRNA and miRNA data are reported individually in Table 2. Also, the confusion matrix and ROC curve are illustrated for miRNA and mRNA data in Figs. 2 and 3, respectively. Results indicated that machine learning-derived mRNAs (Accuracy = 92%) and miRNAs (Accuracy = 95%) panels could significantly distinguish these subtypes from each other with high accuracy.

In similar studies, all subtypes, including ccRCC, pRCC, chRCC, WT (Wilms tumor), and RT (Rhabdoid tumor), were classified based on selected miRNAs in the feature selection process. However, these subtypes are different in cancer nature and patient type. The ccRCC, pRCC, and chRCC are common adult kidney cancer. In contrast, WT and RT are common pediatric kidney cancer. Thus, feature selection and classification of all subtypes may lead to missing information related to pediatric kidney cancer subtypes. In addition, the comparison classification accuracy of similar studies with the proposed method will be difficult due to this difference. Nevertheless, the accuracy of the applied methods is illustrated in Table 3. Also, we do not find any similar studies related to ccRCC, pRCC, and chRCC subtypes based on mRNA data for comparison.

In the association rule mining process, candidate features were scaled into the [0 1] range using min-max normalization. Next, candidate features (mRNAs/miRNAs) were discretized into three categories, including low, medium, and high levels, so that the number of miRNA and mRNA items were equal to 212 and 233, respectively. Then, the FP-Growth algorithm was applied to generate frequent itemsets and association rules. Parameters of the algorithm, including min-support (frequent itemset), max-length (maximum length of frequent itemset), and lift (association rule), were set to 0.1, 4, and 1.1, sequentially.

To discover patterns, association rules that consequently were equal to KIRC/KIRP/KICH were selected. In the miRNA data, the number of association rules related to KIRC/KIRP was equal to 27,635/23,198. Moreover, in the mRNA data, the number of association rules related to KIRC/KIRP was equal to 94,354/28,956. Due to



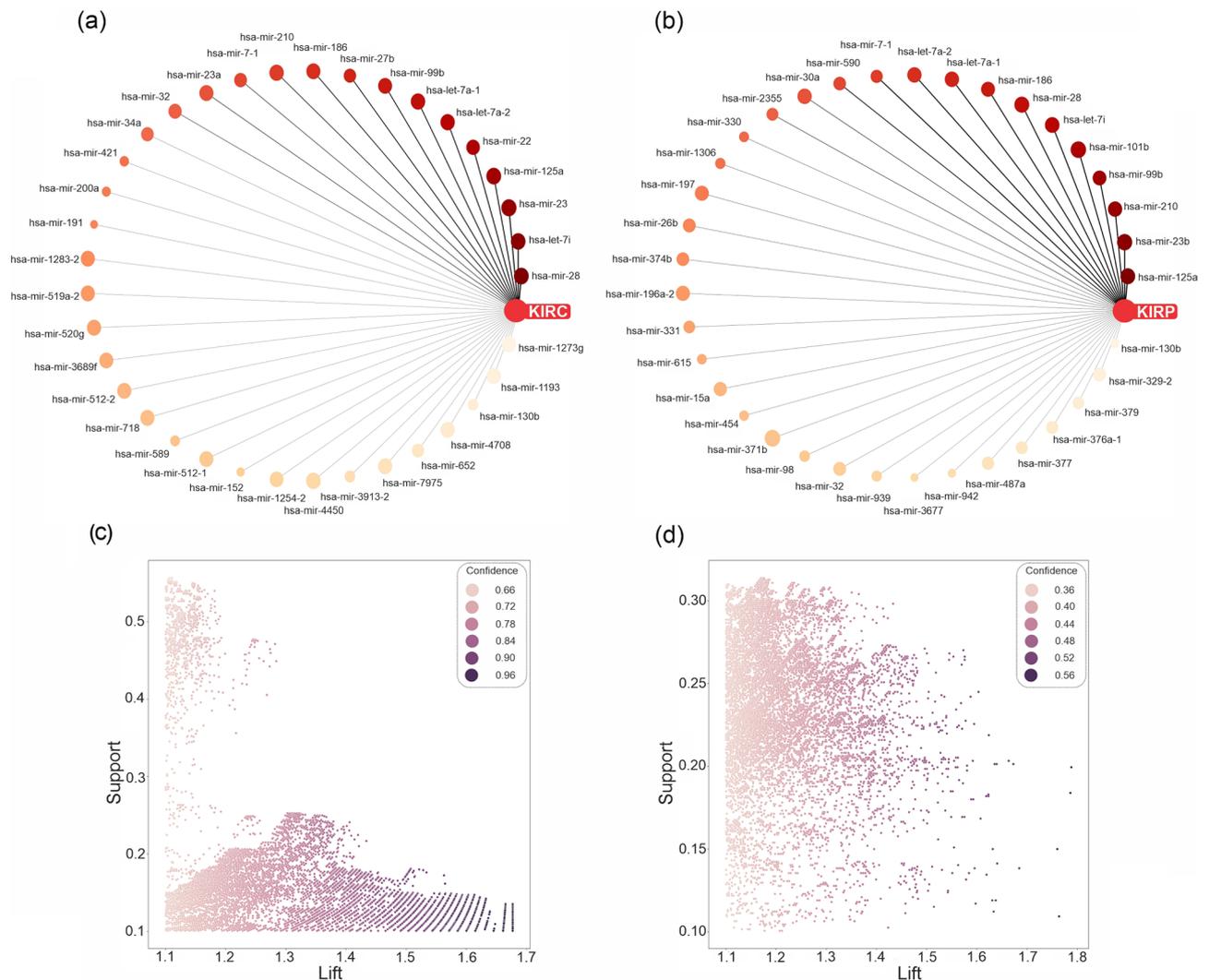
**Figure 3.** The presentation of feature selection and classification steps in the mRNA data. **(a)** Candidate mRNAs (77) are presented by Bar plot based on normalized AMGM values. **(b)** Training and test data confusion matrix. **(c)** ROC curve of training and test data. The AUC-ROC is reported for each curve. 0, 1, and 2 are pointed to KIRP, KIRC, and KICH groups, respectively. AUC-ROC Area under the Curve of Receiver Operating Characteristic.

	miRNA			mRNA		
	Train	Test	Validation	Train	Test	Validation
Accuracy (%)	96	<b>95</b>	96	93	<b>92</b>	91
F <sub>1</sub> -score (KIRP)	0.95	0.93	0.93	0.90	0.90	0.89
F <sub>1</sub> -score (KIRC)	0.98	0.96	0.98	0.95	0.95	0.95
F <sub>1</sub> -score (KICH)	0.92	0.95	0.92	0.83	0.78	0.75
AUC (KIRP)	0.96	0.95	0.94	0.93	0.91	0.91
AUC (KIRC)	0.97	0.95	0.96	0.93	0.94	0.94
AUC (KICH)	0.95	0.97	0.92	0.91	0.89	0.80

**Table 2.** The performance metrics for the classification step. Significant values are in bold. AUC Area under the Curve, KIRC kidney renal clear cell carcinoma, KIRP kidney renal papillary cell carcinoma, KICH kidney chromophobe carcinoma.

