


# گزارش پروژه درس یادگیری ماشین

عنوان و نویسندگان مقاله منبع

## Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach

Oneeb Rehman <sup>1</sup>, Hanqi Zhuang <sup>1,\*</sup>, Ali Muhamed Ali <sup>1</sup> , Ali Ibrahim <sup>1</sup> and Zhongwei Li <sup>2</sup>

سورن سلاجقه – زهرا توکل همدانی

پاییز و زمستان ۹۹

## تعریف مسئله

ریز آران‌ای‌ها (microRNAs) مولکول‌های کوچک non-coding RNA هستند، که براساس جفت شدن جزئی با توالی مکمل خود در mRNA، منجر به بریده شدن و تجزیه mRNA هدف خود شده و به این صورت در رگولاسیون ژن‌های سلول نقش دارند. تاکنون آزمایش‌های فراوانی مشخص کرده‌اند که تغییر در پروفایل بیان miRNA می‌تواند با ایجاد سرطان‌های مختلف در ارتباط باشد. این مشاهدات باعث افزایش توجه نسبت به این مولکول‌ها شده و مطالعات زیادی با هدف شناسایی miRNA‌های مختلف به عنوان بیومارکر سرطان در حال انجام می‌باشد.

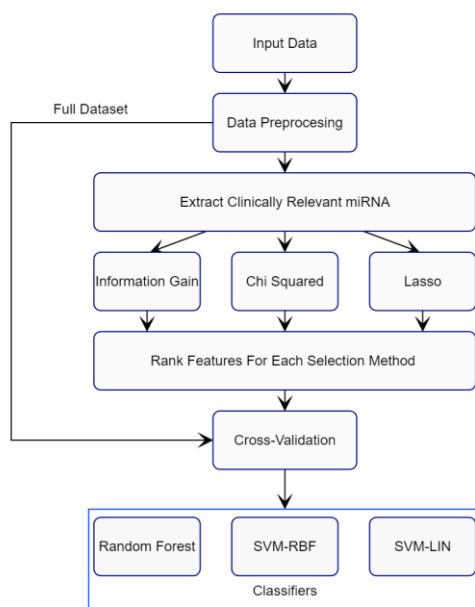
اما شناسایی تعدادی miRNA مشخص که به طور شفافی در شناسایی یک سرطان مهم باشند با چالش‌های فراوانی همراه است. زیرا ممکن است بیان بعضی از miRNA‌ها در یک سرطان خاص افزایش یافته و بیان آن‌ها در نوع دیگری از سرطان کاهش یابد. علاوه بر این بعضی از miRNA‌ها در ایجاد سرطان نقشی حیاتی داشته و برخی دیگر از اهمیت کمتری برخوردارند. بنابراین شناسایی miRNA‌های مرتبط با سرطان context-sensitive بوده و به محل و نوع سرطان وابسته است. با توجه به این موضوعات استفاده از روش‌های آنالیز کامپیوتری روی دیتاست‌های بزرگ miRNA و سرطان می‌تواند در شناسایی miRNA‌های بیومارکر سرطان مفید واقع شود.

تاکنون مطالعات مختلفی با استفاده از روش‌های یادگیری ماشین برای شناسایی و تشخیص سرطان با استفاده از miRNA به عنوان بیومارکر انجام شده است. همچنین با ایجاد پورتال داده Genomic Data Commons (GDC) توسط National Cancer Institute (NCI) حجم داده‌های در دسترس بیان miRNA به شکل چشم‌گیری افزایش یافته و در نتیجه شرایط برای انجام آزمایش‌های دقیق‌تر با استفاده از روش‌های یادگیری ماشین بهبود یافته است. با توجه به وجود داشتن داده‌های مربوط به ۳۴ نوع مختلف سرطان در این پورتال، امکان انجام مدلسازی‌های متنوعی وجود دارد. در این مطالعه هدف استفاده از روش‌های مختلف feature selection برای شناسایی گروهی از miRNA‌ها می‌باشد که در شناسایی سرطان پستان نقش حیاتی دارند. دلیل انتخاب سرطان پستان در این مطالعه این بوده است که حجم بزرگی از دیتای miRNA موجود در پورتال GDC مربوط به این سرطان می‌باشد. دلیل استفاده از روش‌های feature selection، تعداد زیاد ویژگی‌های داده پروفایل بیانی miRNA و امکان over-fit شدن مدل، و همچنین اهمیت رسیدن به دسته‌ای از miRNA‌ها به عنوان بیومارکر می‌باشد. علاوه بر این از نظر کلینیکی این موضوع حائز اهمیت است که ویژگی‌هایی در دست داشته باشیم که بتوان با اندازه‌گیری مستقیم آن‌ها در مورد وضعیت بیمار نتیجه‌گیری کرد. از این نظر روش‌های کاهش بعد feature selection به این دلیل که در شکل ویژگی‌های ورودی تغییری ایجاد نمی‌کنند دارای مطلوبیت بیشتری می‌باشند.

