scientific reports

Check for updates

OPEN

Key therapeutic targets implicated at the early stage of hepatocellular carcinoma identified through machine-learning approaches

Seyed Mahdi Hosseiniyan Khatibi[1,2,3,6], Farima Najjarian[4,6], Hamed Homaei Rad[3], Mohammadreza Ardalan[1], Mohammad Teshnehlab[5], Sepideh Zununi Vahed[1]✉ & Saeed Pirmoradi[2]✉

Hepatocellular carcinoma (HCC) is the most frequent type of primary liver cancer. Early-stage detection plays an essential role in making treatment decisions and identifying dominant molecular mechanisms. We utilized machine learning algorithms to find significant mRNAs and microRNAs (miRNAs) at the early and late stages of HCC. First, pre-processing approaches, including organization, nested cross-validation, cleaning, and normalization were applied. Next, the t-test/ANOVA methods and binary particle swarm optimization were used as a filter and wrapper method in the feature selection step, respectively. Then, classifiers, based on machine learning and deep learning algorithms were utilized to evaluate the discrimination power of selected features (mRNAs and miRNAs) in the classification step. Finally, the association rule mining algorithm was applied to selected features for identifying key mRNAs and miRNAs that can help decode dominant molecular mechanisms in HCC stages. The applied methods could identify key genes associated with the early (e.g., Vitronectin, thrombin-activatable fibrinolysis inhibitor, lactate dehydrogenase D (LDHD), miR-590) and late-stage (e.g., SPRY domain containing 4, regucalcin, miR-3199-1, miR-194-2, miR-4999) of HCC. This research could establish a clear picture of putative candidate genes, which could be the main actors at the early and late stages of HCC.

Hepatocellular carcinoma (HCC) is the third cause of cancer deaths worldwide[1]. The scientific observations have indicated that cirrhosis[2], heavy alcoholism[3], smoking, lifestyle, hepatitis B and C viral infection[4,5], hemochromatosis[6], and alpha-1-antitrypsin deficiency can be important HCC risk factors. Liver function, clinical expertise, availability of treatment resources, and cancer stage can affect treatment procedures. Due to the diagnosis of liver cancer at the late stages, the overall survival rate of HCC patients has not increased despite advancements in treatment[7]. HCC early-stage detection allows clinicians to use a wide range of treatments[8], playing an essential role in making treatment decisions. Moreover, the identification of dominant molecular mechanisms at the early and late stages can improve treatment strategies.

Traditionally, clinicians utilized alpha-fetoprotein (AFP) and AFP-L3 (a glycoform of AFP) as HCC biomarkers in most developing countries[7,9]; however, these biomarkers have no reliability, sufficient sensitivity, and specificity[8]. Another biomarker was Des-gamma-carboxyprothrombin (DCP), which is upregulated at late stages[10–12]. In recent years, next-generation sequencing (NGS) technology and bioinformatics methods have provided promising ways for the identification of biomarkers[13]. Many studies detected the expression of cancer-associated genes and indicated their vital role in hepatocarcinogenesis. Previous studies aimed to identify the

[1]Kidney Research Center, Tabriz University of Medical Sciences, Daneshgah Street, Tabriz 51665118, Iran. [2]Clinical Research Development Unit of Tabriz Valiasr Hospital, Tabriz University of Medical Sciences, Niyayesh Blvd., Tabriz, Iran. [3]Rahat Breath and Sleep Research Center, Tabriz University of Medical Science, Tabriz, Iran. [4]Faculty of Medicine, Tabriz University of Medical Sciences, Tabriz, Iran. [5]Department of Electric and Computer Engineering, K.N. Toosi University of Technology, Tehran, Iran. [6]These authors contributed equally: Seyed Mahdi Hosseiniyan Khatibi and Farima Najjarian. ✉email: sepide.zununi@gmail.com; said.pirmoradi@gmail.com

differentially expressed RNA transcripts, genes, or miRNAs in cancer versus normal or cancer versus other liver diseases.

Recently, Artificial Intelligence has succeeded in many applications, such as health care[14]. In this regard, a few studies proposed machine learning methods to predict the HCC stages based on the genomic profile of samples[15]. Sathipati et al. suggested a support vector machine-cancer stage prediction method and a bi-objective genetic algorithm for miRNA selection. Test accuracy and AUC (area under the receiver operating characteristic curve) of 74.28%, and 0.73 for early (stage I, II) and late (stage III, IV) stages of discrimination were reported based on miRNA data, respectively[16]. Kaur et al. investigate mRNA and methylation data to distinct early (stage I) and late (stage II, III, IV) stages. They utilized different feature selection and classification algorithms and compared the obtained result. Accuracy and AUC of 76% and 0.79 were reported, respectively[17]. In these studies, authors applied hold-out cross-validation for error estimation with an 80:20 ratio for training and test splitting. Moreover, a machine learning approach was applied to the early diagnosis of HCC and classifying patients with HCC and without HCC (CwoHCC)[18]. In another study, authors utilized machine learning and bioinformatic tools to diagnose HCC patients (HCC and non-HCC)[19]. Książek et al.[20] used a two-level feature selection method (NCA-GA-SVM) for HCC fatality prognosis prediction. Recently, Liu et al.[21] proposed a deep-learning model to predict HCC recurrence based on pathology images.

In this study, we applied machine learning algorithms to investigate significant mRNAs and miRNAs separately. First, we applied pre-processing approaches, including organization, nested cross-validation, cleaning, and normalization. Next, the t-test/ANOVA methods and binary particle swarm optimization (PSO) were used as a filter and wrapper method in the feature selection step, respectively. Then, a classifier based on machine learning and deep learning algorithms was utilized to evaluate the selected features (mRNAs and miRNAs) in the classification step. Finally, the association rule mining algorithm was applied to selected features for identifying key mRNAs and miRNAs that can help decode dominant molecular mechanisms at the early and late stages of HCC.

Results

Our primary objective was to identify the significant mRNAs/miRNAs that can classify patients at early-stage and late-stage with the best accuracy in the first phase. Decoding the molecular mechanisms of early- and late-stages and identifying their top mRNAs/miRNAs were our next objectives. In this regard, we applied four steps to mRNA/miRNA data, including preprocessing, feature selection, classification, and association rule mining as shown in Fig. 1. Moreover, we used Python and its libraries, including Numpy, Pandas, Matplotlib, Sickit-learn, Scipy, Pytorch, Pyswarms, and Mlxtend to implement the proposed algorithms.

In the feature selection step, t-test (filter method) and binary PSO (wrapper method) were used for selecting significant mRNAs and miRNAs. Finally, 77 miRNAs among 1881 miRNAs were selected, presented in Table S1. Furthermore, 123 mRNAs among 60,483 mRNAs were selected (Table S2). The binary PSO parameters, including α, β, θ, number of particles, and number of iterations, were set to 2, 2, 0.9, 35, and 100, respectively.

In the classification step, we employed seven classifiers, including SVM, KNN, NB, RF, deep Self-Organizing Auto-Encoder (SOAE), Logistic Regression, and XgBoost to evaluate the importance of selected features (mRNAs/miRNAs) based on their discrimination power between early and late stages. The average performance of each classifier was represented using accuracy, F1-score, MCC, sensitivity, and specificity for train/validation/test folds of miRNA and mRNA data in Tables 1 and 2, respectively. Moreover, we reported the performance of classifiers based on both selected mRNAs and miRNAs by concatenating chosen features (Table 3).

The performance of classifiers based on miRNA features illustrated that SVM with 70% accuracy and 0.7 AUC was the best model. SVM was also the best classifier in mRNA features with 74.7% accuracy and 0.75 AUC. In addition, concatenating mRNAs and miRNAs improved the classification performance with an accuracy of 76.9 and an AUC of 0.77. Also, all measures were calculated based on the nCV, which is the most accurate error estimation approach in the real world. In the association rule mining step, we discovered interesting relations including feature(s)-feature(s) (mRNAs/miRNAs) and feature(s)-target (early-stage/late-stage). Moreover, we selected significant mRNAs/miRNAs based on the repeat count of these features in generated rules and studied their role in the early and late stages of HCC tumors.

miRNA data association rule mining analysis. Twenty-eight top miRNAs involved in the consequence of early-stage and late-stage rules were presented based on the repeat counts in Table 4. In miRNA data, parameters of the algorithm, including lift (association rule), max-length (maximum length of frequent itemset), and min-support (frequent itemset) were set to 1.1, 4, and 0.3, respectively. Also, twenty of the top early-stage and late-stage rules were presented as the if–then form in Supplementary Tables S3 and S4, respectively.

In Supplementary Fig. S1a,b, the Spearman correlation for five top miRNAs of early-stage and late-stage was shown based on the heatmap plot, respectively. Also, the strength distribution of early-stage and late-stage association rules according to their lift, support, and confidence was shown in Supplementary Fig. S1c,d, respectively. We displayed the repeat count of 28 top miRNAs as ring bar plots in Supplementary Fig. S1e,f for early-stage and late-stage rules, respectively. Moreover, the boxplots of the top three miRNAs with a high repeat count at the early-stage and late-stage association rules were shown in Supplementary Fig. S2a,b, respectively.

In addition, we showed features-phenotype associations according to association rules in the graph network (Fig. 2). In Fig. 2a, it is obvious that early-stage phenotype, based on early-stage association rules, is highly dependent on miR-590, miR-23a, miR-4443, and miR-4764. In Fig. 2b, it is obvious that the late-stage phenotype is dependent on miR-3199-1 and miR-194.2.