Methods	Accuracy (%)	Ref.
One way ANOVA + logistic regression (all subtypes)	86	32
One way ANOVA + decision tree (all subtypes)	87	33
Resampling + NCA + LSTM (all subtypes)	95.5	34
AMGM + deep neuro-fuzzy (all subtypes)	93.2	35
Proposed method (ccRCC, pRCC, and chRCC subtypes)	<b>95</b>	–

**Table 3.** Comparison of classification accuracy based on miRNA data. Significant values are in bold.

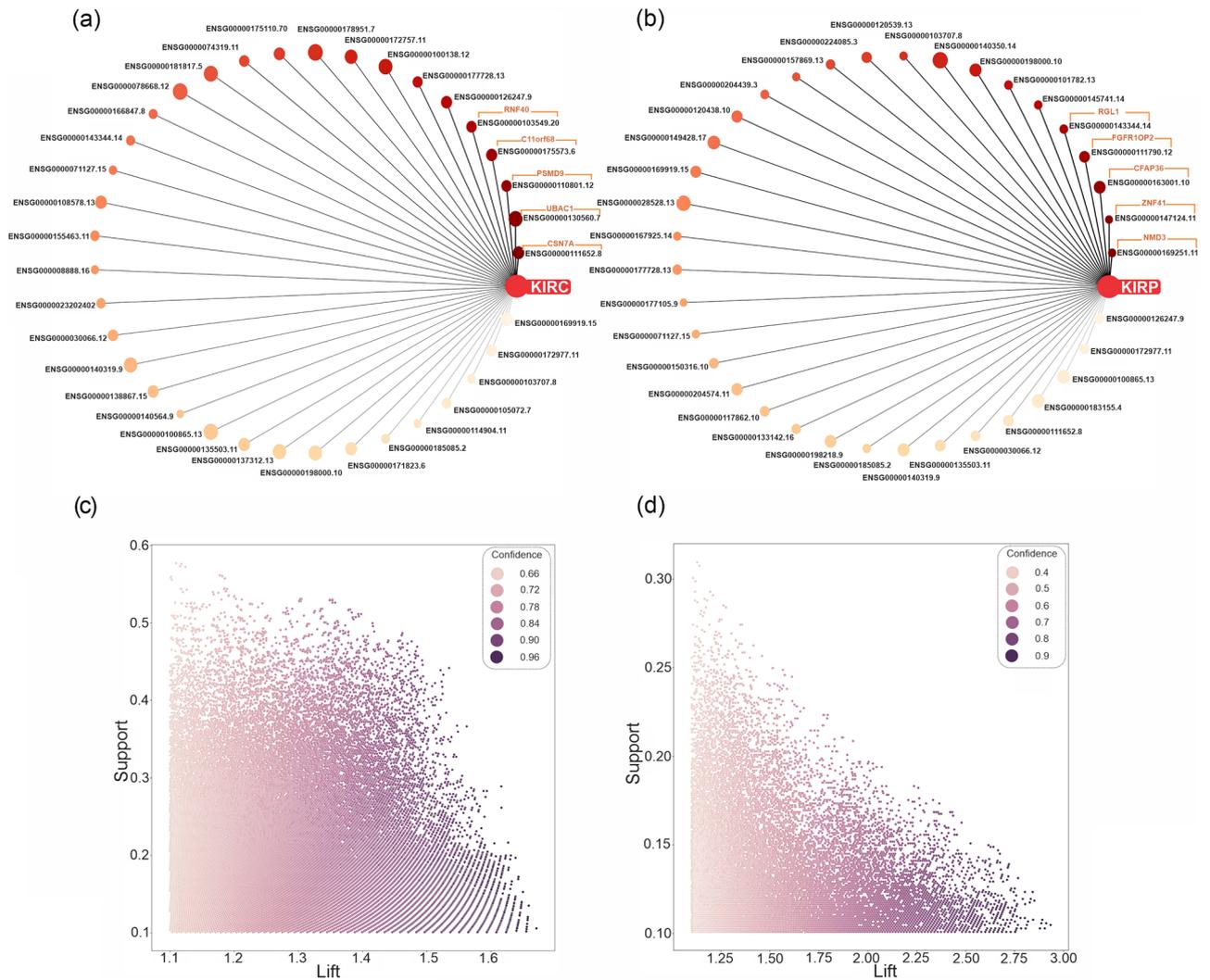


**Figure 4.** The relationship and specification of the KIRC/KIRP association rules in the miRNA data. Graph network of top miRNAs with high repeat count in (a) KIRC and (b) KIRP association rules. The most frequent miRNAs were displayed with the weight of the edge based on repeat counts. Also, the AMG value of miRNAs was illustrated with the size of the node. Strength distribution of (c) KIRC and (d) KIRP association rules according to their support, lift, and confidence. *KIRC* kidney renal clear cell carcinoma, *KIRP* kidney renal papillary cell carcinoma, *KICH* kidney chromophobe carcinoma.

the lack of samples in the *KICH* subtype, frequent itemsets and related association rules were not generated; the support count of itemsets was less than the min-support threshold. In the antecedent part of the selected association rules, mRNAs and miRNAs with the highest repeat count were considered significant features in each RCC subtype. In Figs. 4a,b and 5a,b, significant miRNAs and mRNAs are shown as a graph network based on repeat count in KIRC/KIRP rules. Moreover, the strength distribution of KIRC/KIRP association rules according to their support, lift, and confidence is illustrated in Figs. 4c,d and 5c,d for miRNA and mRNA, respectively. In this regard, mRNAs and miRNAs were identified based on sorted repeat count (Supplementary Table 3). We hypothesized that these top features with the highest repeat counts may play a fundamental role in the pathogenesis of specific subtypes.

Box plots of five top miRNAs and mRNAs in KIRC and KIRP subtypes are illustrated in Fig. 6. The comparison of medians, interquartile ranges, and whiskers in box plots of significant miRNAs and mRNAs demonstrated a notable difference in all RCC subtypes. The pair plot of five top miRNAs and mRNAs of KIRC/KIRP rules are illustrated in Supplementary Figs. 1c,d and 2c,d, respectively. Moreover, the correlation of the top ten miRNAs and mRNAs of KIRC/KIRP rules is shown in Supplementary Figs. 1a,b and 2a,b, respectively based on the Spearman correlation.

CSN7A (ENSG00000111652.8) and NMD3 (ENSG00000169251.11) were the most frequent itemset with 5774 and 2306 repeat counts in KIRC and KIRP association rules, respectively. Moreover, miR-28 and miR-125a were the most frequent itemset with 2642 and 2591 repeat counts in KIRC and KIRP association rules, respectively. As a result, we decided to examine these miRNAs and mRNAs more closely using other association rules to investigate their relation with another feature. More in-depth coverage of these findings is available in