## مراحل و روش‌های به کار رفته در مقاله

در تصویر زیر مراحل کلی این مطالعه قابل مشاهده است. ابتدا داده‌های مورد نیاز از پورتال GDC دریافت شده، و پس از preprocessing و تمیز کردن داده‌ها، تنها داده‌های مربوط به miRNA‌هایی که در wet lab به عنوان بیومارکرهای احتمالی

برای سرطان پستان شناسایی شده‌اند، به عنوان feature ها نگه داشته شده‌اند. (به این موارد بیومارکرهای clinically verified گفته می‌شود).



در ادامه از سه روش feature selection یعنی، Chi Squared، LASSO و برای رتبه‌بندی miRNA ها استفاده شده است. سپس تعدادی از miRNA هایی که با استفاده از هر یک از سه روش feature selection مورد بررسی، بهترین رتبه‌ها را کسب کرده بودند در دسته‌هایی با اندازه‌های مختلف (استفاده از تمامی miRNA ها، ۱۰ ویژگی برتر، ۵ ویژگی برتر، یا ۳ ویژگی برتر به دست آمده از هر یک از سه روش انتخاب ویژگی) در نظر گرفته شده، و از آن‌ها به عنوان ویژگی برای دسته‌بندی با سه روش Random Forest و SVM خطی و SVM با کرنل RBF و ارزیابی براساس 10-fold cross validation استفاده شده است. (یعنی برای هر یک از ۳ روش دسته‌بندی به ازای ۳×۳ حالت در نظر گرفتن دسته‌های با اندازه‌های ۱۰ و ۵ و ۳ ویژگی برتر انتخاب شده توسط ۳ روش انتخاب ویژگی، به علاوه ۱ حالت در نظر گرفتن تمامی ویژگی‌ها، آزمایش (در مجموع ۳۰ آزمایش) انجام شده است).

معیارهای گزارش شده به عنوان نتیجه هر آزمایش accuracy، sensitivity، specificity و AUC محاسبه شده بوده است.

نحوه محاسبه accuracy، sensitivity و specificity در ادامه به همین ترتیب قابل مشاهده است.

$$\frac{TN}{TN + FP} \quad \frac{TP}{TP + FN} \quad \frac{TP + TN}{TP + TN + FP + FN}$$

نکته: افراد سالم negative و افراد سرطانی positive در نظر گرفته شده‌اند.

با توجه به imbalance شدیدی که بین کلاس‌ها وجود دارد (۱۰۴ مورد از ۱۲۰۷ مورد سالم بوده‌اند). در همه آزمایش‌های انجام شده معیار accuracy مقدار مناسبی به دست آورده است. در این دسته‌بندی چالش اصلی مربوط به دسته‌بندی درست کلاس افراد سالم است. بنابراین معیاری که در میان آزمایش‌ها نسبت به ویژگی‌های انتخاب شده حساسیت بیشتری نشان می‌دهد specificity

است. در ادامه برای بررسی این موضوع که آیا می‌توان تنها براساس تعداد کمی از miRNA دسته‌بندی مناسبی ارائه داد، لیست ده ویژگی برتر انتخاب شده توسط هر یک از ۳ روش انتخاب ویژگی در کنار هم قرار داده شده‌اند. لیست به دست‌آمده از دو روش Information gain و Chi2 تنها نسبت به یکدیگر یک جابه‌جایی داشته‌اند. بنابراین در ادامه بررسی‌های بیشتر بر مبنای این ۱۰ ویژگی انجام شده است. ۸ زیرمجموعه ۳ تایی از مجموعه ۱۰ ویژگی برتر به دست آمده با استفاده از این دو روش، به شکل زیر تعریف شده است:

زیرمجموعه اول: ویژگی‌های برتر ۱ تا ۳

زیرمجموعه دوم: ویژگی‌های برتر ۲ تا ۴

....