The hsa-mir-590 was the most frequent itemset in early-stage association rules (1330 repeat counts). Therefore, this miRNA was investigated based on association rules in the graph network to find its relation with other miRNAs (Fig. 3). As shown in Fig. 3a, it is obvious that the most frequent miRNA (miR-590) at the early-stage

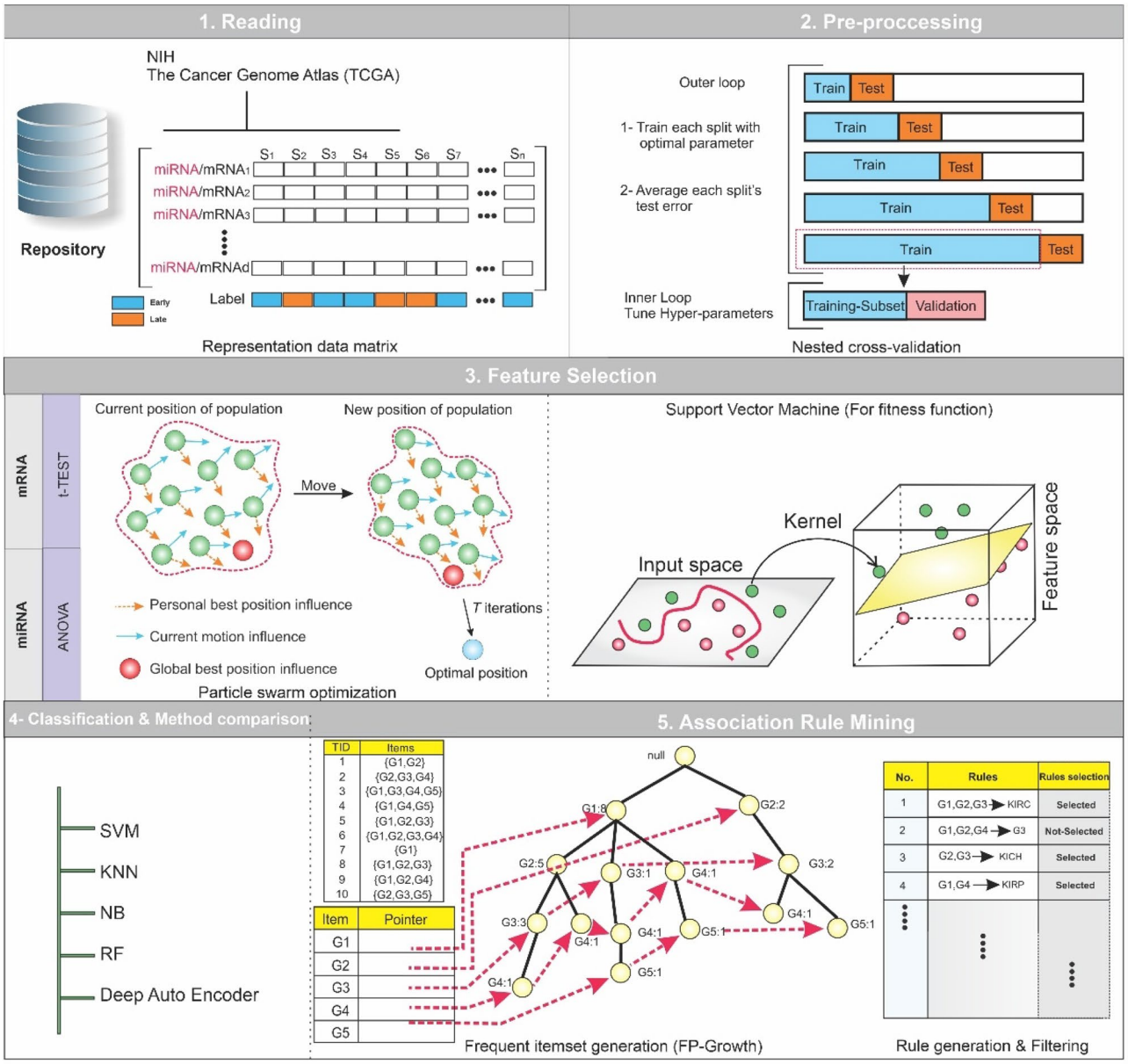

Figure 1. The overview of the proposed method. Five main steps were applied to miRNA and mRNA expression data separately, including reading, preprocessing, feature selection, classification, and association rule mining. (1) In the reading step, each dataset was downloaded from the TCGA repository. (2) The preprocessing step includes two sub-steps, nested cross-validation, and normalization. (3) The feature selection step contains two sub-steps: the filter method based on t-test for mRNA data and ANOVA for miRNA data, and the wrapper method based on binary particle swarm optimization (PSO) for both mRNA and miRNA data, in which candidate miRNAs/mRNAs with more relevance to early-stage and late-stage Hepatocellular Carcinoma (HCC) were selected. (4) multiclassifier models were utilized to evaluate the discrimination power of selected miRNAs/mRNAs. (5) The Association Rule Mining method discovered the hidden relationship between selected miRNAs/mRNAs at the early-stage and late-stage of HCC in the first level and the complex relationship among selected miRNAs/mRNAs in the second level.

association rules is associated with miR-3691, miR-21, and miR-126. The hsa-mir-3199-1 was the most frequent itemset in late-stage association rules (with 351 repeat counts) that has a high dependency on miR-21 and miR-126 (Fig. 3b).

mRNA data association rule mining analysis. Twenty-eight top mRNAs involved in the consequence of early-stage and late-stage rules were presented based on the repeat counts in Table 4. In mRNA data, parameters of the algorithm, including lift, max-length, and min-support were set to 1.1, 4, and 0.2, respectively. More-

Classifier	Folds	Accuracy	AUC-ROC	F1-score (Early stage)	F1-score (Late stage)	MCC	Sn	Sp
SVM	Train	88.8	0.88	0.89	0.88	0.78	0.81	0.95
	Validation	71.3	0.71	0.72	0.7	0.43	0.67	0.75
	Test	**70**	**0.7**	**0.7**	**0.68**	**0.4**	**0.68**	**0.73**
KNN	Train	72.2	0.72	0.75	0.67	0.47	0.56	0.87
	Validation	56.2	0.56	0.62	0.46	0.14	0.39	0.74
	Test	57.1	0.56	0.63	0.45	0.14	0.37	0.75
NB	Train	70	0.7	0.74	0.63	0.43	0.51	0.89
	Validation	66.1	0.66	0.7	0.59	0.34	0.5	0.81
	Test	66.3	0.65	0.67	0.59	0.32	0.53	0.78
RF	Train	91.4	0.91	0.91	0.9	0.83	0.85	0.97
	Validation	65	0.65	0.67	0.61	0.3	0.59	0.74
	Test	65	0.66	0.67	0.61	0.32	0.58	0.74
AE	Train	75	0.74	0.75	0.74	0.5	0.71	0.78
	Validation	66	0.66	0.67	0.64	0.33	0.62	0.7
	Test	65.1	0.65	0.66	0.62	0.3	0.6	0.7
Logistic regression	Train	82.7	0.82	0.82	0.82	0.65	0.8	0.84
	Validation	62.7	0.63	0.63	0.61	0.26	0.6	0.65
	Test	62.3	0.62	0.62	0.6	0.24	0.6	0.64
XgBoost	Train	98.5	0.98	0.98	0.98	0.97	0.97	0.99
	Validation	64.2	0.64	0.64	0.63	0.29	0.61	0.67
	Test	63	0.63	0.63	0.6	0.26	0.59	0.66

Table 1. The performance of classifiers based on 77 selected miRNAs. Significant values are in bold. AUC: The area under the curve, ROC: receiver operating characteristic curve, MCC: Matthews Correlation Coefficient, Sn: sensitivity, Sp: specificity.

Classifier	Folds	Accuracy	AUC-ROC	F1-score (Early stage)	F1-score (Late stage)	MCC	Sn	Sp
SVM	Train	93	0.93	0.93	0.93	0.86	0.95	0.91
	Validation	77.6	0.77	0.77	0.77	0.55	0.74	0.8
	Test	**74.7**	**0.75**	**0.74**	**0.74**	**0.5**	**0.73**	**0.76**
KNN	Train	70	0.7	0.6	0.76	0.46	0.94	0.46
	Validation	60	0.6	0.44	0.7	0.36	0.9	0.31
	Test	59	0.58	0.41	0.67	0.2	0.85	0.31
NB	Train	79	0.79	0.79	0.78	0.58	0.77	0.81
	Validation	70	0.7	0.72	0.68	0.42	0.64	0.77
	Test	68	0.7	0.7	0.62	0.38	0.59	0.78
RF	Train	92.3	0.92	0.92	0.92	0.84	0.89	0.95
	Validation	65.3	0.65	0.66	0.64	0.31	0.62	0.68
	Test	63.6	0.64	0.63	0.63	0.28	0.6	0.68
AE	Train	79.5	0.79	0.79	0.79	0.59	0.78	0.8
	Validation	71.4	0.71	0.71	0.71	0.43	0.7	0.72
	Test	70	0.71	0.7	0.7	0.42	0.68	0.73
Logistic regression	Train	95.2	0.95	0.95	0.95	0.9	0.95	0.94
	Validation	68.7	0.68	0.68	0.68	0.37	0.68	0.68
	Test	67.7	0.68	0.65	0.69	0.36	0.71	0.64
XgBoost	Train	92.4	0.92	0.92	0.92	0.85	0.89	0.94
	Validation	60	0.6	0.6	0.58	0.2	0.57	0.62
	Test	61.2	0.61	0.59	0.61	0.23	0.62	0.61

Table 2. The performance of classifiers based on 123 selected mRNAs. Significant values are in bold. AUC: The area under the curve, ROC: receiver operating characteristic curve, MCC: Matthews Correlation Coefficient, Sn: sensitivity, Sp: specificity.

over, twenty of the top early-stage and late-stage rules were presented as the if–then form in Supplementary Tables S5 and S6, respectively.

Classifier	Folds	Accuracy	AUC-ROC	F1-score (Early stage)	F1-score (Late stage)	MCC	Sn	Sp
SVM	Train	96.1	0.96	0.96	0.96	0.92	0.94	0.97
	Validation	75.4	0.75	0.76	0.74	0.51	0.7	0.8
	Test	**76.9**	**0.77**	**0.78**	**0.74**	**0.54**	**0.7**	**0.83**
KNN	Train	76.2	0.76	0.72	0.79	0.54	0.88	0.63
	Validation	64	0.64	0.59	0.68	0.29	0.79	0.48
	Test	63.2	0.64	0.56	0.67	0.29	0.8	0.47
NB	Train	78.7	0.76	0.8	0.72	0.56	0.59	0.93
	Validation	68	0.67	0.72	0.61	0.38	0.5	0.85
	Test	67.7	0.67	0.63	0.64	0.36	0.64	0.71
RF	Train	94	0.94	0.94	0.93	0.83	0.9	0.97
	Validation	67.8	0.68	0.7	0.64	0.36	0.6	0.76
	Test	69.5	0.69	0.71	0.66	0.4	0.63	0.75
AE	Train	82	0.81	0.82	0.81	0.64	0.8	0.84
	Validation	74	0.74	0.74	0.73	0.48	0.72	0.76
	Test	75	0.74	0.76	0.72	0.5	0.7	0.79
Logistic regression	Train	100	1	1	1	1	1	1
	Validation	71	0.7	0.7	0.7	0.41	0.71	0.7
	Test	74.7	0.74	0.75	0.73	0.49	0.75	0.74
XgBoost	Train	95	0.95	0.95	0.94	0.9	0.91	0.97
	Validation	64.6	0.65	0.65	0.62	0.3	0.6	0.69
	Test	63	0.63	0.64	0.6	0.26	0.6	0.65

Table 3. The performance of classifiers based on 200 selected mRNAs and miRNAs. Significant values are in bold. AUC: The area under the curve, ROC: receiver operating characteristic curve, MCC: Matthews Correlation Coefficient, Sn: sensitivity, Sp: specificity.

In Supplementary Fig. S3a,b, the Spearman correlation for five top mRNAs of early-stage and late-stage was shown based on the heatmap plot, respectively. The strength distribution of early-stage and late-stage association rules in line with their support, lift, and confidence was demonstrated in Supplementary Fig. S3c,d, respectively. We displayed the repeat count of 28 top mRNAs as ring bar plots in Supplementary Fig. S3e,f for early-stage and late-stage rules. Also, the boxplots of three top mRNAs with a high repeat count at the early-stage and late-stage association rules were shown in Supplementary Fig. S2c,d, respectively.