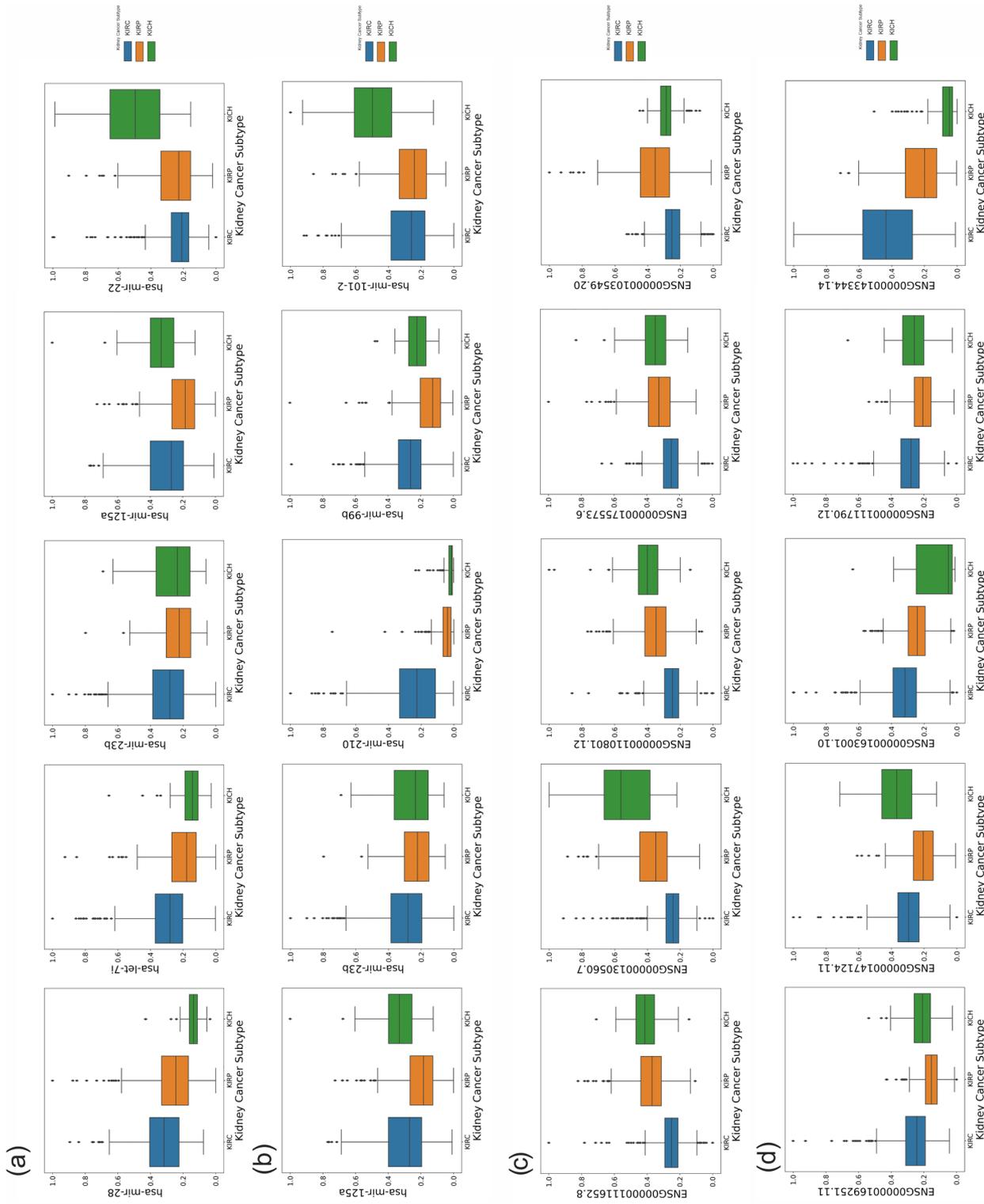
**Figure 5.** The KIRC/KIRP association rules in the mRNA data. Graph network of (a) KIRC and (b) KIRP association rules of top mRNAs with a high repeat count. The most frequent mRNAs were displayed with the weight of the edge based on repeat counts. The mRNAs’ AMGM value is illustrated with the size of the node. Strength distribution of (c) KIRC and (d) KIRP association rules according to their support, lift, and confidence. *KIRC* kidney renal clear cell carcinoma, *KIRP* kidney renal papillary cell carcinoma, *KICH* kidney chromophobe carcinoma.

the discussion section of the study. In this regard, we showed these relations based on association rules in the graph network (Figs. 7 and 8).

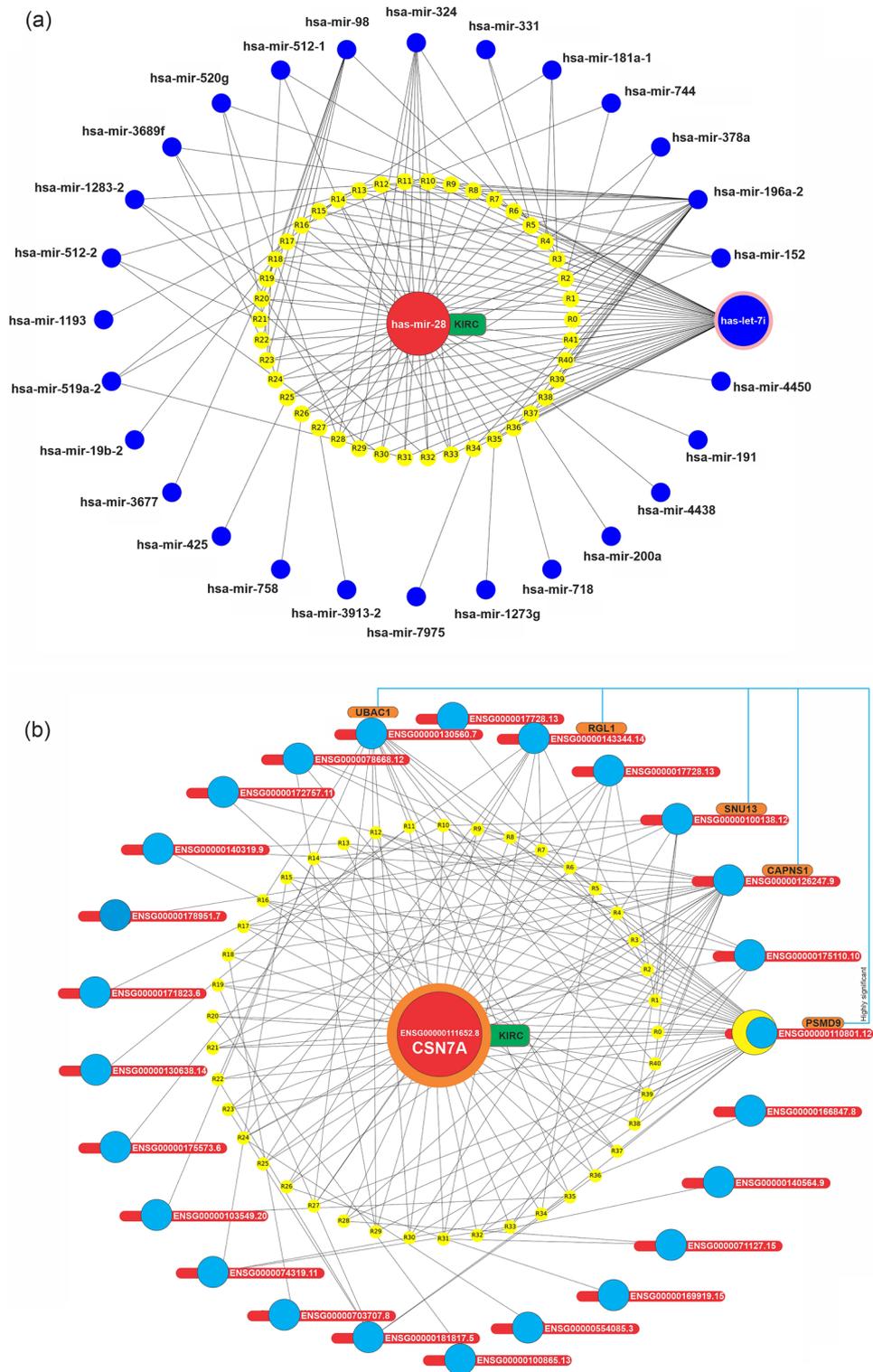
In Fig. 7a, it is obvious that hsa-miR-28, the most frequent miRNA in KIRC association rules, has a high dependency on hsa-let-7i, hsa-miR-196a-2, hsa-miR-324, hsa-miR-198, and hsa-miR-152, respectively. In Fig. 8a, it is clear that hsa-miR-125a, the most frequent miRNA in KIRP association rules, has a high dependency on hsa-miR-99b, hsa-miR-374b, hsa-miR-186, hsa-miR-28, and hsa-miR-32, respectively. In Fig. 7b, it is noticeable that CSN7A, the most frequent mRNA in KIRC association rules, has a high dependency on ENSG00000110801.12 (PSMD9), ENSG00000126247.9 (CAPNS1), ENSG00000130560.7 (UBAC1), ENSG00000100138.12 (SNU13), and ENSG00000143344.14 (RGL1: Ral guanine nucleotide dissociation stimulators like 1), respectively. In Fig. 8b, it is observable that ENSG00000169251.11 (NMD3), the most frequent miRNA in KIRP association rules, has a high dependency on ENSG00000126247.9 (CAPNS1), ENSG00000185085.2 INTS5 (integrator complex subunit 5), and ENSG00000163001.10 (CFAP36), respectively.