زیرمجموعه هشتم: ویژگی‌های برتر ۸ تا ۱۰

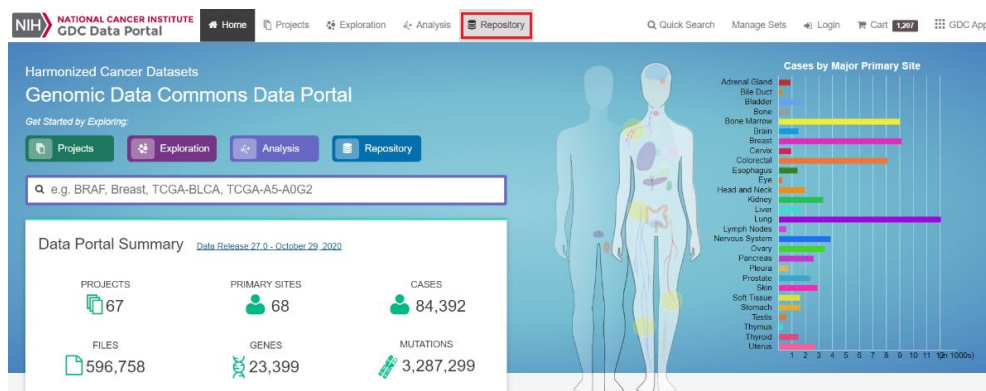
سپس با در نظر گرفتن هر کدام از این ۸ زیرمجموعه به عنوان ویژگی، با استفاده از یک مدل SVM با کرنل RBF، و یک مدل random forest، مجموعاً ۱۶ بار ارزیابی به روش 10 fold cross-validation انجام شده و نمودار specificity نسبت به شماره زیرمجموعه رسم شده است. با توجه به چالش موجود برای پیش‌بینی برچسب درست موارد سالم، انتظار این بوده است که با افزایش شماره زیرمجموعه استفاده شده به عنوان ویژگی، specificity به دست آمده نیز کاهش پیدا کرده و نمودارها شکل نزولی داشته باشند.

## پایه‌سازی مراحل

### دریافت داده‌ها

برای دریافت داده‌ها به سایت <https://portal.gdc.cancer.gov> مراجعه کرده و از نوار بالای صفحه وارد بخش Respiratory

می‌شویم.



Files **Cases**

[Add a File Filter](#)

Search Files ?

Q e.g. 142682.bam, 4f6e2e7a-b...

Data Category

☐ transcriptome profiling # Files 1,207

Data Type

☐ Aligned Reads # Files 1,207

☐ Isoform Expression Quantification # Files 1,207

☒ miRNA Expression Quantification # Files 1,207

Experimental Strategy

☒ miRNA-Seq # Files 1,207

Workflow Type

☐ BCGSC miRNA Profiling # Files 1,207

در صفحه باز شده ابتدا در ستون سمت چپ فیلترهای مشخص شده در تصویر مقابل را اعمال کرده و سپس روی Cases کلیک می‌کنیم.

در فیلترهای بخش Cases نیز موارد مشخص شده در تصویر زیر را انتخاب می‌کنیم. با انتخاب این موارد در مجموع با در نظر گرفتن همه این فیلترها به ۱۲۰۷ فایل با فرمت txt. می‌رسیم، که هر کدام یک پروفایل بیانی miRNA بوده و در آن‌ها بیان ۱۸۸۱ مورد miRNA در یک آزمایش اندازه‌گیری شده است. با توجه به اینکه در مقاله نیز به همین تعداد ۱۲۰۷ آزمایش اشاره شده است، به احتمال زیاد در اعمال فیلترها درست عمل کرده‌ایم. حالا برای دریافت فایل‌ها ابتدا روی Add All Files to Cart کلیک کرده و سپس در بالای صفحه روی Cart کلیک می‌کنیم.

NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home Projects Exploration Analysis Repository

Quick Search Manage Sets Login **Cart 1,207** GDC Apps

Files **Cases**

[Add a Case/Biospecimen Filter](#)

Search Cases ?

Q e.g. TCGA-A5-A0G2, 432fe4a9-2...

[Upload Case Set](#)

Case ID

eg. TCGA-DD\*, \*DD\*, TCGA-DD-AAVP Go!