In addition, we showed features-phenotype associations based on association rules in the graph network (Fig. 4). Early-stage phenotype had a high dependency on ENSG00000109072 (Vitronectin) and ENSG00000175600 (SUGCT, succinyl-CoA:glutarate-CoA transferase), Fig. 4a. In Fig. 4b, it is obvious that late-stage phenotype, based on late-stage association rules, is highly dependent on ENSG0000055957, ENSG00000178301, ENSG00000130988, ENSG00000173269, ENSG0000080618, and ENSG00000116816.

Vitronectin was the most frequent itemset in early-stage association rules (with 1533 repeat counts). Hence, to investigate its associations with other features, its relations based on association rules were studied in the graph network (Fig. 5). In Fig. 5a, it is noticeable that Vitronectin, the most frequent mRNA at the early-stage association rules, is dependent on ENSG00000125730 (Complement C3). Furthermore, the ENSG00000176422 (SPRY domain containing 4) was the most frequent itemset with 7297 repeat counts in late-stage association rules. In Fig. 5b, it is obvious that the SPRY domain containing 4, the most frequent mRNA in the late-stage association rules, is associated with ENSG00000017248, ENSG000000137806, and ENSG000000166816. More in-depth biological functions of these findings are provided in the "Discussion" section.

Discussion and conclusion

Accurate prediction and stage classification of HCC are vital for the management of patients since the proper HCC treatment decisions are impacted by the degree of liver impairment and tumor stage. In this study, aberrantly expressed mRNA and microRNA patterns were identified by deep learning that can discriminate early stage from the late stage of cancerous HCC with high accuracy. Utilizing ARM analysis, top candidate mRNAs and microRNAs were found in early and late HCC association rules. Vitronectin, thrombin-activatable fibrinolysis inhibitor (TAFI), lactate dehydrogenase D (LDHD), and miR-590 were identified as top transcripts involved at the early stage of HCC. A SPRY domain containing 4, regucalcin, and miR-3199-1 were identified to play important roles at the late stage of HCC.

The crosstalk between cancer cells and their microenvironment is the first stage in the expansion of metastasis. In the present study, vitronectin was the first identified mRNA by association rule mining to be implicated at the early stage of HCC. Vitronectin is an adhesive multifunctional glycoprotein that links cells to the extracellular matrix (ECM) via different ligands such as urokinase plasminogen activator receptor (uPAR), plasminogen activator inhibitor-1 (PAI-1), and integrins. Vitronectin is mainly synthesized by hepatocytes[22] and plays major roles in cell growth, cell adhesion, differentiation, progression, migration, regulation of the innate immune

Early-stage rules		Late-stage rules	
miRNA ID	**Repeat count**	**miRNA ID**	**Repeat count**
hsa-mir-590	1330	hsa-mir-3199-1	351
hsa-mir-23a	827	hsa-mir-194-2	168
hsa-mir-4443	662	hsa-mir-4999	108
hsa-mir-3691	448	hsa-mir-885	85
hsa-mir-877	447	hsa-mir-151b	70
hsa-mir-331	427	hsa-mir-4654	52
hsa-mir-6515	396	hsa-mir-216b	38
hsa-mir-629	376	hsa-mir-22	37
hsa-mir-4764	355	hsa-mir-126	33
hsa-mir-7850	273	hsa-mir-3926-1	19
hsa-let-7e	256	hsa-mir-4526	18
hsa-mir-4523	238	hsa-mir-330	17
hsa-mir-1289-1	213	hsa-mir-641	17
hsa-mir-1255a	212	hsa-mir-3622a	15
hsa-mir-6888	211	hsa-mir-4673	13
hsa-mir-6801	206	hsa-mir-6845	13
hsa-mir-4752	206	hsa-mir-548v	12
hsa-mir-4487	192	hsa-mir-6728	12
hsa-mir-5706	179	hsa-mir-3155a	12
hsa-mir-95	171	hsa-mir-3936	11
hsa-mir-423	166	hsa-mir-6783	10
hsa-mir-4746	165	hsa-mir-4735	10
hsa-mir-183	158	hsa-mir-548s	9
hsa-mir-1254-2	145	hsa-mir-3680-1	9
hsa-mir-643	141	hsa-mir-3926-2	9
hsa-mir-561	133	hsa-mir-1257	9
hsa-mir-4478	127	hsa-mir-4757	9
hsa-mir-658	121	hsa-mir-124-1	9
mRNA ID	**Repeat count**	**mRNA ID**	**Repeat count**
ENSG00000109072.12	1553	ENSG00000176422.12	7297
ENSG00000080618.12	1398	ENSG00000130988.11	7086
ENSG00000166816.12	1039	ENSG00000080618.12	6969
ENSG00000137806.7	931	ENSG00000109072.12	6949
ENSG00000146416.15	254	ENSG00000166816.12	6871
ENSG00000130307.10	254	ENSG00000137806.7	6668
ENSG00000255987.1	250	ENSG00000055957.9	6661
ENSG00000245954.5	250	ENSG00000036473.6	6372
ENSG00000245164.5	250	ENSG00000167711.12	6355
ENSG00000236213.1	250	ENSG00000125730.15	6042
ENSG00000233387.1	250	ENSG00000146416.15	5915
ENSG00000211751.6	250	ENSG00000161944.15	5723
ENSG00000211749.1	250	ENSG00000121410.10	5706
ENSG00000246084.2	250	ENSG00000188338.13	5588
ENSG00000163815.5	250	ENSG00000163631.15	5531
ENSG00000237702.2	250	ENSG00000244414.5	5256
ENSG00000124203.5	247	ENSG00000147647.11	5243
ENSG00000113263.11	247	ENSG00000139597.15	5194
ENSG00000264468.1	247	ENSG00000167701.12	5162
ENSG00000010319.5	237	ENSG00000134240.10	5075
ENSG00000178343.4	233	ENSG00000185305.9	4933
ENSG00000273328.4	229	ENSG00000172482.4	4890
ENSG00000264419.1	225	ENSG00000178301.3	4460
ENSG00000197921.5	219	ENSG00000213995.10	4328
ENSG00000270412.1	216	ENSG00000157379.12	4248
ENSG00000174990.4	213	ENSG00000154734.13	4230
Continued			

mRNA ID	Repeat count	mRNA ID	Repeat count
ENSG00000231690.2	213	ENSG00000173269.12	4134
ENSG00000272789.1	204	ENSG00000170989.8	3992

Table 4. Top miRNAs and mRNAs based on repeat count in early-stage and late-stage rules.

system, complement activation, and angiogenesis under different biological and pathological circumstances[23]. Moreover, vitronectin participates in other biological processes such as controlling tissue remodeling, wound healing, and coagulation pathway (fibrinolysis and thrombosis). Some tumor cells have been reported to secrete vitronectin[24,25] to promote ECM degradation and cell migration.

The role of vitronectin in the pathogenesis of HCC has been reported previously. Cytokines and/or growth factors can stimulate the synthesis and secretion of vitronectin in hepatocarcinoma cells[26] and promote the adhesion and migration of cancer cells[27]. Within the liver tumor microenvironment, expressed vitronectin can support the recruitment and preservation of effector lymphocytes by a uPAR-mechanism[24]. uPAR is an anchored receptor, interacting with uPA and some molecules, such as vitronectin and integrins. Evidence indicates that in different cancers, the uPAR-uPA system (by activating plasminogen and fibrinolysis) is linked with tumor progression, peritoneal dissemination, and metastasis[28]. Abnormal levels of uPAR might induce EMT by vitronectin binding and easing tumor invasion and metastasis[29]. An increased serum level of vitronectin represents high diagnostic and prognostic values for HCC[30] since it is associated with clinicopathological factors and early recurrence[31], cell migration[27], and the malignant growth of the tumor. Vitronectin when freed from the cancer cells complex guarded by fibrinogen, functions as a pro-migratory factor for directing metastasis of cancer cells to low-fibrinogen body cavities or lymphatics in a uPAR-dependent manne[32]. Suppression of vitronectin can inhibit HCC in vitro and in vivo[33], therefore, it can potentially be considered a therapeutic target for the treatment of HCC.

Cases with advanced HCC have irregular fibrinolysis and coagulation that is associated with tumor progression where cancer-associated thrombosis is an important cause of mortality. Venous thromboembolism, mainly portal vein tumor thrombus, is a challenging and common complication in the HCC that can be the earliest sign of an underlying malignancy[34,35]; it indicates a worse prognosis and less tolerance to treatment. In this regard, thrombin-activatable fibrinolysis inhibitor (TAFI) was identified as the second top transcript at an early stage of HCC by our analysis. TAFI, also called carboxypeptidase B2, is a plasma glycoprotein that is activated by plasmin or thrombin during the coagulation cascade. It acts as a molecular link between fibrinolysis and coagulation and can also regulate the interaction between inflammation and coagulation[36]. The binding of thrombin to thrombomodulin, a regulator of hemostasis that plays an anti-metastatic role in cancer, is essential for TAFI activation. An elevated level of TAFI is associated with several types of cancer and a more advanced cancer stage[37–39], signifying that TAFI can play a role in the pathogenesis of thrombosis in cancer. Beyond activation of systemic coagulation, TAFI secretion from cancer cells elevates the intra-tumoral deposition of fibrin, promoting the growth and dissemination of tumor cells[40]. It is proposed that the production of TAFI can be mediated by directly malignant cells or indirectly by liver/endothelial cells that are induced by cancer-induced inflammatory cytokines. Modulation of TAFI may hinder migration and invasion of cancer cells[41,42]; therefore, TAFI can be another valuable molecular target for the treatment of HCC.

Lactate dehydrogenase D (LDHD) was the 3rd identified mRNA at the early stage of HCC in this study. It is responsible for the mitochondrial metabolism of D-lactate (a less common form of lactate) in humans[43] and is supposed to produce by cancer cells[44]. The LDHD preferentially uses NADPH as a coenzyme that differs from the coenzyme that is used by other LDHs (A–C). In cancer cells, the mitochondrial LDHD metabolism is more active than in normal cells[45] and its elevated level was detected in clear cell renal cell carcinoma[46], prostate cancer[47], and uterine sarcoma[48]. The methylglyoxal (MG) pathway produces an end-product, LDHD, to eliminate the toxic glycolysis-derived MG[49], fatty acid synthesis, and scavenge reactive oxygen species, all of which are vital for cancer cell proliferation and viability[47]. Based on the available studies, D- lactate metabolism can represent a target for the development of an anticancer therapeutic strategy in the HCC.

NADH dehydrogenase 1 alpha subcomplex assembly factor 1 (*NDUFAF1*), the 4th identified mRNA, is a chaperone protein in mitochondria that are implicated in the assembly of the NADH[50]. Its downregulation is connected with the recurrence of HCC[51].microRNAs (miRs) are non-coding, small RNAs that regulate gene expression negatively. Their abnormal expression, as oncogenes or tumor suppressors, is involved in the initiation, development, and metastasis of HCC. Evidence suggests that certain subsets of miRs can be therapeutic targets for HCC. In our association rule mining analysis, top miRs including miR-590, miR-23a, miR-4443, miR-3691, and miR-877 were identified to be involved at the early stage of the HCC, the roles of which have been reported previously. miR-590 plays a tumor suppressor role in HCC by targeting a variety of transcripts such as transcriptional enhancer activator domain 1 (TEAD1)[52], Wnt pathway[53], TGF-beta RII[54], and ROCK2[55]. A bioinformatics analysis in HCC cell lines indicated that SOX2, CX3CL1, E-cadherin, N-cadherin, and FOXA2 are the potential downstream target genes of miR-590-3p in HCC[56]. This microRNA can be a potential target molecule for the treatment of HCC.