### Discussion

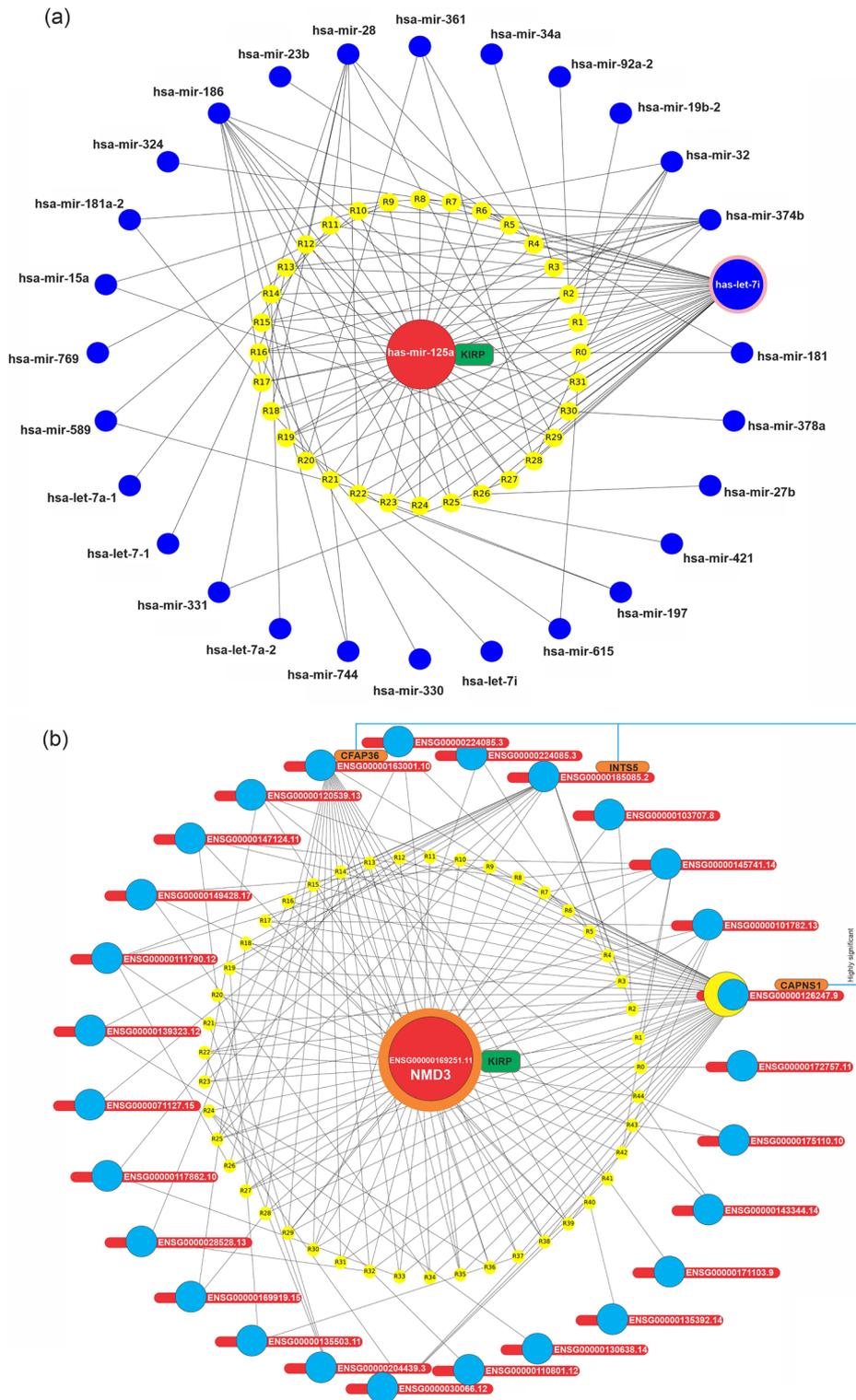
We carried out a comprehensive machine learning analysis of clinically significant patterns of the miRNAs and mRNAs within the RCC subtypes. A panel of 77 candidate mRNAs and a panel of 73 miRNAs could discriminate KIRC from KIRP with high accuracy. The association rule mining analysis could identify top mRNAs and miRNAs with the highest repeat counts, suggesting their possible pathological roles in each RCC subtype. The CSN7A and miR-28 along with the NMD3 and miR-125a were the most frequent itemsets in the KIRC and KIRP association rules, respectively. The roles of these mRNAs have not been studied before in the RCC. In the



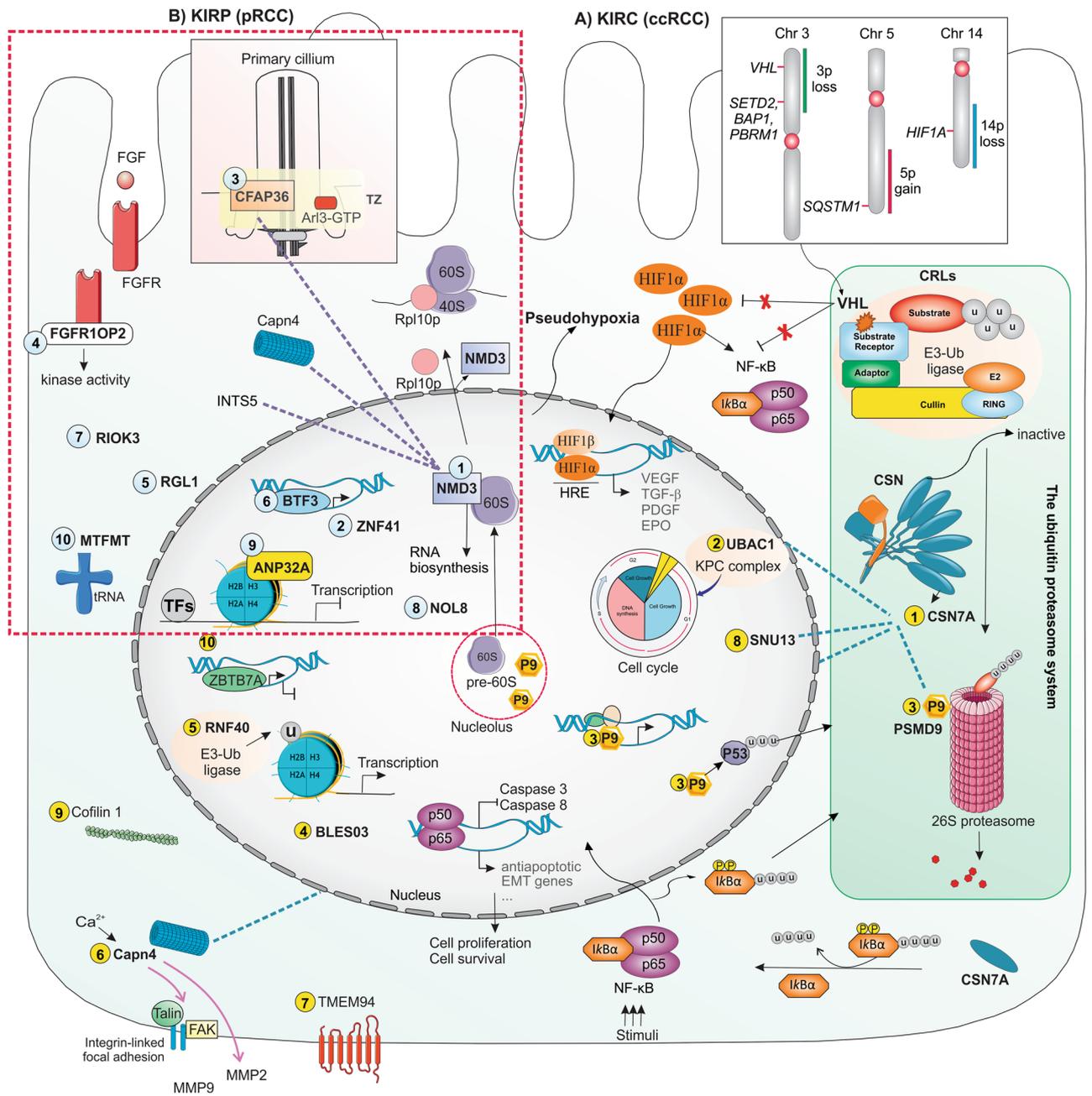
**Figure 6.** The selected miRNA and mRNA are based on the association rule mining step. Box plot of selected miRNAs for (a) KIRC and (b) KIRP groups. Box plot of selected mRNAs for (c) KIRC and (d) KIRP groups. Expression values of miRNAs and mRNAs were normalized in the range of 0 to 1 by the min-max method. KIRC kidney renal clear cell carcinoma, KIPN kidney renal papillary cell carcinoma, KICH kidney chromophobe carcinoma.