Primary Site

☒ breast # Cases 1,079

Program

☒ TCGA # Cases 1,079

Project

☒ TCGA-BRCA # Cases 1,079

☐ TCGA-DLBC 1

Disease Type

# Cases

[Clear](#) [Primary Site](#) [IS](#) [breast](#) [AND](#) [Program Name](#) [IS](#) [TCGA](#) [AND](#) [Project Id](#) [IS](#) [TCGA-BRCA](#) [AND](#) [Data Type](#) [IS](#) [miRNA Expression Quantification](#) [AND](#) [Experimental Strategy](#) [IS](#) [miRNA-Seq](#) [Advanced Search](#)

[Add All Files to Cart](#) [Manifest](#) [View 1,079 Cases in Exploration](#) [View Images](#)

Files (1,207) Cases (1,079)

Primary Site Project Data Category Data Type Data Format

Show More

Showing 1 - 20 of 1,207 files 60.61 MB

Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
open	732ee0df-c0fa-4fa3-aa77-989a11172285.mirbase21.mimas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.28 KB	0
open	cd88d4aa-99b6-495e-97fb-1a062aed48a2.mirbase21.mimas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.05 KB	0
open	e11fe3de-a466-4cdf-89d2-3f5c0e86a304.mirbase21.mimas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.2 KB	0
open	5d6655d3-c63d-44df-b928-33105a0b0033.mirbase21.mimas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.2 KB	0

در صفحه باز شده امکان دانلود انواع داده‌ها از جمله داده‌های کلینیکال مربوط به افراد، داده‌های مربوط به نمونه‌برداری و نحوه انجام تست‌ها، و البته فایل‌های پروفایل بیانی **miRNA**ها وجود دارد. با توجه به اینکه در این مطالعه تنها داده‌های مربوط به سطح بیان **miRNA**ها به عنوان ویژگی مورد استفاده قرار می‌گیرند، تنها به داده‌های پروفایل بیانی **miRNA**ها نیاز داریم. برای دانلود این ۱۲۰۷ فایل، ابتدا روی **Download** کلیک کرده و سپس روی **Cart** کلیک می‌کنیم. یک فایل فشرده شده با فرمت **.gz** شامل ۱۲۰۷ فولدر مربوط به هر یک از آزمایش‌ها دانلود می‌شود، که در هر یک از این ۱۲۰۷ فولدر یک فایل **.txt** با نامی متفاوت نسبت به نام فولدر قرار گرفته است. در این فایل‌های **.txt** اطلاعاتی در رابطه با فرد آزمایش‌شونده وجود ندارد و بنابراین نمی‌توانیم برچسب مربوط به هر آزمایش (اینکه فرد سرطان داشته یا خیر) را از این فایل‌ها به دست آوریم. بنابراین برای رسیدن به برچسب هر آزمایش روی **Sample Sheet** کلیک می‌کنیم. یک فایل **.txt** دیگر دانلود می‌شود.

The screenshot shows the GDC Data Transfer Tool interface. On the left, it displays 'FILES 1207', 'CASES 1079', and 'FILE SIZE 60.61 MB'. The main area is divided into 'File Counts by Project' and 'File Counts by Authorization Level'. The 'Project' table shows TCGA-BRCA with 1079 cases, 1,207 files, and 60.61 MB. The 'Authorization Level' table shows 'Authorized' with 1207 files and 60.61 MB. On the right, there are instructions on how to download files in the Cart, including a 'Download Manifest' section and a 'Download Cart' section. At the bottom, there are buttons for 'Biospecimen', 'Clinical', 'Sample Sheet' (highlighted with a red box), 'Metadata', 'Download', and 'Remove From Cart'. A dropdown menu for 'Download' is open, showing 'Manifest' and 'Cart' (highlighted with a red box). Below this, there is a 'Cart Items' section showing a list of files with columns for 'Remove', 'Access', 'File Name', 'Cases', 'Project', 'Data Category', 'Data Format', 'File Size', and 'Annotations'.

Remove	Access	File Name	Cases	Project	Data Category	Data Format	File Size	Annotations
	open	732ee0df-c0fa-4fa3-aa77-989a11172285.mirbase21.mirnas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.28 KB	0
	open	cd88d4aa-99b6-495e-97fb-1a062aed48a2.mirbase21.mirnas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.05 KB	0
	open	e1f1e3de-a466-4cdf-89d2-3f5c0e86a304.mirbase21.mirnas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.2 KB	0
	open	5d6655d3-c63d-44df-b928-33105a0b0033.mirbase21.mirnas.quantification.txt	1	TCGA-BRCA	Transcriptome Profiling	TXT	50.2 KB	0