This work has some limitations. We did not validate the results on other cancer genomic datasets including, gene expression omnibus (GEO). It is suggested to validate the results in other datasets in future works. Further bioinformatics analysis is needed to be performed to find the targets of the identified microRNAs and to understand their correlations with the identified mRNAs.

Identification of therapeutic targets is essential for the effective development of drugs for HCC. In this study, we applied an AI-based framework to highlight putative mRNA and microRNA targets for HCC. The applied methods could identify key genes associated with the early (e.g., Vitronectin, TAFI, LDH-D, miR-590) and late-stage (e.g., SPRY domain containing 4, regucalcin, miR-3199-1, miR-194-2, miR-4999) of HCC. Applying

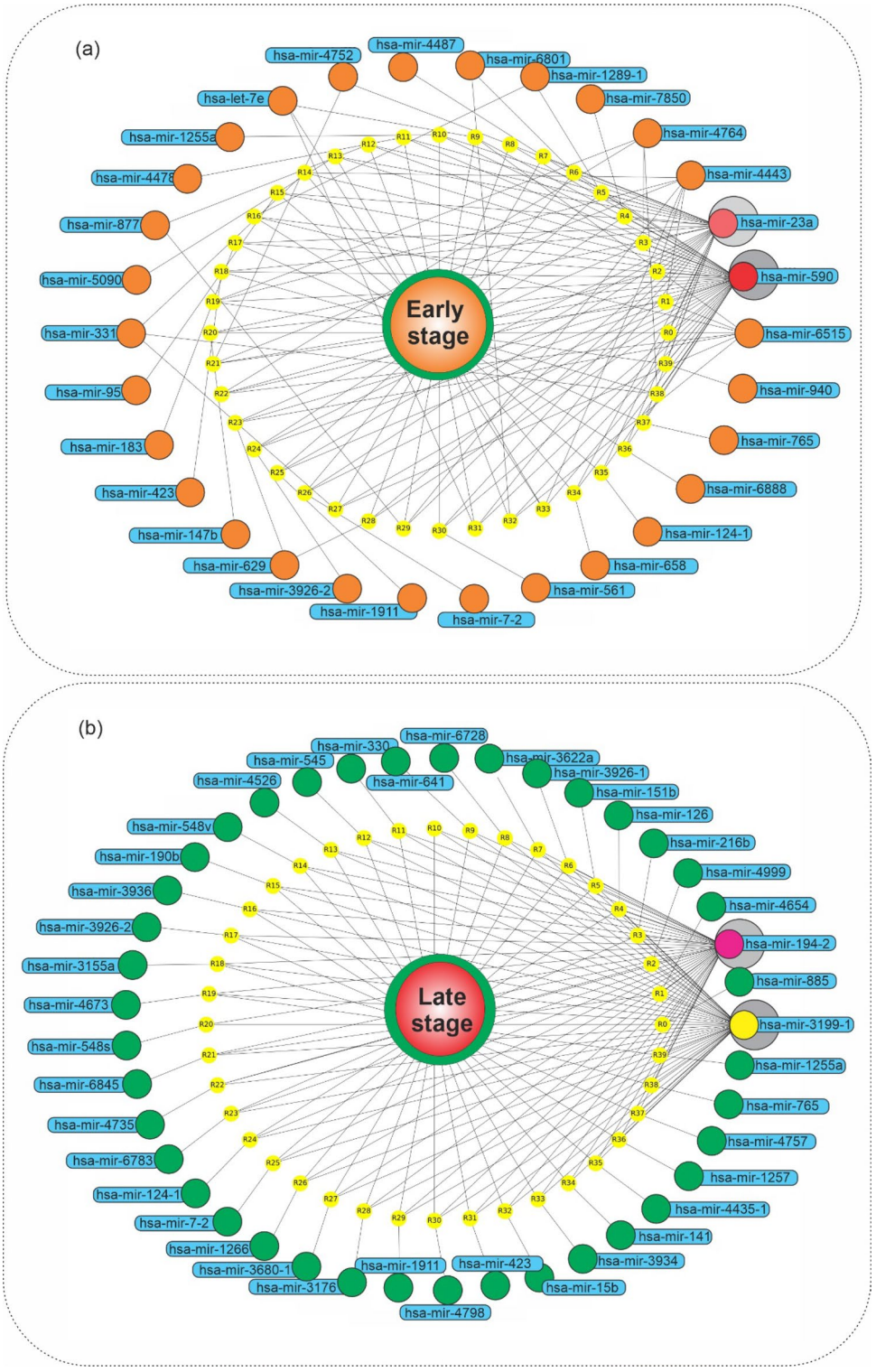

Figure 2. Graph network of miRNAs at early- and late-stage of HCC. Graph network of (**a**) early-stage related association rules (with lift > 1.16) and (**b**) late-stage related association rules (with lift > 1.2), in which the early-stage phenotype, its rules, and related miRNAs were presented, by orange, yellow, and blue colors, respectively. Python programming language (version 3.9) and Matplotlib library (version 3.6.0) were used to draw the plots, all of them are open sources.

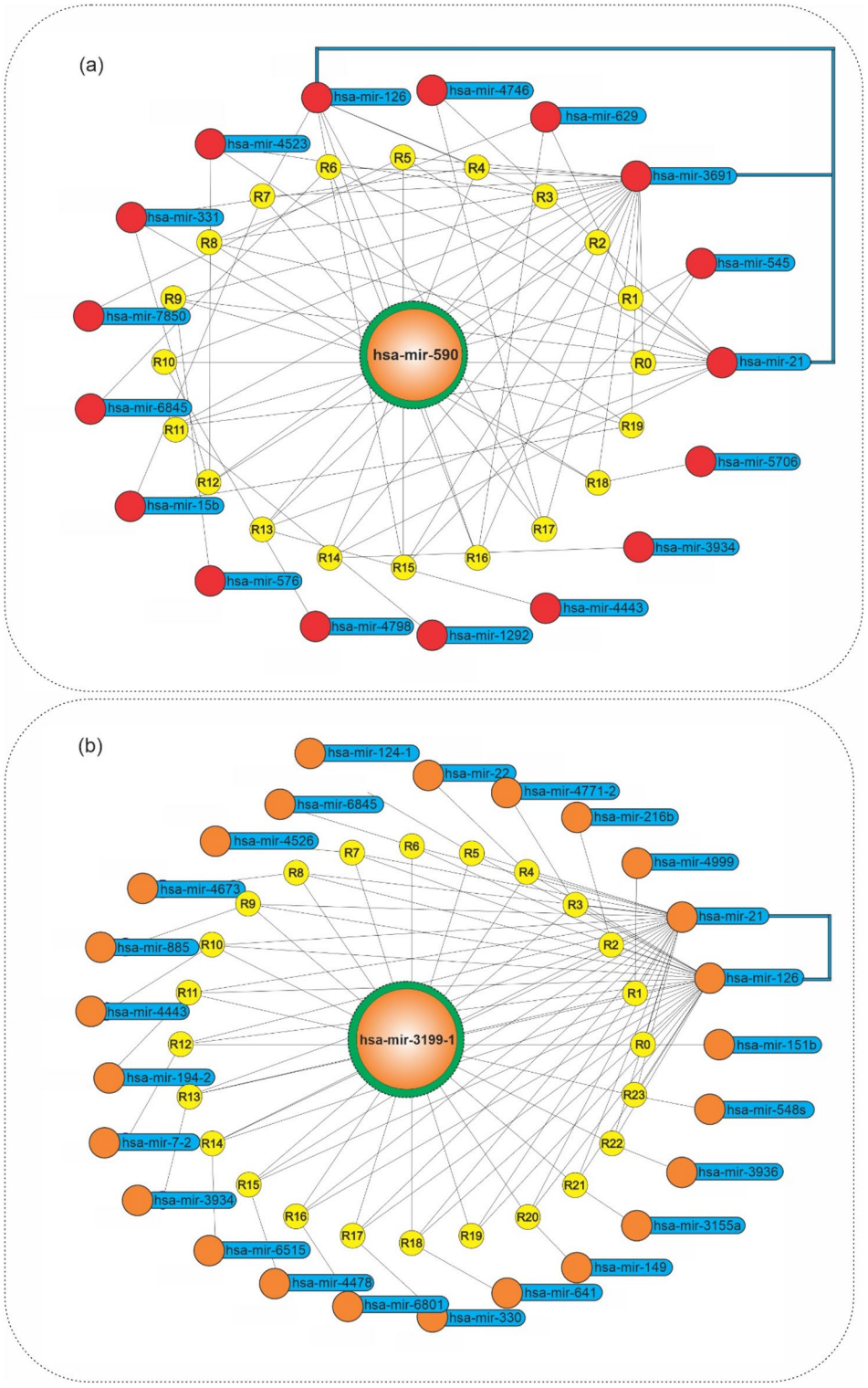

Figure 3. Graph network of has-mir-590 and has-mir-3199-1 in HCC. Graph network of (**a**) has-mir-590 (with lift > 1.14) and (**b**) has-mir-3199-1 (with lift > 1.126) related association rules, in which the has-mir-590 and has-mir-3199-1, their rules, and their related miRNAs were presented, by orange, yellow, and blue colors, respectively. Python programming language (version 3.9) and Matplotlib library (version 3.6.0) were used to draw the plots, all of them are open sources.