**Figure 7.** The relationship and specification of the miRNAs and mRNAs are based on the KIRC association rules. Graph network of (a) has-miR-28 (with support > 1.66) related association rules, in which the identified miRNA, its rules, and miRNAs were presented, in orange, yellow, and blue colors, respectively. (b) Graph network of ENSG00000111652.8 related association rules (with support > 0.3 and lift > 1.4), in which ENSG00000111652.8, rules, mRNAs were presented, in red, yellow, and blue colors, respectively.



**Figure 8.** The relationship and specification of the miRNAs and mRNAs are based on the KIRP association rules. Graph network of (a) has-miR-125a (with support > 1.26) related association rules, in which the identified miRNA, its rules, and miRNAs were presented, in orange, yellow, and blue colors, respectively. (b) Graph network of ENSG00000169251.1 related association rules (with support > 0.2 and lift > 1.145), in which ENSG00000169251.1, rules, mRNAs were presented, in red, yellow, and blue colors, respectively.



◀ **Figure 9.** Possible roles of the deep learning-derived mRNAs and miRNAs in the pathogenesis of KIRC and KIRP (identified by deep learning methods). **(A)** The *CSN7A*, *UBAC1*, *PSMD9*, *RNF40*, *Capn4*, *TMEMs*, *CFL1*, *CFL1*, and *ZBTB7A* were identified as candida mRNAs (indicated by 1–10) in KIRC with the highest repeat counts by the Association Rule Mining analysis. The UPS is composed of Ub, proteasome, ubiquitinating (E1, E2s, and E3s), and deubiquitination enzymes. The VHL is the substrate recognition subunit of a CRL complex that polyubiquitylates its targets (HIF and NF- $\kappa$ B) to be degraded. In KIRC, the *VHL* gene is mutated or downregulated that result in the accumulation of these targets, making a pseudohypoxia state and highly vascularized tumors. The *CSN7A*, *UBAC1*, and *PSMD9* are components of the UPS identified in this study. (1) *CSN7A* via I $\kappa$ B $\alpha$  deubiquitinylation can suppress the transcription activity of the NF- $\kappa$ B. (2) The *UBAC1* is involved in ubiquitination and degradation of the cell cycle proteins. (3) the *PSMD9* (P9) may be involved in the maintenance of integrity and morphology of nucleolus and impact cell survival and cell cycle regulation. The *COPS7A* in the KIRC association rules has a high dependency on *PSMD9*, *Capn4*, and *UBAC1*, *SNU13*, respectively (presented by dashed lines). **(B)** The *NMD3*, *ZNF41*, *CFAP36*, *FGFR1OP2*, *RGL1*, *BTF3*, *RIOK3*, *NOL8*, *ANP32A*, and *MTFMT* were identified as the top ten mRNA with high association rules in the KIRP. (1) The *NMD3* is a nuclear adaptor protein that transports the ribosomal 60S subunit into the cytoplasm and participates in the cytoplasmic maturation of ribosomal 60S particles. It also affects RNA biosynthesis, mainly ribosomal RNA, and may impact tumorigenesis. (2) *ZFP41* is a regulator of the transcription of different genes. (3) The *CFAP36* is a binding partner of the *Arl3*, a small GTPase in the primary cilia. The primary cilia participate in the cilia formation, regulation of the cell cycle, and mitosis of cancer cells. The *NMD3* in KIRP has a high dependency on *CAPNS1*, *INTS5*, and *CFAP36*, respectively (presented by dashed lines). For more details, please see the main text and Supplementary file, discussion part. *ANP32A* acidic nuclear phosphoprotein 32 family member A, *BLES03* basophilic leukemia-expressed protein, *BTF3* basic transcription factor 3, *CFAP36* Cilia and flagella-associated protein 36, *CAPN4* calpain small subunit 1, *CFL1* Cofilin1, *CRLs* E3 ligases are cullin-RING ligases, *CSN7A* COP9 signalosome subunit 7A, *FAK* focal adhesion kinase, *FGF* fibroblast growth factor, *FGFR1OP2* FGF receptor 1 oncogene partner 2, *HIF1* hypoxia inducible factors, *INTS5* integrator complex subunit 5, *MMP2/9* matrix metalloproteinase 2/9, *MTFMT* mitochondrial methionyl-tRNA formyltransferase, *NF- $\kappa$ B* nuclear factor kappa B, *NOL8* nucleolar protein 8, *PSMD9* proteasome 26S subunit, non-ATPase 9, *RGL1* Ral guanine nucleotide dissociation stimulators like 1, *RNF40* ring finger protein 40, *RIOK3* RIO kinase 3, *SNU13* small nuclear ribonucleoprotein 13, *TMEM94* transmembrane protein 94, *TZ* transition zone, *Ub* ubiquitin, *UBAC1* UBA domain containing 1, *UPS* ubiquitin–proteasome system, *ZBTB7A* zinc finger and BTB domain-containing 7A, *ZFP41* Zinc finger protein 41.

following sections, we present a brief discussion on the possible roles of these mRNAs and microRNAs in the pathogenesis of the KIRC and KIRP based on the published literature.

**Candidate RNAs in KIRC.** The *CSN7A*, *UBAC1*, *PSMD9*, *RNF40*, and *Capn4* were identified as candida mRNAs in the KIRC with the highest repeat counts by the association rule mining analysis. Machine learning approaches could detect novel targets in the field of the KIRC and bring the UPS (ubiquitin–proteasome system) and its components (*CSN7A*, *UBAC1*, *PSMD9*, and *RNF40*) into the focus of interest. The degradation of regulatory proteins by the UPS has an important function in controlling the cell cycle progression, DNA repair, response to extracellular stress, and signal transduction. The ubiquitin, proteasome, ubiquitinating enzymes (including ubiquitin-activating (E1), -conjugating (E2s), and -ligases (E3s) enzymes) along with deubiquitinating enzymes are the key components of the UPS system<sup>36</sup>. Most of the E3 ligases are cullin-RING ligases (CRLs) that determine the substrate-selectivity in response to specific stimuli either to degrade by the proteasome or modify their function. The COP9 signalosome (CSN), another component of the UPS, is a multi-subunit (CSN 1–8) metalloprotease complex<sup>37</sup> that plays roles in gene expression, cell-cycle control, and DNA-damage response<sup>38</sup>, and increases the stability of some proteins including EGFR<sup>39</sup>. Furthermore, the CSN plays a role in controlling the activity of NF- $\kappa$ B, an inflammatory transcriptional regulator involved in cell survival, proliferation, and transformation<sup>40,41</sup>.