## خوانش داده‌ها و preprocessing

برای خواندن فایل‌های تکست درون فایل فشرده دانلود شده در قسمت قبلی، **tarfile** فراخوانی شده و با استفاده از تابع **extractall()** ۱۲۰۷ فولدر درون فایل فشرده **extract** شده‌اند. در هر کدام از این فولدرها یک فایل تکست وجود دارد. با استفاده از تابع **pandas.read\_csv()** اطلاعات این ۱۲۰۷ فایل خوانده شده و در یک **dataframe** ذخیره می‌شود، به شکلی که در **dataframes[i]** اطلاعات مربوط به فایل **i**ام قرار گرفته باشد. در تصویر زیر ساختار اطلاعات هر فایل قابل مشاهده است:

	miRNA_ID	read_count	reads_per_million_miRNA_mapped	cross-mapped
0	hsa-let-7a-1	5641	6743.308393	N
1	hsa-let-7a-2	5707	6822.205460	N
2	hsa-let-7a-3	5587	6678.756248	N
3	hsa-let-7b	9112	10892.576862	N
4	hsa-let-7c	1452	1735.735470	N
...	...	...	...	...
1876	hsa-mir-9500	0	0.000000	N
1877	hsa-mir-96	27	32.276073	N
1878	hsa-mir-98	114	136.276752	N
1879	hsa-mir-99a	395	472.186991	Y
1880	hsa-mir-99b	37435	44750.177220	N

ستون اول مشخص‌کننده miRNA بوده و در ستون دوم تعداد readهای شمرده شده آن miRNA نوشته شده است. همچنین تعداد readهای نرمال شده نسبت به کل readها در ستون سوم قرار گرفته است. در مجموع در هر آزمایش سطح بیان ۱۸۸۱ مورد miRNA مشخص شده و ۱۲۰۷ مورد داده به همین شکل در dataframes ذخیره شده است.

**نکته:** درون تعداد کمی از ۱۲۰۷ فولدر موجود در فایل فشرده دانلود شده بیش از یک فایل تکست موجود است. فایل(های) دیگر با نام annotations.txt مشخص شده‌اند که در آن‌ها معمولاً اصلاحاتی در رابطه با داده‌های کلینیکال فرد مربوط به آن آزمایش ثبت شده است. بنابراین در خوانش داده‌ها به این فایل‌ها کاری نداریم.

برای به دست آوردن برچسب داده‌ها فایل تکست sample sheet را می‌خوانیم. ساختار داده‌های این فایل به شکل زیر است.

	File ID	File Name	Data Category	...	Case ID	Sample ID	Sample Type
0	ced45efd-cd54-44ea-b80e-ff04b1f061a6	732ee0df-c0fa-4fa3-aa77-989a11172285.mirbase21...	Transcriptome Profiling	...	TCGA-A7-A3IY	TCGA-A7-A3IY-01A	Primary Tumor
1	b8647788-d1b1-406d-baa6-607fb43d8056	cd88d4aa-99b6-495e-97fb-1a062aed48a2.mirbase21...	Transcriptome Profiling	...	TCGA-A8-A080	TCGA-A8-A080-01A	Primary Tumor
2	0174aa45-bcca-4677-b619-593ed2119a05	e1f1e3de-a466-4cdf-89d2-3f5c0e86a304.mirbase21...	Transcriptome Profiling	...	TCGA-EW-A1P0	TCGA-EW-A1P0-01A	Primary Tumor
3	d4976851-cef0-47d5-9ea6-8c72ece9ed7a	5d6655d3-c63d-44df-b928-33105a0b0033.mirbase21...	Transcriptome Profiling	...	TCGA-E2-A153	TCGA-E2-A153-01A	Primary Tumor
4	a607c74a-4ed5-4e2b-bf5b-3cc9e3214335	6380cd41-2628-484f-acd3-2240d9903fba.mirbase21...	Transcriptome Profiling	...	TCGA-A0-A03L	TCGA-A0-A03L-01A	Primary Tumor
...	...	...	...	...	...	...	...
1202	c645f59d-0770-4609-b4a7-0e2c38ee0ebc	c9b9ed6a-ff3e-4f10-bd30-2b746dc43404.mirbase21...	Transcriptome Profiling	...	TCGA-A0-A030	TCGA-A0-A030-01A	Primary Tumor
1203	2bb3034f-d37a-4eb4-b633-a3952ab947dd	1f4d0bb9-23e8-4ba0-93cb-8e5efdcfcf9a.mirbase21...	Transcriptome Profiling	...	TCGA-BH-A18R	TCGA-BH-A18R-11A	Solid Tissue Normal
1204	89bbfa2d-ba3d-4f14-bf01-7292e5958215	636c9329-44fb-4dd4-b0d6-2bf0bc9531de.mirbase21...	Transcriptome Profiling	...	TCGA-A8-A07R	TCGA-A8-A07R-01A	Primary Tumor
1205	2a0bad5d-d833-4434-9a0f-f8eda09cb5c9	90e83adc-6d01-4959-b6bd-08d156f45b0f.mirbase21...	Transcriptome Profiling	...	TCGA-AN-A046	TCGA-AN-A046-01A	Primary Tumor
1206	4054a92c-e4c6-4742-99fc-2e9bc0b99ed5	5af377ab-7c3e-4a8d-9583-51b667d82700.mirbase21...	Transcriptome Profiling	...	TCGA-BH-A18K	TCGA-BH-A18K-11A	Solid Tissue Normal