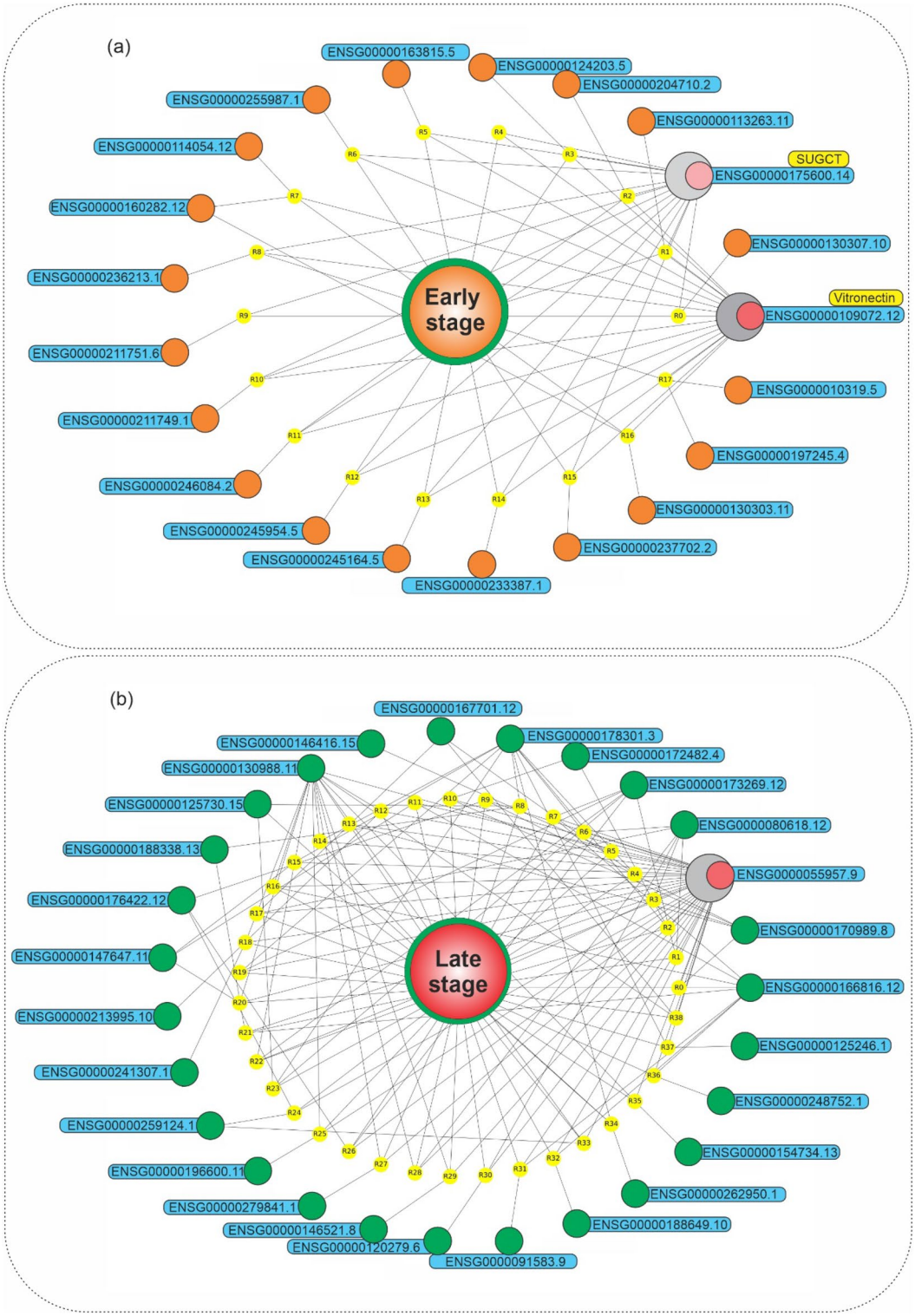

Figure 4. Graph network of mRNAs at early- and late-stages of HCC. Graph network of (**a**) early-stage related association rules (with lift > 1.21) and (**b**) late-stage related association rules (with lift > 1.38), in which the early-stage phenotype, its rules, and related mRNAs were presented, by orange, yellow, and blue colors, respectively. Python programming language (version 3.9) and Matplotlib library (version 3.6.0) were used to draw the plots, all of them are open sources.

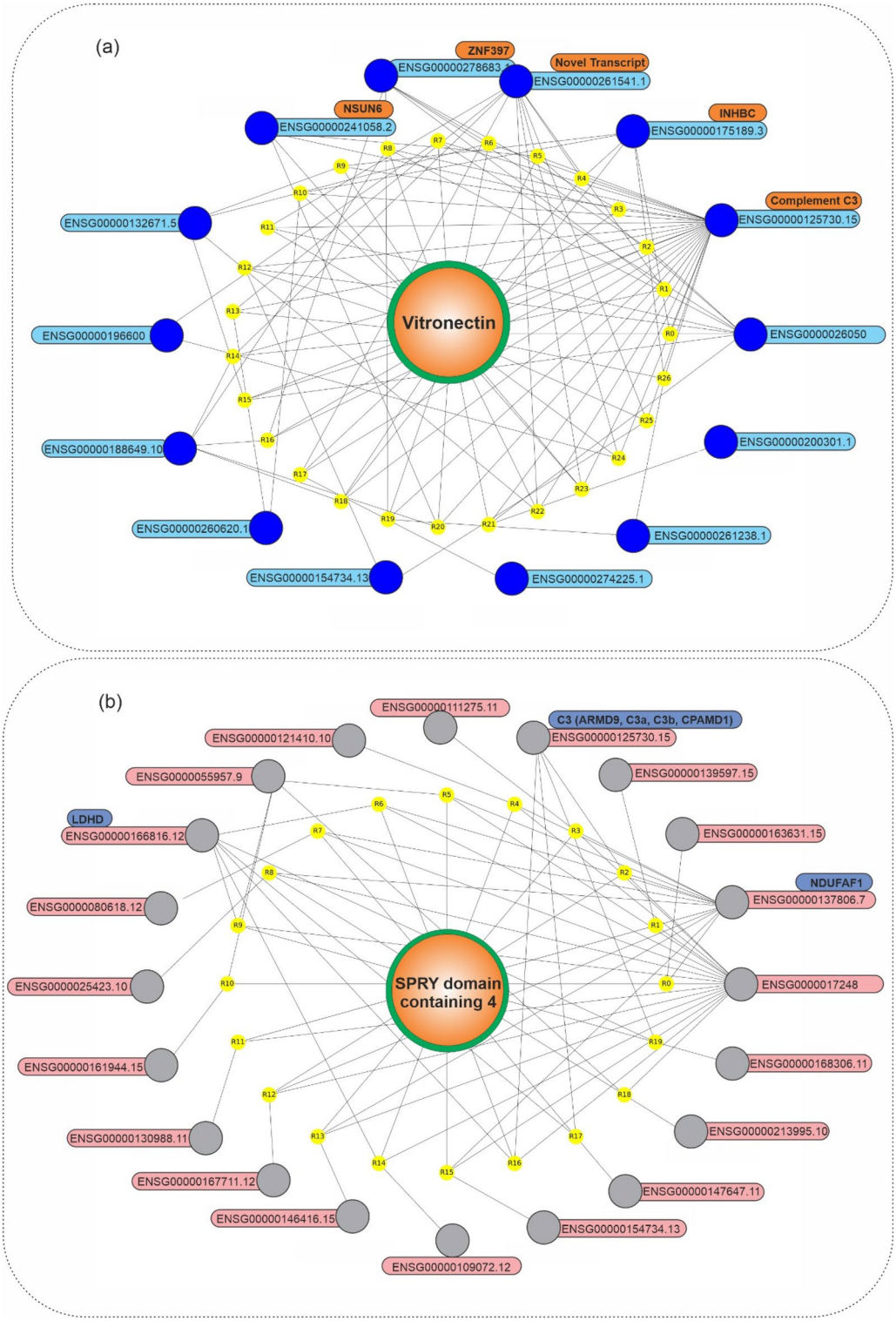

Figure 5. Graph network of Vitronectin and SPRY domain containing 4 in HCC. (**a**) Graph network of Vitronectin (with lift > 1.42) related association rules, in which the Vitronectin, its rules, and related mRNAs were presented, by orange, yellow, and blue colors, respectively. Vitronectin, the most frequent mRNA in the early-stage association rules, has a high dependency on ENSG00000125730 (Complement C3). (**b**) Graph network of the SPRY domain containing 4 (with lift > 1.46) related association rules, in which the SPRY domain containing 4, its rules, and related mRNAs were presented, by orange, yellow, and blue colors, respectively. Python programming language (version 3.9) and Matplotlib library (version 3.6.0) were used to draw the plots, all of them are open sources.

targeted molecular therapy at an early stage and proper time will improve the outcome of patients with HCC and lessen their mortality rate. This research could establish a likely clear picture of putative candidate genes which could be the main actors at the early and late-stage of HCC.

Methods

Dataset. We obtained the mRNA and miRNA profiles of HCC samples from The Cancer Genome Atlas (TCGA) database; which is accessible in the GDC data portal (https://portal.gdc.cancer.gov/). Furthermore, clinical data were downloaded to extract the sample's HCC stage based on its Biospecimen Core Resource ID. The mRNA expression was reported in terms of FPKM values for 60,483 RNA transcripts. In the miRNA profile, 1881 miRNA expression values were recorded using the Illumina HiSeq 2000 platform. The HCC stage system was defined based on the TNM system; (T) the size of the primary tumor, (M) the distant metastasis, and (N) the spread of cancer to lymph nodes. In this study, we considered stage I as an early-stage class and stages II, III, and IV as a late-stage class. The information on mRNA/miRNA data is displayed in Table 5 in more detail.

Method

The whole process was displayed in Fig. 1. The proposed algorithms were applied to mRNA and miRNA data separately.

In the pre-processing step, first, we organized the mRNA and miRNA data into a matrix form with 381/382 rows and 60,483/1881 columns for mRNA/miRNA data that present the number of samples and features, respectively. Next, the nested cross-validation (CV) technique was applied to data for accurate error estimation in the real world. In the nested CV, the number of folds in the outer and inner loops were considered 10 and 5, respectively. Then, features with the same value in all samples of training folds of the inner loops were removed. Finally, z-score and min–max methods were applied for normalization in feature selection/classification and association rule mining steps, respectively. In the z-score, we mapped the distribution of features into the normal distribution, and in the min–max, we scaled features into the [0 1] range.

In the feature selection step, filter and wrapper methods were applied. The filter methods reduce the number of features (mRNAs/miRNAs) by removing the irrelevant attributes and decreasing computational cost and time for the wrapper step. These methods evaluate features individually in the selection procedure and are classifier-independent. In the filter method, T-test and ANOVA were used for mRNA and miRNA data based on their performance in feature selection, respectively. We applied filter methods to training folds of each inner loop, and we selected 25 top features based on their p-values. Next, this process was repeated ten (tenfold in outer-loop) products five (fivefold in inner-loop) times. Finally, 261 mRNAs and 150 miRNAs were selected based on the union of selected features obtained from training folds of inner-outer loops.

The wrapper methods considered the interaction of features and due to using the classifier in the selection procedure, they are classifier-dependent. We applied binary particles swarm optimization (PSO) for feature selection as a wrapper method. In binary PSO, the support vector machine (SVM) was utilized for the fitness function evaluation. Due to being robust and reliable, we defined the fitness function based on AUC, shown in Eq. (1). Besides, the fitness function value, including mean and standard deviation of AUC, was calculated based on inner validation folds of each outer fold. Ultimately, 123 significant mRNAs and 77 miRNAs were selected based on binary PSO output.

$$fitness\ value = \big(1 - mean(AUC\ in\ all\ validation\ folds\big) + mean\big(standard\ deviation\ of\ AUC\ in\ all\ validation\ folds\big) \tag{1}$$

In the classification step, we evaluated the discrimination power of selected features (mRNAs/miRNAs) by classifying early-late stages groups. The performance of the classifier represents how much-selected features are significant. SVM, random forest (RF), K-nearest neighbor (KNN), Naive Bayes (NB), Deep self-organizing auto-encoder (SOAE), logistic regression (LR), and XgBoost were used for the classification task. Furthermore, the performance of classifiers was reported using accuracy, AUC, MCC, F_1-score, sensitivity, and specificity.