The VHL is the substrate recognition subunit of the E3 ubiquitin ligase complex that polyubiquitylates the hydroxylated targets under normoxia<sup>42</sup>. Under the loss of the VHL, the degradation of its specific targets, hypoxia inducible factors (HIF1 $\alpha$  and HIF2 $\alpha$ ) does not occur, making a pseudohypoxia state. Both VHL deficiency and the accumulation of the HIF-1 promote the NF- $\kappa$ B activity that subsequently stimulates an NF- $\kappa$ B/PI3K/AKT/TGF- $\beta$ /EGFR/IKK signaling cascade, resulting in the activation of proliferation and glycolytic pathways, a pro-angiogenic and apoptosis-resistant phenotype, and highly vascularized tumors<sup>43,44</sup>. It is reported that the CSN elevates the efficacy of the VHL-mediated HIF-1 $\alpha$  recognizing, ubiquitination, and degradation<sup>45</sup>.

In our deep learning analysis, the *CSN7A* (*COPS7A*), 7a subunit of the CSN, was identified as the first top mRNA with the highest role in the KIRC. It is reported that in tumor tissues, the *CSN7A* level is decreased which may be associated with the oxidative phosphorylation pathway<sup>46</sup> and the transcription-coupled nucleotide excision DNA repair<sup>47</sup>. Moreover, an overexpressed *CSN7A* could stimulate the I $\kappa$ B $\alpha$  deubiquitinylation and; consequently, suppressing the transcriptional activity of the NF- $\kappa$ B<sup>48</sup>. The mechanism by which the *CSN7A* can impact KIRC has not been defined; however, the *CSN7A* may function alike in the KIRC.

The *UBAC1*, the second identified mRNA in this study, is a subunit of the KPC complex, an E3 ubiquitin-protein ligase. The *UBAC1* acts with the proteasome and ubiquitinated proteins such as the NF- $\kappa$ B<sup>49,50</sup>. The *UBAC1* also contributes to the inflammatory signal transduction pathways and affects cell proliferation and viability in keratinocytes<sup>51</sup>. The role of the *UBAC1* in the KIRC needs to be clarified. The *PSMD9* was 3<sup>rd</sup> top KIRC-related mRNA found in our rule mining analysis. The *PSMD9*, as a part of the 26S proteasome, regulates protein degradation<sup>52</sup>. The *PSMD9* is overexpressed in tumor tissues and associated with cell proliferation, hostile

tumor outcome, and resistance to the therapy<sup>53–55</sup>. Since different tumor suppressors and oncogenes are controlled by the Ub- and proteasome-mediated degradation, the CSN7A, UBAC1, PSMD9, and RNF40 may play important roles in the pathogenesis of the KIRC. The CSN7A in the KIRC association rules has a high dependency on PSMD9, CAPNS1, and UBAC1, SNU13, respectively. Our study may open an innovative horizon to investigate the role of the CSN7A, UBAC1, PSMD9, and RNF40 in the pathogenesis of the KIRC.

microRNAs are critical managers of the development and progression of the RCC. They function as oncomirs or anti-oncomirs. miR-28, let-7i, miR-23b, miR-125a, miR-22 were top five identified miRNAs that were dysregulated in KIRC. For more details see Supplementary file, discussion part.

**Candidate RNAs in KIRP.** In the Rule mining analysis, the NMD3, ZNF41, CFAP36, FGFR1OP2, and RGL1 were identified as the top five mRNA with high association rules in the KIRP ( $\geq 2195$ ).

The transcript patterns of the ribosomal proteins are tumor- and tissue-specific<sup>56</sup> and their induced abnormal translation can provoke a malignant phenotype independent of chromatin remodeling and the deregulation of the transcriptional process<sup>57</sup>. The 60S ribosomal export protein NMD3 (NMD3) was the first identified mRNA involved in the pathogenesis of the KIRP in our study. The NMD3 is a nuclear adaptor protein that transports the 60S subunit into the cytoplasm via the nuclear pore complex and participates in the cytoplasmic maturation of 60S particles<sup>58,59</sup>. In the cytoplasm, the release of the NMD3p from 60S subunits needs a GTPase and the ribosomal protein (Rpl10p). Any mutation in these proteins leads to cytoplasmic retention of the NMD3 on pre-60S subunits, blocking ribosome assembly and biogenesis<sup>60,61</sup>. The NMD3 also exerts a significant effect on RNA biosynthesis, mainly ribosomal RNA, and consequently may impact tumorigenesis<sup>62</sup>. A direct role of the NMD3 needs to be elucidated in the KIRP.

Respectively, miR-125a, miR-23b, miR-210, miR-99b, and miR-101-2 were the top five identified miRNAs dysregulated in KIRP. The possible pathological roles of the candidate mRNAs and miRNAs identified in this study are presented in Fig. 9 and explained in the Supplementary file in more detail. Although much remains to elucidate the KIRP mechanism, the roles of the NMD3, ZNF41, CFAP36, FGFR1OP2, and RGL1 are of considerable interest. The NMD3 in KIRP has a high dependency on CAPNS1, INTS5, and CFAP36, respectively.

The molecular features presented in this study will offer new insights into the underlying mechanisms that are responsible for the initiation and progression of KIRC and KIRP. Moreover, the ability to diagnose KIRC and KIRP with a high certainty level will help pathologists accurately differentiate the most common subtypes of RCC. In this way, appropriate clinical decision-making strategies can be obtained. Furthermore, the ability of artificial intelligence for accurate differentiation of RCC may reduce unnecessary intervention rates. Our attempts to develop molecular discriminating patterns had some limitations. First of all, we did not evaluate the molecular mechanism of the candidate RNAs in the RCC models. In vitro and in vivo studies are needed to be performed to achieve this goal. Second, we did not validate the identified RNAs in clinical samples. Future studies would address this issue in large-sample-sized studies. We believe that adding other features from mutations, polymorphisms, alterations in copy number, and DNA methylation platforms would effectively tackle the discriminating problem of RCC subtypes and improve their early detection.

## Conclusion

In this paper, deep learning-driven biomarkers were presented for discriminating common subtypes of RCC. Panels of 77 mRNAs and 73 miRNAs could discriminate the KIRC, KIRP, and KICH subtypes from each other with high accuracy. The CSN7A and miR-28 along with the NMD3 and miR-125a were the most frequent itemsets in the KIRC and KIRP association rules, respectively. Due to a frequent mutation in protein-coding regions and an elevated burden of unfolded proteins; an elevated protein turnover was necessary for those speedily dividing cancer cells. Hence, the inhibition of the UPS components appeared to be a hopeful strategy for KIRC therapy. The identified mRNAs and microRNAs in this study can regulate signal transduction, cell cycle machinery, and apoptosis and all are relevant contributors to carcinogenesis and cancer progression. Therefore, they may provide further insight into the pathogenesis, diagnosis, prognosis, and molecular-targeted therapy in RCC subtypes (Fig. 9).

## Data availability

The data obtained from the artificial intelligence approaches will be available from the corresponding authors upon request.