ستون اول (File ID) نام فولدر هر یک از آزمایش‌ها را مشخص می‌کند. ستون دوم نام فایل تکست هر آزمایش بوده و در ستون آخر نیز برچسب مورد نظرمان برای آن آزمایش ثبت شده است. برچسب‌ها به شکل Primary Tumor، Solid Tissue Normal و Metastatic هستند. در این مطالعه موارد Primary Tumor و Metastatic به عنوان موارد دارای سرطان (کلاس ۱) و موارد Solid Tissue Normal به عنوان موارد سالم (کلاس ۰) در نظر گرفته شده‌اند. با شمارش موارد Solid Tissue Normal متوجه می‌شویم که از این ۱۲۰۷ مورد ۱۰۴ مورد مربوط به کلاس موارد سالم بوده است، که این عدد با عدد گزارش شده در مقاله تطابق دارد.

در اینجا چالش اصلی عدم تطابق ترتیب آزمایش‌ها در sample sheet با ترتیب خوانش و ذخیره سازی نتایج آزمایش‌ها در dataframes است. برای اینکه sample sheet را با همان ترتیبی که ۱۲۰۷ فایل را خوانده‌ایم مرتب کنیم تا تناظر بین برچسب هر آزمایش با داده مربوط به آن آزمایش به راحتی برقرار شود. برای این کار ردیف‌های sample sheet را نسبت به مقادیر ستون File ID مرتب می‌کنیم. پس از انجام این کار می‌توان دید که ترتیب قرارگیری نام فایل‌های تکست در ستون دوم sample sheet (File Name) مطابق ترتیب خوانده شدن و ذخیره‌سازی این فایل‌ها در dataframes می‌باشد. بنابراین حالا برچسب مربوط به جدول ۱۸۸۱×۴ نام ذخیره شده در dataframes از ستون آخر ردیف نام sample sheet به دست می‌آید. درنهایت dataframes در قالب یک Dataframe با نام table، همراه با ردیف‌های نام‌گذاری شده بر اساس شماره sample و ستون‌های نامگذاری شده بر اساس نام miRNA، ذخیره شده است.

برچسب‌ها نیز در Y ذخیره شده‌اند. به این شکل که از ردیف اول تا آخر sample sheet مرتب‌شده در صورتی که Solid Tissue Normal ذخیره شده باشد، عدد ۰؛ و در غیر این صورت عدد ۱ در Y ذخیره شده است.


با توجه به توضیحات مقاله تنها از ستون سوم داده‌های بیان miRNA یعنی تعداد read ها نرمال شده، به عنوان feature استفاده شده است. برای هر یک از ۱۲۰۷ آزمایش، عدد ۱ به عنوان pseudo count به مقادیر ستون دوم اضافه شده و سپس از این مقادیر لگاریتم در پایه ۲ گرفته شده است (دلیل اضافه کردن یک این است که لگاریتم در پایه دو برای miRNA هایی که reads\_per\_million\_miRNA\_mapped برای آن‌ها صفر بوده است منفی بی‌نهایت نشود). این مقادیر در آرایه دو بعدی features (۱۲۰۷×۱۸۸۱) ذخیره شده‌اند. حالا ۱۸۸۱ ستون features را استاندارد می‌کنیم، به شکلی که میانگین مقادیر هر ستون ۰ و واریانس آن ۱ باشد. در نهایت features در قالب یک dataframe با نام table\_scaled، همراه با ردیف‌های نام‌گذاری شده بر اساس شماره sample و ستون‌های نامگذاری شده براساس نام miRNA ها، ذخیره شده است. پس از استانداردسازی داده‌ها ستون‌هایی که همه ۱۲۰۷ مقدار آن‌ها صفر نبوده‌اند را می‌یابیم. صفر بودن یا نبودن هر یک از ستون‌های table\_scaled به شکل Boolean در non\_zero ذخیره شده است. مقادیر ۲۷۷ ستون صفر می‌باشد، بنابراین از ۱۸۸۱ miRNA بررسی شده ۱۶۰۴ مورد مقادیر غیر صفر داشته‌اند. با انتخاب ستون‌های با مقادیر غیر صفر از table\_scaled دیتافریم table\_scaled\_no\_zero می‌رسیم.