In the next step, significant relationships were discovered by association rule mining. We extracted association rules concerning selected mRNAs/miRNAs and early/late HCC stage groups. In this regard, the early/late stages group was added as a new feature to mRNA/miRNA data, and features (mRNA/miRNA) were categorized into three parts, namely low, medium, and high expression levels. Next, the FP-Growth algorithm was utilized to generate association rules in two phases, including frequent itemsets and rules generation. Next, early and late stages association rules were obtained based on the consequent part of rules with early and late-stage values, respectively. Then, we studied the antecedent part of early-stage association rules and reported mRNAs/miRNAs based on their repeat count in early-stage association rules. Finally, we studied five top miRNAs/mRNAs of early-stage and late-stage, based on the repeat count, more in-depth from a biological point of view.

The concept of the above algorithms was explained comprehensively in the following sections. In addition, the output of each step was reported and displayed in the "Result" section.

mRNA data		miRNA data	
Early-stage	Late-stage	Early-stage	Late-stage
189	192	190	192

Table 5. Information of mRNA and miRNA data.

Nested cross-validation. Feature selection and classification are the main actors in the machine learning and data mining areas. The quality of the classifier is dependent on the quality of selected features. The classifier performance is calculated based on testing data, which is not used for training and validating the model. Combining too many irrelevant features may lead to low generalization in testing data and high variance error estimation (overfitting) in the training process. In contrast, a lack of significant features may lead to high bias error estimation (underfitting). In this regard, the precise error estimation method plays a crucial role in classification and feature selection procedures.

Cross-validation (CV) is a fundamental action for the classifier accuracy/error estimation in a given dataset by splitting data into training and testing sets. Various versions of CV have been implemented to apply in feature selection and classifier parameters tuning, including leave-one-out CV, repeated double CV, and nested CV. The nested CV is a reliable way for classifier accuracy/error estimation[57]. The data is split into k outer folds in nCV and the remaining k-1 folds were merged and split into inner folds for inner training and validation. Training outer folds, including inner training folds and validation folds, are used for feature selection and model parameter tuning. Finally, general classifier accuracy/error is estimated based on testing outer folds.

z-score and min–max normalization. Z-score and min–max normalization are implemented by Eqs. (2) and (3), respectively. In Eq. (2), μ and σ are the mean and standard deviation values of x, and in Eq. (3), min and max are the minima and maximum values of x (feature). The Z-score process alters data distribution and converts it to normal. While the min–max method does not change data distribution and only scales data to the [0 1] range.

$$y = \frac{x - \mu}{\sigma} \tag{2}$$

$$y = \frac{1}{Max - Min}(x - Min) \tag{3}$$

T-test and ANOVA. The t-test and ANOVA are a type of statistical tests employed to compare the mean of two groups. They are parametric statistical hypotheses, which are widely used in medical data. In parametric methods, there are some assumptions about the distribution of probability variables and parameters of the distribution. Conditions of normality, equal variance, and independence of samples are the principal assumptions in the t-test.

In the t-test method[58], the t-statistic, based on Eq. (4), is calculated.

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{4}$$

In Eq. (4), $\overline{x}_i$, S_i^2, n_i, and μ_i is the sample mean, sample variance, the number of samples, and ith population mean, respectively. Equation (4) converts to Eq. (5) by considering $\mu_1 - \mu_2 = 0$ based on the null hypothesis.

$$t = \frac{(\overline{x}_1 - \overline{x}_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \tag{5}$$

In the ANOVA method[59], F-statistic, based on the following steps, is calculated and compared with a threshold to find important features.

Step 1 Calculating the variation between groups (Eqs. 6 and 7):

$$Between\ sum\ of\ squares\ (BSS) = \sum_{i=1}^{C} n_i(\overline{x}_i - \overline{x})^2 \tag{6}$$

$$Between\ mean\ squares\ (BMS) = \frac{BSS}{df_B} \tag{7}$$

where n_i, $\overline{x}_i$, and $\overline{x}$ are the number of samples and mean of samples in ith group, and mean of all samples, respectively. Also, $df_B = K - 1$ is the degree of freedom.

Step 2 Calculating the variation within groups (Eqs. 8 and 9):

$$Within\ sum\ of\ squares\ (WSS) = \sum_{i=1}^{C} (n_i - 1)\sigma_i^2 \tag{8}$$

$$Within\ mean\ squares\ (WMS) = \frac{WSS}{df_w} \tag{9}$$

σ is the standard deviation, and $df_w = (N - K)$ where N and K a are the number of total samples and groups, respectively.

Step 3 Calculating F-test statistic (Eq. 10):

$$F = \frac{BMS}{WMS} \tag{10}$$

The F-statistic demonstrates features of discriminative capabilities and its higher values mean that the variation among means of groups is less likely to happen by chance.

Particle swarm optimization (PSO). Many algorithms utilize swarm intelligence to solve optimization problems, such as PSO, ant colony optimization (ACO), etc. PSO is a widely used algorithm for optimization among swarm intelligence-based algorithms[60]. It is well-known as an easy and flexible method from the implementation point of view. This algorithm uses a mathematical simulated model based on swarm behavior such as bird flocking in nature.

The PSO discovers the objective function space for finding optimum points by updating the position and tuning the velocity of individual agents, called particles. Updating of the particle position is implemented by Eq. (11). Adjusting of movement is defined by Eq. (12), composed of three components based on its own best location (X_i^*), global best location (g^*), and previous velocity (v_i^t). In Eqs. (11) and (12), X_i^t and v_i^t are the position and velocity of ith particle in t time, respectively.

$$X_i^{t+1} = X_i^t + v_i^{t+1} \tag{11}$$

$$v_i^{t+1} = \theta v_i^t + \alpha\epsilon_1\left[g^* - X_i^t\right] + \beta\epsilon_2\left[X_i^* - X_i^t\right] \tag{12}$$

In Eq. (12), ϵ_1 and ϵ_2 are two random vectors, which the values of their elements are between 0 and 1. α and β are user-defined parameters, which can typically be $\alpha \approx 2$ *and* $\beta \approx 2$. θ is the value between 0 and 1, which in the simplest case is defined by $\theta \approx 0.5 \sim 0.9$. The pseudo-code of PSO is illustrated in Table 6 in more detail.

In the standard PSO, position and velocity are based on continuous values. However, many real-world optimization problems search space are defined based on discrete values, such as binary problems. In this regard, Kennedy and Eberhart presented a new version of standard PSO for discrete optimization in 1997[62]. They applied sigmoid and uniform transformations to the velocity vector and position vector, as shown in Eqs. (13) and (14).

$$S\left(v_i^k\right) = \frac{1}{1 + \exp\left(-v_i^k\right)} \quad k = 1, 2, \ldots, d \tag{13}$$

$$x_i^k = \begin{cases} 1 & if\ r < S\left(v_i^k\right) \\ 0 & otherwise \end{cases} \tag{14}$$

where r is a random variable in the [0 1] range. The value of each velocity element v_i^k is defined as the probability of taking one value by x_i^k. The binary PSO (BPSO) significantly varies from the standard continuous PSO.

Classification. Classifier models are powerful tools that apply well-known machine-learning algorithms for classification tasks. In this study, we used SVM, NB, KNN, RF, LR, and XgBoost as classic classifiers, and Deep self-organizing auto-encoder (SOAE)[63] to assess the discrimination power of selected features.

The predictive performance of the classifier was evaluated using the following evaluation metrics (Eqs. 15–19), including accuracy, F_1-score, Matthews correlation coefficient (MCC), sensitivity (Sn), and specificity (Sp). False negative (FN), false positive (FP), true negative (TN), and true positive (TP).

Particle Swarm Optimization Algorithm
Objective function $f(X)$, $X = (x_1, x_2, ..., x_d)^T$
Initialize locations x_i and velocity v_i of n particles
Find g^* from min $(f(X_1), f(X_2), ..., f(X_n))$ at t=0
While (criterion):
For loop over all n particles and all d dimensions:
Generate new velocity v_i^{t+1} using Eq. 10.
Calculate new locations $X_i^{t+1} = X_i^t + v_i^{t+1}$
Evaluate objective functions at new locations X_i^{t+1}
Find the current best for each particle X_i^*
Endfor
Find the current global best g^*
Update $t = t + 1$ (pseudo time or iteration counter)
Endwhile
Output the final results g^*

Table 6. Pseudo code of particle swarm optimization[61].

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{15}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{16}$$

$$F_1 = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{17}$$

$$sensitivity = \frac{TP}{TP + FN} \times 100 \tag{18}$$

$$specificity = \frac{TN}{TN + FP} \times 100 \tag{19}$$

Association rule mining. Association rule mining is a potent data mining tool that presents the hidden association in the form of rules by discovering associated frequently co-occurring items in the dataset. Market basket analysis[64–66] and bioinformatics[67] are two main areas that apply association rule mining for the extraction of significant associations in marketing and genomic data, respectively. The interpretation of gene expression data (mRNA/miRNA), annotations, detection of protein interaction, and biomolecular localization prediction are some applications of association rule mining in bioinformatics[67].

Association rule mining has two main steps, frequent itemset mining (FIM) and association rule generation. FIM extracts frequently co-occur sets of items (i.e., frequent itemsets). If itemset support is more than the minimum support threshold, itemset is called a frequent itemset. Next, the association rule generation step creates the rules from the discovered frequent itemsets (FIs). If the support/confidence/lift of the rule is no less than the minimum support/confidence/lift threshold, the generated rule is called the association rule. These thresholds are user-defined parameters.

Association rule mining is an NP-hard problem, in which finding the results is challenging in a reasonable time. Introducing the Apriori algorithm addressed the computational problem in most regular-sized data[68]. Since then, many types of research have been done to develop new algorithms such as FP-Growth[69] and Eclat[70]. These algorithms improved the scalability of the Apriori algorithm. However, the computational cost of association rule mining in the FIM stage for high-dimension data and big data is a challenging subject.

There are some principal terms in association rule mining, which are mentioned in the above section. In the following, we describe and formalize these basic concepts of frequent itemset and association rules. The related theories are available in[71] with more details. Let $I = \{i_1, i_2, \ldots, i_d, y\}$ is a set of items, $D = \{d_1, d_2, \ldots, d_n\}$ is a dataset of n instances, $F = \{f_1, f_2, \ldots, f_m\}$ is the features space with m features, and $Y = \{0, 1\}$ is the user-defined phenotype. The d_i can be presented as a tuple (X_i, y_i), where $X_i \in f_1 \times f_2 \times \cdots \times f_m$ and $y_i \in Y$.

Definition 1. (Length of an itemset)
Let X be an itemset, which has K-distinct items, the length of the X is defined as $|X| = K$.

Definition 2. (Support count and support of an itemset)
The total number of samples including X itemset is defined support count of an itemset X. Also, support of an item set X is the ratio of support count to the total number of samples.

Definition 3. (Frequent itemset)
An itemset X is called a frequent itemset if and only if its support is no less than the minimum support, which is the user-defined threshold.