Received: 21 February 2022; Accepted: 19 September 2022

Published online: 30 September 2022

## References

1. Bray, F. *et al.* Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries (vol 68, pg 394, 2018). *CA Cancer J. Clin.* **70**(4), 313–313 (2020).
2. Mytsyk, Y. *et al.* "Potential clinical applications of microRNAs as biomarkers for renal cell carcinoma. *Cent. Eur. J. Urol.* **71**(3), 295–303. <https://doi.org/10.5173/cej.2018.1618> (2018).
3. Howlader, N. *et al.* SEER cancer statistics review, 1975–2016, National Cancer Institute. *Bethesda, MD* **2020**, 1–10 (2019).
4. Capitanio, U. *et al.* Epidemiology of renal cell carcinoma. *Eur. Urol.* **75**(1), 74–84 (2019).
5. Vasudev, N. S., Selby, P. J. & Banks, R. E. Renal cancer biomarkers: The promise of personalized care. *BMC Med.* **10**(1), 1–10 (2012).
6. Capitanio, U. *et al.* A critical assessment of the prognostic value of clear cell, papillary and chromophobe histological subtypes in renal cell carcinoma: a population-based study. *BJU Int.* **103**(11), 1496–1500 (2009).
7. Cairns, P. Renal cell carcinoma. *Cancer Biomark.* **9**(1–6), 461–473 (2011).

8. Braga, E. A., Fridman, M. V., Loginov, V. I., Dmitriev, A. A. & Morozov, S. G. Molecular mechanisms in clear cell renal cell carcinoma: Role of miRNAs and hypermethylated miRNA genes in crucial oncogenic pathways and processes. *Front. Genet. Rev.* **10**(320). <https://doi.org/10.3389/fgene.2019.00320> (2019).
9. Négrier, S. *et al.* Axitinib in first-line for patients with metastatic papillary renal cell carcinoma: Results of the multicentre, open-label, single-arm, phase II AXIPAP trial. *Eur. J. Cancer* **129**, 107–116 (2020).
10. Pang, J. S. *et al.* The underlying molecular mechanism and potential drugs for treatment in papillary renal cell carcinoma: a study based on TCGA and Cmap datasets. *Oncol. Rep.* **41**(4), 2089–2102 (2019).
11. Ng, K. L. *et al.* Differentiation of oncocytoma from chromophobe renal cell carcinoma (RCC): can novel molecular biomarkers help solve an old problem?. *J. Clin. Pathol.* **67**(2), 97–104. <https://doi.org/10.1136/jclinpath-2013-201895> (2014).
12. Ge, Y.-Z. *et al.* MicroRNA expression profiles predict clinical phenotypes and prognosis in chromophobe renal cell carcinoma. *Sci. Rep.* **5**(1), 1–8 (2015).
13. Sim, S. *et al.* Prognostic utility of pre-operative circulating osteopontin, carbonic anhydrase IX and CRP in renal cell carcinoma. *Br. J. Cancer* **107**(7), 1131–1137 (2012).
14. Sabatino, M. *et al.* Serum vascular endothelial growth factor and fibronectin predict clinical response to high-dose interleukin-2 therapy. *J. Clin. Oncol.* **27**(16), 2645 (2009).
15. Li, G., Feng, G., Gentil-Perret, A., Genin, C. & Tostain, J. Serum carbonic anhydrase 9 level is associated with postoperative recurrence of conventional renal cell cancer. *J. Urol.* **180**(2), 510–514 (2008).
16. Zhao, H. *et al.* Gene expression profiling predicts survival in conventional renal cell carcinoma. *PLoS Med.* **3**(1), e13 (2006).
17. Morrissey, J. J. *et al.* Urine aquaporin 1 and perilipin 2 differentiate renal carcinomas from other imaged renal masses and bladder and prostate cancer. *Mayo Clin. Proc.* **90**(1), 35–42 (2015).
18. Rossi, S. H., Klatte, T., Usher-Smith, J. & Stewart, G. D. Epidemiology and screening for renal cancer. *World J. Urol.* **36**(9), 1341–1353 (2018).
19. N.-N. C. Institute. The Cancer Genome Atlas (TCGA). <https://portal.gdc.cancer.gov>
20. Bakhshandeh, S., Azmi, R. & Teshnehlab, M. Symmetric uncertainty class-feature association map for feature selection in microarray dataset. *Int. J. Mach. Learn. Cybern.* **11**(1), 15–32. <https://doi.org/10.1007/s13042-019-00932-7> (2020).
21. Tsai, C.-F. & Sung, Y.-T. Ensemble feature selection in high dimension, low sample size datasets: Parallel and serial combination approaches. *Knowl. Based Syst.* **203**, 106097 (2020).
22. Ferreira, A. J. & Figueiredo, M. A. Efficient feature selection filters for high-dimensional data. *Pattern Recogn. Lett.* **33**(13), 1794–1804 (2012).
23. Azadifar, S. & Monadjemi, S. A. Feature selection using social network techniques. In *2015 7th Conference on Information and Knowledge Technology (IKT)*, 1–6 (IEEE, 2015).
24. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**(10), P10008 (2008).
25. Boppana, R. & Halldórsson, M. M. Approximating maximum independent sets by excluding subgraphs. *BIT Numer. Math.* **32**(2), 180–196. <https://doi.org/10.1007/BF01994876> (1992).
26. Pirmoradi, S., Teshnehlab, M., Zarghami, N. & Sharifi, A. A self-organizing deep auto-encoder approach for classification of complex diseases using SNP genomics data. *Appl. Soft Comput.*, 106718. <https://doi.org/10.1016/j.asoc.2020.106718> (2020).
27. Agrawal, R. & Srikant, R. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, vol. 1215, 487–499 (Citeseer, 1994).
28. Han, J., Pei, J. & Yin, Y. Mining frequent patterns without candidate generation. *ACM SIGMOD Rec.* **29**(2), 1–12 (2000).
29. Han, J., Pei, J., Yin, Y. & Mao, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Disc.* **8**(1), 53–87 (2004).
30. Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. & Verkamo, A. I. Fast discovery of association rules. *Adv. Knowl. Discov. Data Mining* **12**(1), 307–328 (1996).
31. Kuok, C. M., Fu, A. & Wong, M. H. Mining fuzzy association rules in databases. *ACM SIGMOD Rec.* **27**(1), 41–46 (1998).
32. Wach, S. *et al.* MicroRNA profiles classify papillary renal cell carcinoma subtypes. *ACM SIGMOD Rec.* **109**(03), 714 (2013).
33. Youssef, Y. M. *et al.* Accurate molecular classification of kidney cancer subtypes using MicroRNA signature. *Eur. Urol.* **59**(5), 721–730. <https://doi.org/10.1016/j.eururo.2011.01.004> (2011).
34. Muhamed Ali, A. *et al.* A machine learning approach for the classification of kidney cancer subtypes using miRNA genome data. *Eur. Urol.* **8**(12), 2422 (2018).
35. Pirmoradi, S., Teshnehlab, M., Zarghami, N. & Sharifi, A. A self-organizing deep neuro-fuzzy system approach for classification of kidney cancer subtypes using miRNA genomics data. *Comput. Methods Programs Biomed.* **206**, 106132. <https://doi.org/10.1016/j.cmpb.2021.106132> (2021).
36. Glickman, M. H. & Ciechanover, A. The ubiquitin-proteasome proteolytic pathway: destruction for the sake of construction. *Physiol. Rev.* **82**(2), 373–428. <https://doi.org/10.1152/physrev.00027.2001> (2002).
37. Li, L. & Deng, X. W. The COP9 signalosome: An alternative lid for the 26S proteasome?. *Trends Cell Biol.* **13**(10), 507–509. <https://doi.org/10.1016/j.tcb.2003.08.002> (2003).
38. Wei, N., Serino, G. & Deng, X. W. "The COP9 signalosome: More than a protease. *Trends Biochem. Sci.* **33**(12), 592–600. <https://doi.org/10.1016/j.tics.2008.09.004> (2008).
39. Hou, J. *et al.* CSN6 controls the proliferation and metastasis of glioblastoma by CHIP-mediated degradation of EGFR. *Oncogene* **36**(8), 1134–1144. <https://doi.org/10.1038/onc.2016.280> (2017).
40. Zhao, W. *et al.* Litopenaeus vannamei Notch interacts with COP9 signalosome complex subunit 1 (CNS1) to negatively regulate the NF- $\kappa$ B pathway. *J. Proteomics* **232**, 104074. <https://doi.org/10.1016/j.jprot.2020.104074> (2021).
41. Schweitzer, K., Bozko, P. M., Dubiel, W. & Naumann, M. CSN controls NF-kappaB by deubiquitinylation of IkkappaBalpha. *EMBO J.* **26**(6), 1532–1541. <https://doi.org/10.1038/sj.emboj.7601600> (2007).
42. Mitchell, T. J. *et al.* Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx renal. *Cell* **173**(3), 611–623.e17. <https://doi.org/10.1016/j.cell.2018.02.020> (2018).
43. Pantuck, A. J., An, J., Liu, H. & Rettig, M. B. NF-kappaB-dependent plasticity of the epithelial to mesenchymal transition induced by Von Hippel-Lindau inactivation in renal cell carcinomas. *Can. Res.* **70**(2), 752–761. <https://doi.org/10.1158/0008-5472.Can-09-2211> (2010).
44. An, J., Fisher, M. & Rettig, M. B. VHL expression in renal cell carcinoma sensitizes to bortezomib (PS-341) through an NF-kappaB-dependent mechanism. *Oncogene* **24**(9), 1563–1570. <https://doi.org/10.1038/sj.onc.1208348> (2005).
45. Miyauchi, Y., Kato, M., Tokunaga, F. & Iwai, K. The COP9/signalosome increases the efficiency of von Hippel-Lindau protein ubiquitin ligase-mediated hypoxia-inducible factor-alpha ubiquitination. *J. Biol. Chem.* **283**(24), 16622–16631. <https://doi.org/10.1074/jbc.M710599200> (2008).
46. Wicker, C. A. & Izumi, T. Analysis of RNA expression of normal and cancer tissues reveals high correlation of COP9 gene expression with respiratory chain complex components. *BMC Genomics* **17**(1), 983. <https://doi.org/10.1186/s12864-016-3313-y> (2016).
47. Susak, H. *et al.* Efficient and flexible Integration of variant characteristics in rare variant association studies using integrated nested Laplace approximation. *PLoS Comput. Biol.* **17**(2), e1007784. <https://doi.org/10.1371/journal.pcbi.1007784> (2021).