در جدول ۱ مقاله تعدادی miRNA که در آزمایشگاه ارتباط آن‌ها با سرطان پستان مشخص شده معرفی شده‌اند. در ادامه مراحل مطالعه، تنها ستون‌های مربوط به میزان بیان این موارد را نگه داشته شده‌اند. در واقع مراحل feature selection و سپس classification روی این miRNA های clinically verified انجام شده است. برای انتخاب اطلاعات مربوط به این miRNA ها باید بدانیم کدام یک از ستون‌های features مربوط به این miRNA ها هستند. بنابراین miRNA های مشخص شده در جدول یک را کپی کرده، و یک لیست می‌سازیم که هر کدام از اعضای آن یکی از این miRNA ها باشد. سپس از ردیف اول ستون اول یکی از داده‌های dataframes تا ردیف ۱۸۸۱ پیمایش کرده و شماره ردیف‌هایی که نام miRNA آن ردیف در لیست miRNA های clinically verified موجود باشد را در wetlab\_miRNA\_index ذخیره می‌کنیم. طول لیست wetlab\_miRNA\_index ۳۵ خواهد بود، در حالی که در جدول ۱ مقاله ۳۸ miRNA مشخص شده بود. با انجام بررسی‌های بیشتر متوجه می‌شویم که در جدول ۱ مقاله دو مورد از miRNA ها، یعنی hsa-mir-17 و hsa-mir-206، دوبار نوشته شده‌اند و همچنین hsa-mir-213 به اشتباه به شکل has-mir-213 تایپ شده است.

Table 1. Clinically Verified miRNA.

miRNA [14]			
hsa-mir-10b	hsa-let-7d	hsa-mir-206	hsa-mir-34a
hsa-mir-125b-1	hsa-let-7f-1	hsa-mir-17	hsa-mir-27b
hsa-mir-145	hsa-let-7f-2	hsa-mir-335	hsa-mir-126
hsa-mir-21	hsa-mir-206	hsa-mir-373	hsa-mir-101-1
hsa-mir-125a	hsa-mir-30a	hsa-mir-520c	hsa-mir-101-2
hsa-mir-17	hsa-mir-30b	hsa-mir-27a	hsa-mir-146a
hsa-mir-125b-2	hsa-mir-203a	hsa-mir-221	hsa-mir-146b
hsa-let-7a-2	hsa-mir-203b	hsa-mir-222	hsa-mir-205
hsa-let-7a-3	has-mir-213	hsa-mir-200c	
hsa-let-7c	hsa-mir-155	hsa-mir-31	



 <span>miRBase</span>	
<a href="#">Home</a> <a href="#">Search</a> <a href="#">Browse</a> <a href="#">Help</a> <a href="#">Download</a> <a href="#">Blog</a> <a href="#">Submit</a> <a href="#">hsa-mir-181a-1</a>	
Stem-loop sequence hsa-mir-181a-1	
Accession	MI0000289 ( <a href="#">change log</a> )
Previous IDs	<b>hsa-mir-213</b>
Symbol	HGNC:MIR181A1
Description	<i>Homo sapiens</i> miR-181a-1 stem-loop
Gene family	MIPF0000007; <a href="#">mir-181</a>

با حذف دو مورد اضافه و تصحیح موردی که اشتباه تایپی داشته است مجدداً بررسی‌ها انجام شده‌اند. این بار متوجه می‌شویم که اصلاً موردی با نام hsa-mir-213 در میان miRNAهای مشخص شده در ستون اول فایل‌ها وجود نداشته است. برای انجام

بررسی‌های بیشتر در رابطه با این miRNA ابتدا به مقاله رفرنس این جدول مراجعه شد، که نتیجه‌ای به دست نیامد، زیرا در این مقاله نیز از همین نام برای این miRNA استفاده شده بود. بنابراین در ادامه نام این miRNA را در دیتابیس miRBase سرچ کردیم و متوجه شدیم که hsa-mir-213 نام قدیمی hsa-mir-181a-1 بوده است. با سرچ کردن hsa-mir-181a-1 در ستون اول فایل‌های مورد استفاده می‌بینیم که در فایل‌ها نیز از نام جدید این miRNA استفاده شده است.