Definition 4. (Association rule)
An association rule is defined as a form of $A \rightarrow C$, where A and C are itemsets and $A \cup C = \varphi, A \subset X, C \subset X$. In the $A \rightarrow C$, A and C are called the Antecedent and Consequent, respectively. Also, $A \rightarrow C$ displays the association that if all items in Antecedent occur, then all items in Consequent co-occur. The generated association rules are filtered out based on the user-defined threshold, such as support, confidence, and lift.

Definition 5. (Support of rule)
The support of rule $A \rightarrow C$ is the percentage of samples in D (as shown in Eq. 20). This measure presents the usefulness of the rule.

$$Support(A \rightarrow C) = \frac{support(A \cup C)}{n} \tag{20}$$

Definition 6. (Confidence of rule)
The confidence of rule $A \rightarrow C$ is the percentage value that displays how frequently C occurs among all the examples containing A (as shown in Eq. 21). This measure shows the certitude of the rule.

$$Confidence(A \rightarrow C) = P(C|A) = \frac{support(A \cup C)}{support(A)} \tag{21}$$

Definition 7. (Lift of rule)
The Lift of rule $A \rightarrow C$ defines that the occurrence of itemsets A is dependent to the C. When the Lift value is less (more) than 1, the occurrence of A is negatively (positively) associated with the occurrence of C. A and C are independent when the Lift value is equal to 1. The Lift value is shown in Eq. (22).

$$Lift(A \rightarrow C) = \frac{P(A \cup C)}{P(A)P(C)} \tag{22}$$

Here the FP-Growth approach was applied for association rule mining. Also, the pseudo-codes of the two stages are available in Tables 7 and 8.

Algorithm: Frequent itemset generation in FP-Growth algorithm

Input: A database DB, represented by FP-tree constructed, and a minimum support threshold ξ.

output: The complete set of frequent patterns

Method: FP-growth (Tree, α)

(1) **if** Tree contains a single prefix path, **then**: // Mining single prefix-path FP-tree
(2) let P be the single prefix-path part of Tree;
(3) let Q be the multipath part with the top branching node replaced by a null root;
(4) **for** each combination (denoted as β) of the nodes in the path P **do**:
(5) generate pattern $\beta \cup \alpha$ with support = minimum support of nodes in β;
(6) let freq_pattern_set(P) be the set of patterns so generated;
(7) **end for**
(8) **else** let Q be Tree:
(9) **for** each item ai in Q **do**: // Mining multipath FP-tree
(10) generate pattern $\beta = a_i \cup \alpha$ with support = a_i.support;
(11) construct β's conditional pattern-base and then β's conditional FP-tree Treeβ;
(12) **if** Treeβ = ∅ **then**:
(13) call FP-growth (Treeβ, β);
(14) let freq_pattern_set(Q) be the set of patterns so generated;
(15) **end if**
(16) **end for**
(17) **end if**
(18) return (freq_pattern_set(P) ∪ freq_pattern_set(Q) ∪ (freq_pattern_set(P) × freq_pattern_set(Q)))

Table 7. Pseudocode of frequent itemset generation step in FP-Growth algorithm[69].

Algorithm: Rules generation in FP-Growth algorithm

Method:

(1) **For** each frequent itemset k-itemset F_k, k≥2 **do:**

(2) $H_1 = \{i | i \in F_k\}$; // 1-item consequents of the rule.

(3) call rules_generator (F_k, H_1);

(4) **End for**

Procedure rules_generator (F_k, H_m)

(1) $k = |F_k|$; // size of frequent itemset

(2) $m = |H_m|$; // size of rule consequent

(3) **if** k>m+1 **then:**

(4) $H_m = m + 1 - item\ consequents\ generated\ from\ H_m$;

(5) **For** each $h_{m+1} \in H_{m+1}$ **do:**

(6) $Confidance = \frac{Support(F_k)}{Support(F_k - H_{m+1})}$;

(7) **if** Confidence ≥ min_confidence **then:**

(8) output: the rule ($F_k - h_{m+1}$) → h_{m+1};

(9) **Else:**

(10) delete h_{m+1} from H_{m+1};

(11) **End if**

(12) **End for**

(13) call ap-genrules(F_k, H_{m+1})

(14) **End if**

Table 8. Pseudocode of rules generation step in FP-Growth algorithm[72].

Ethical approval. This study was approved by the Ethics Committee of Tabriz University of Medical Sciences, Tabriz, Iran (Ethical code: IR.TBZMED.VCR.REC.1400.291).

Data availability

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

Received: 8 November 2022; Accepted: 28 February 2023
Published online: 07 March 2023

References

1. El-Serag, H. B. Hepatocellular carcinoma. *N. Engl. J. Med.* **365**, 1118–1127. https://doi.org/10.1056/NEJMra1001683 (2011).
2. Forner, A. & Bruix, J. Hepatocellular carcinoma—Authors' reply. *The Lancet* **380**, 470–471 (2012).
3. Lin, C.-W. *et al.* Heavy alcohol consumption increases the incidence of hepatocellular carcinoma in hepatitis B virus-related cirrhosis. *J. Hepatol.* **58**, 730–735 (2013).
4. De Martel, C. *et al.* Global burden of cancers attributable to infections in 2008: A review and synthetic analysis. *Lancet Oncol.* **13**, 607–615 (2012).
5. Beasley, R. P. Hepatitis B virus. The major etiology of hepatocellular carcinoma. *Cancer* **61**, 1942–1956 (1988).

6. Crownover, B. K. & Covey, C. J. Hereditary hemochromatosis. *Am. Fam. Phys.* **87**, 183–190 (2013).
7. Blum, H. E. Treatment of hepatocellular carcinoma. *Best Pract. Res. Clin. Gastroenterol.* **19**, 129–145. https://doi.org/10.1016/j.bpg.2004.11.008 (2005).
8. Marrero, J. A. Current treatment approaches in HCC. *Clin. Adv. Hematol. Oncol.* **11**, 15–18 (2013).
9. Chen, C.-H. *et al.* Long-term trends and geographic variations in the survival of patients with hepatocellular carcinoma: Analysis of 11 312 patients in Taiwan. *J. Gastroenterol. Hepatol.* **21**, 1561–1566. https://doi.org/10.1111/j.1440-1746.2006.04425.x (2006).
10. Yanaihara, N. *et al.* Unique microRNA molecular profiles in lung cancer diagnosis and prognosis. *Cancer Cell* **9**, 189–198 (2006).
11. Mohamed, A. A. *et al.* MicroRNAs and clinical implications in hepatocellular carcinoma. *World J. Hepatol.* **9**, 1001 (2017).
12. Volinia, S. *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc. Natl. Acad. Sci.* **103**, 2257. https://doi.org/10.1073/pnas.0510565103 (2006).
13. Iorio, M. V. *et al.* MicroRNA signatures in human ovarian cancer. *Can. Res.* **67**, 8699. https://doi.org/10.1158/0008-5472.CAN-07-1936 (2007).
14. Huang, W., Zhang, Y. & Wan, S. A sorting fuzzy min–max model in an embedded system for atrial fibrillation detection. *ACM Trans. Multim. Comput. Commun. Appl. (TOMM)* **18**, 1–18 (2022).
15. Ahn, J. C., Qureshi, T. A., Singal, A. G., Li, D. & Yang, J. D. Deep learning in hepatocellular carcinoma: Current status and future perspectives. *World J. Hepatol.* **13**, 2039–2051. https://doi.org/10.4254/wjh.v13.i12.2039 (2021).
16. Yerukala Sathipati, S. & Ho, S.-Y. Novel miRNA signature for predicting the stage of hepatocellular carcinoma. *Sci. Rep.* **10**, 14452. https://doi.org/10.1038/s41598-020-71324-z (2020).
17. Kaur, H., Bhalla, S. & Raghava, G. P. S. Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. *PLoS ONE* **14**, e0221476. https://doi.org/10.1371/journal.pone.0221476 (2019).
18. Zhang, Z.-M. *et al.* Early diagnosis of hepatocellular carcinoma using machine learning method. *Front. Bioeng. Biotechnol.* **8**, 254 (2020).
19. Cheng, B., Zhou, P. & Chen, Y. Machine-learning algorithms based on personalized pathways for a novel predictive model for the diagnosis of hepatocellular carcinoma. *BMC Bioinform.* **23**, 248. https://doi.org/10.1186/s12859-022-04805-9 (2022).
20. Książek, W., Turza, F. & Pławiak, P. NCA-GA-SVM: A new two-level feature selection method based on neighborhood component analysis and genetic algorithm in hepatocellular carcinoma fatality prognosis. *Int. J. Numer. Methods Biomed. Eng.* **38**, e3599. https://doi.org/10.1002/cnm.3599 (2022).
21. Liu, Z. *et al.* Deep learning for prediction of hepatocellular carcinoma recurrence after resection or liver transplantation: A discovery and validation study. *Hepatol. Int.* **16**, 577–589. https://doi.org/10.1007/s12072-022-10321-y (2022).
22. Seiffert, D., Geisterfer, M., Gauldie, J., Young, E. & Podor, T. J. IL-6 stimulates vitronectin gene expression in vivo. *J. Immunol. (Baltimore, Md.:1950)* **155**, 3180–3185 (1995).
23. Preissner, K. T. The role of vitronectin as multifunctional regulator in the hemostatic and immune systems. *Blut* **59**, 419–431. https://doi.org/10.1007/bf00349063 (1989).
24. Edwards, S., Lalor, P. F., Tuncer, C. & Adams, D. H. Vitronectin in human hepatic tumours contributes to the recruitment of lymphocytes in an alpha v beta3-independent manner. *Br. J. Cancer* **95**, 1545–1554. https://doi.org/10.1038/sj.bjc.6603467 (2006).
25. Yasumitsu, H. *et al.* Vitronectin secretion by hepatic and non-hepatic human cancer cells. *In Vitro Cell. Dev. Biol. Anim.* **29a**, 403–407. https://doi.org/10.1007/bf02633989 (1993).
26. Koli, K., Lohi, J., Hautanen, A. & Keski-Oja, J. Enhancement of vitronectin expression in human HepG2 hepatoma cells by transforming growth factor-beta 1. *Eur. J. Biochem.* **199**, 337–345. https://doi.org/10.1111/j.1432-1033.1991.tb16129.x (1991).
27. Nejjari, M. *et al.* Expression, regulation, and function of alpha V integrins in hepatocellular carcinoma: An in vivo and in vitro study. *Hepatology (Baltimore, MD)* **36**, 418–426. https://doi.org/10.1053/jhep.2002.34611 (2002).
28. Bifulco, K. *et al.* Urokinase receptor promotes ovarian cancer cell dissemination through its 84–95 sequence. *Oncotarget* **5**, 4154–4169. https://doi.org/10.18632/oncotarget.1930 (2014).
29. Madsen, C. D. & Sidenius, N. The interaction between urokinase receptor and vitronectin in cell adhesion and signalling. *Eur. J. Cell Biol.* **87**, 617–629. https://doi.org/10.1016/j.ejcb.2008.02.003 (2008).
30. Mohamed, S. Y., Esmaiel, A. E., Shabana, M. A. & Ibrahim, N. F. J. G. I. Assessment of plasma vitronectin as diagnostic and prognostic marker of hepatocellular carcinoma in patients with hepatitis C virus cirrhosis. *Gastroenterol. Insights* **13**, 9–19 (2022).
31. Yang, X. P. *et al.* Diagnostic and prognostic roles of serum vitronectin in hepatitis B-related hepatocellular carcinoma. *Cancer Biomark.: Sect. A Dis. Mark.* **17**, 271–279. https://doi.org/10.3233/cbm-160639 (2016).
32. Schneider, G. *et al.* Evidence that vitronectin is a potent migration-enhancing factor for cancer cells chaperoned by fibrinogen: A novel view of the metastasis of cancer cells to low-fibrinogen lymphatics and body cavities. *Oncotarget* **7**, 69829–69843. https://doi.org/10.18632/oncotarget.12003 (2016).
33. Zhu, W. *et al.* Vitronectin silencing inhibits hepatocellular carcinoma in vitro and in vivo. *Future Oncol. (London, England)* **11**, 251–258. https://doi.org/10.2217/fon.14.202 (2015).
34. Zanetto, A. *et al.* Cancer-associated thrombosis in cirrhotic patients with hepatocellular carcinoma. *Cancers* **10**, 450. https://doi.org/10.3390/cancers10110450 (2018).
35. Zanetto, A. *et al.* More pronounced hypercoagulable state and hypofibrinolysis in patients with cirrhosis with versus without HCC. *Hepatol. Commun.* **5**, 1987–2000. https://doi.org/10.1002/hep4.1781 (2021).
36. Lin, J. H. *et al.* Identification of human thrombin-activatable fibrinolysis inhibitor in vascular and inflammatory cells. *Thromb. Haemost.* **105**, 999–1009. https://doi.org/10.1160/th10-06-0413 (2011).
37. Balcik, O. S. *et al.* Serum thrombin activatable fibrinolysis inhibitor levels in patients with newly diagnosed multiple myeloma. *Blood Coagul. Fibrinolysis: Int. J. Haemost. Thromb.* **22**, 260–263. https://doi.org/10.1097/MBC.0b013e3283442cf9 (2011).
38. Fawzy, M. S., Mohammed, E. A., Ahmed, A. S. & Fakhr-Eldeen, A. Thrombin-activatable fibrinolysis inhibitor Thr325Ile polymorphism and plasma level in breast cancer: A pilot study. *Meta Gene* **4**, 73–84. https://doi.org/10.1016/j.mgene.2015.03.004 (2015).
39. Hataji, O. *et al.* Increased circulating levels of thrombin-activatable fibrinolysis inhibitor in lung cancer patients. *Am. J. Hematol.* **76**, 214–219. https://doi.org/10.1002/ajh.20079 (2004).
40. Fawzy, M. S. & Toraih, E. A. Data supporting the structural and functional characterization of Thrombin-Activatable Fibrinolysis Inhibitor in breast cancer. *Data Brief* **5**, 981–989. https://doi.org/10.1016/j.dib.2015.10.043 (2015).
41. Yu, C., Luan, Y., Wang, Z., Zhao, J. & Xu, C. Suppression of TAFI by siRNA inhibits invasion and migration of breast cancer cells. *Mol. Med. Rep.* **16**, 3469–3474. https://doi.org/10.3892/mmr.2017.7031 (2017).
42. Bazzi, Z. A. *et al.* Activated thrombin-activatable fibrinolysis inhibitor (TAFIa) attenuates breast cancer cell metastatic behaviors through inhibition of plasminogen activation and extracellular proteolysis. *BMC Cancer* **16**, 328. https://doi.org/10.1186/s12885-016-2359-1 (2016).
43. Monroe, G. R. *et al.* Identification of human D lactate dehydrogenase deficiency. *Nat. Commun.* **10**, 1477. https://doi.org/10.1038/s41467-019-09458-6 (2019).
44. Santel, T. *et al.* Curcumin inhibits glyoxalase 1: A possible link to its anti-inflammatory and anti-tumor activity. *PLoS ONE* **3**, e3508. https://doi.org/10.1371/journal.pone.0003508 (2008).
45. Meng, H. *et al.* Engineering a d-lactate dehydrogenase that can super-efficiently utilize NADPH and NADH as cofactors. *Sci. Rep.* **6**, 24887. https://doi.org/10.1038/srep24887 (2016).
46. Wang, Y., Li, G., Wan, F., Dai, B. & Ye, D. Prognostic value of D-lactate dehydrogenase in patients with clear cell renal cell carcinoma. *Oncol. Lett.* **16**, 866–874. https://doi.org/10.3892/ol.2018.8782 (2018).