48. Zheng, J., Zhang, H., Ma, R., Liu, H. & Gao, P. Long non-coding RNA KRT19P3 suppresses proliferation and metastasis through COP57A-mediated NF- $\kappa$ B pathway in gastric cancer. *Oncogene* **38**(45), 7073–7088. <https://doi.org/10.1038/s41388-019-0934-z> (2019).
49. Kravtsova-Ivantsiv, Y. *et al.* KPC1-mediated ubiquitination and proteasomal processing of NF- $\kappa$ B p105 to p50 restricts tumor growth. *Cell* **161**(2), 333–347. <https://doi.org/10.1016/j.cell.2015.03.001> (2015).
50. Kravtsova-Ivantsiv, Y. *et al.* Excess of the NF- $\kappa$ B p50 subunit generated by the ubiquitin ligase KPC1 suppresses tumors via PD-L1- and chemokines-mediated mechanisms. *Proc. Natl. Acad. Sci. USA*. **117**(47), 29823–29831. <https://doi.org/10.1073/pnas.2019604117> (2020).
51. Mazzone, P. *et al.* UBAC1/KPC2 regulates TLR3 signaling in human keratinocytes through functional interaction with the CARD14/CARMA2sh-TANK Complex. *Int. J. Mol. Sci.* **21**(24). <https://doi.org/10.3390/ijms21249365> (2020).
52. Sahu, I., Sangith, N., Ramteke, M., Gadre, R. & Venkatraman, P. A novel role for the proteasomal chaperone PSMD9 and hnRNP A1 in enhancing I $\kappa$ B $\alpha$  degradation and NF- $\kappa$ B activation: Functional relevance of predicted PDZ domain-motif interaction. *FEBS J.* **281**(11), 2688–2709. <https://doi.org/10.1111/febs.12814> (2014).
53. Köster, F. *et al.* PSMD9 expression correlates with recurrence after radiotherapy in patients with cervical cancer. *Oncol. Lett.* **20**(1), 581–588. <https://doi.org/10.3892/ol.2020.11622> (2020).
54. Banz-Jansen, C., Münchow, B., Diedrich, K. & Finas, D. Bridge-1 is expressed in human breast carcinomas: Silencing of Bridge-1 decreases Smad2, Smad3 and Smad4 expression in MCF-7 cells, a human breast cancer cell line. *Arch. Gynecol. Obstet.* **284**(6), 1543–1549. <https://doi.org/10.1007/s00404-011-1875-0> (2011).
55. Zhou, C. *et al.* Prognostic value and molecular mechanisms of proteasome 26S subunit, non-ATPase family genes for pancreatic ductal adenocarcinoma patients after pancreaticoduodenectomy. *J. Invest. Surg.*, 1–17, <https://doi.org/10.1080/08941939.2020.1863527> (2020).
56. Dolezal, J. M., Dash, A. P. & Prochownik, E. V. Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. *BMC Cancer* **18**(1), 275. <https://doi.org/10.1186/s12885-018-4178-z> (2018).
57. Gentilella, A., Kozma, S. C. & Thomas, G. A liaison between mTOR signaling, ribosome biogenesis and cancer. *Biochem. Biophys. Acta.* **1849**(7), 812–820. <https://doi.org/10.1016/j.bbagr.2015.02.005> (2015).
58. Zhou, Y., Musalgaonkar, S., Johnson, A. W. & Taylor, D. W. Tightly-orchestrated rearrangements govern catalytic center assembly of the ribosome. *Nat. Commun.* **10**(1), 958. <https://doi.org/10.1038/s41467-019-08880-0> (2019).
59. Patchett, S., Musalgaonkar, S., Malyutin, A. G. & Johnson, A. W. The T-cell leukemia related rpl10-R98S mutant traps the 60S export adapter Nmd3 in the ribosomal P site in yeast. *PLoS Genet.* **13**(7), e1006894. <https://doi.org/10.1371/journal.pgen.1006894> (2017).
60. Hedges, J., West, M. & Johnson, A. W. Release of the export adapter, Nmd3p, from the 60S ribosomal subunit requires Rpl10p and the cytoplasmic GTPase Lsg1p. *EMBO J.* **24**(3), 567–579. <https://doi.org/10.1038/sj.emboj.7600547> (2005).
61. De Keersmaecker, K. *et al.* Exome sequencing identifies mutation in CNOT3 and ribosomal genes RPL5 and RPL10 in T-cell acute lymphoblastic leukemia. *Nat. Genet.* **45**(2), 186–190. <https://doi.org/10.1038/ng.2508> (2013).
62. Bai, B., Moore, H. M. & Laiho, M. CRM1 and its ribosome export adaptor NMD3 localize to the nucleolus and affect rRNA synthesis. *Nucleus* **4**(4), 315–325. <https://doi.org/10.4161/nucl.25342> (2013).

## Acknowledgements

The authors acknowledge the Kidney Research Center for its scientific support. Also, We would like to thank the Clinical Research Development Unit of Tabriz Valiasr Hospital, Tabriz University of Medical Sciences, Tabriz, Iran for their assistance in this research.

## Author contributions

Conceptualization: S.Z.V., Data acquisition: S.M.H.K. and S.P. Data analysis: S.P. and S.M.H.K., Writing original draft: S.P., S.M.H.K., and S.Z.V., Review, and editing: all authors, Final Revision: M.R.A. and M.T.

## Funding

This work was financially supported by the Kidney Research Center, Tabriz University of Medical Sciences, Tabriz, Iran (#68374).

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-20783-7>.

**Correspondence** and requests for materials should be addressed to S.Z.V. or S.P.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022