47. de Bari, L., Moro, L. & Passarella, S. Prostate cancer cells metabolize d-lactate inside mitochondria via a D-lactate dehydrogenase which is more active and highly expressed than in normal cells. *FEBS Lett.* **587**, 467–473. https://doi.org/10.1016/j.febslet.2013.01.011 (2013).
48. Song, K. J. *et al.* Expression and prognostic value of lactate dehydrogenase-A and -D subunits in human uterine myoma and uterine sarcoma. *Medicine* **97**, e0268. https://doi.org/10.1097/md.0000000000010268 (2018).
49. Rulli, A. *et al.* Expression of glyoxalase I and II in normal and breast cancer tissues. *Breast Cancer Res. Treat.* **66**, 67–72. https://doi.org/10.1023/a:1010632919129 (2001).
50. Vogel, R. O. *et al.* Human mitochondrial complex I assembly is mediated by NDUFAF1. *FEBS J.* **272**, 5317–5326. https://doi.org/10.1111/j.1742-4658.2005.04928.x (2005).
51. Yang, J. D. *et al.* Genes associated with recurrence of hepatocellular carcinoma: Integrated analysis by gene expression and methylation profiling. *J. Korean Med. Sci.* **26**, 1428–1438. https://doi.org/10.3346/jkms.2011.26.11.1428 (2011).
52. Ge, X. & Gong, L. MiR-590-3p suppresses hepatocellular carcinoma growth by targeting TEAD1. *Tumour Biol.: J. Int. Soc. Oncodev. Biol. Med.* **39**, 1010428317695947. https://doi.org/10.1177/1010428317695947 (2017).
53. Shan, X. *et al.* MiR-590-5P inhibits growth of HepG2 cells via decrease of S100A10 expression and inhibition of the Wnt pathway. *Int. J. Mol. Sci.* **14**, 8556–8569. https://doi.org/10.3390/ijms14048556 (2013).
54. Jiang, X. *et al.* MicroRNA-590-5p regulates proliferation and invasion in human hepatocellular carcinoma cells by targeting TGF-β RII. *Mol. Cells* **33**, 545–551. https://doi.org/10.1007/s10059-012-2267-4 (2012).
55. You, L. N. *et al.* Exosomal LINC00161 promotes angiogenesis and metastasis via regulating miR-590-3p/ROCK axis in hepatocellular carcinoma. *Cancer Gene Ther.* **28**, 719–736. https://doi.org/10.1038/s41417-020-00269-2 (2021).
56. Elfar, M. & Amleh, A. miR-590-3p and its downstream target genes in HCC cell lines. *Anal. Cell. Pathol. (Amst.)* **3234812**, 2019. https://doi.org/10.1155/2019/3234812 (2019).
57. Parvandeh, S., Yeh, H.-W., Paulus, M. P. & McKinney, B. A. Consensus features nested cross-validation. *Bioinformatics* **36**, 3093–3098. https://doi.org/10.1093/bioinformatics/btaa046 (2020).
58. Kim, T. K. T test as a parametric statistic. *Korean J. Anesthesiol.* **68**, 540–546. https://doi.org/10.4097/kjae.2015.68.6.540 (2015).
59. Kim, T. K. Understanding one-way ANOVA using conceptual figures. *Korean J. Anesthesiol.* **70**, 22–26. https://doi.org/10.4097/kjae.2017.70.1.22 (2017).
60. Kennedy, J. & Eberhart, R. in *Proceedings of ICNN'95-international conference on neural networks.* 1942–1948 (IEEE).
61. Yang, X.-S. in *Nature-Inspired Optimization Algorithms (Second Edition)* (ed Yang, X.-S.) 111–121 (Academic Press, 2021).
62. Kennedy, J. & Eberhart, R. C. in *1997 IEEE International conference on systems, man, and cybernetics. Computational cybernetics and simulation*. 4104–4108 (IEEE).
63. 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.* https://doi.org/10.1016/j.asoc.2020.106718 (2020).
64. Kaur, M. & Kang, S. Market basket analysis: Identify the changing trends of market data using association rule mining. *Procedia Comput. Sci.* **85**, 78–85 (2016).
65. Aghayousefi, R. *et al.* A diagnostic miRNA panel to detect recurrence of ovarian cancer through artificial intelligence approaches. *J. Cancer Res. Clin. Oncol.* https://doi.org/10.1007/s00432-022-04468-2 (2022).
66. Hosseiniyan Khatibi, S. M., Ardalan, M., Teshnehlab, M., Vahed, S. Z. & Pirmoradi, S. Panels of mRNAs and miRNAs for decoding molecular mechanisms of Renal Cell Carcinoma (RCC) subtypes utilizing Artificial Intelligence approaches. *Sci. Rep.* **12**, 16393. https://doi.org/10.1038/s41598-022-20783-7 (2022).
67. Naulaerts, S. *et al.* A primer to frequent itemset mining for bioinformatics. *Brief. Bioinform.* **16**, 216–231 (2015).
68. Agrawal, R. & Srikant, R. in *Proc. 20th int. conf. very large data bases, VLDB.* 487–499 (Citeseer).
69. Han, J., Pei, J., Yin, Y. & Mao, R. Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Min. Knowl. Disc.* **8**, 53–87 (2004).
70. Zaki, M. J. Scalable algorithms for association mining. *IEEE Trans. Knowl. Data Eng.* **12**, 372–390 (2000).
71. Li, H. & Sheu, P. C. Y. A scalable association rule learning heuristic for large datasets. *J. Big Data* **8**, 86. https://doi.org/10.1186/s40537-021-00473-3 (2021).
72. Xie, J., Wu, J. & Qian, Q. in *2009 Eighth IEEE/ACIS International Conference on Computer and Information Science.* 357–362 (IEEE).

Acknowledgements

This work was financially supported by the Kidney Research Center, Tabriz University of Medical Sciences, Tabriz, Iran (#68424). Also, the authors would like to thank the Clinical Research Development Unit of Tabriz Valiasr Hospital for their assistance in this research.

Author contributions

Conception and design study: S.ZV., SMHK, and S.P.; Data analysis: S.P., SM.HK, S.Z.V., H.HR; Writing original draft: S.P., SM.H.K., S.ZV, F.N.; Review and editing: all authors, Final Revision: MR.A. and M.T.

Funding

This research was funded by Tabriz University of Medical Sciences (Grant No: 68424).

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/s41598-023-30720-x.

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.

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) 2023