بنابراین نام جدید این miRNA را جایگزین hsa-mir-213 کرده، این ۳۶ miRNA تصحیح شده را به عنوان miRNAهای clinically verified در نظر می‌گیریم.

با استفاده از لیست wetlab\_miRNAs\_list اصلاح شده (شامل ۳۶ نام miRNA) ستون‌های مربوط به miRNAهای clinically verified را از table\_scaled\_no\_zero انتخاب کرده و در table\_scaled\_no\_zero\_clinical ذخیره می‌کنیم. در مراحل بعدی از table\_scaled\_no\_zero\_clinical برای انتخاب ویژگی و دسته‌بندی استفاده خواهیم کرد.

## روش انتخاب ویژگی Information Gain

روش Information Gain (IG) یک روش انتخاب ویژگی مبتنی بر تئوری اطلاعات است، که کاهش آنتروپی ناشی از داشتن اطلاعات از یک ویژگی را می‌سنجد. برای دیتاست  $X$  با  $n$  کلاس مختلف، آنتروپی Shannon که سنج‌های از underpredictability است، با رابطه زیر مشخص می‌شود.

$$H(X) = - \sum_{i=1}^n p_i \log_2 p_i.$$

که در آن  $P_i$  احتمال کلاس  $i$  در دیتاست  $X$  است. IG کاهش آنتروپی ناشی از دانستن ویژگی  $A$  می‌باشد که با رابطه زیر نشان داده می‌شود.

$$IG(X, A) = H(x) - H(X|A)$$

که در آن:

$$H(X|A) = \sum_{i=1}^v \frac{X_i}{X} H(X_i)$$

که  $X_i$  زیرمجموعه‌ای از دیتاست  $X$  است که برای ویژگی  $A$  دارای مقداری خاص است.  $V$  تعداد مقادیر متفاوت موجود برای ویژگی  $A$  بوده و  $H(X_i)$  آنتروپی آنتروپی زیرمجموعه  $A$ م هنگام تقسیم  $X$  بر مبنای مقادیر ویژگی  $A$  است، بنابراین می‌توان به روش IG به صورت تفاوت میان آنتروپی اولیه و آنتروپی پس از تقسیم دیتاست اولیه بر مبنای ویژگی  $A$  نگاه کرد.

برای به دست آوردن Information Gain، تفاوت آنتروپی بین دسته اصلی با دسته‌هایی که در نهایت به دست می‌آیند محاسبه می‌شود. در تابع آنتروپی، با توجه به فرمول فوق، ابتدا باید تعداد تکرار هر عنصر به دست آید که با دستور np.unique و با True قرار دادن return\_counts انجام می‌شود. در واقع این تابع با گرفتن ستون مورد نظر، آنتروپی ستون را خروجی می‌دهد. فرمول آنتروپی برای عددهای به دست آمده اعمال می‌شود و در نهایت تابع مقدار آنتروپی دسته‌ها را خروجی می‌دهد. در تابع InfoGain، ابتدا آنتروپی دسته اولیه که شامل تمام افراد بیمار و سالم است محاسبه شده و در متغیر total\_entropy قرار می‌گیرد. این تابع دو متغیر را به عنوان ورودی دریافت می‌کند، که یکی داده‌ها و دیگری ستون هدفی می‌باشد که می‌خواهیم آنتروپی را براساس آن حساب کنیم. ستون‌های هدف در واقع read count های process شده miRNA های ما هستند که مقادیر مثبت یا منفی دارند (به دلیل حذف کردن صفرها در مراحل pre-process) و همچنین مقادیر نزدیک به صفر دارند (به دلیل نرمالسازی در مراحل pre-process). می‌خواهیم بر اساس این ویژگی‌ها محاسبه کنیم که اگر هر کدام از این ستون‌ها برای دسته‌بندی در نظر گرفته شوند، آنتروپی در دسته‌های حاصل شده چه مقدار کاهش خواهد یافت، و یا به عبارت دیگر اطلاعات بیشتری به دست خواهیم آورد. برای دسته‌بندی داده‌ها به دو دسته، داده‌هایی که برای ویژگی مورد بررسی دارای مقدار مثبت هستند در متغیر positive و داده‌هایی که برای این ویژگی مقداری منفی دارند، در متغیر negative ریخته می‌شوند. سپس، آنتروپی هر یک از دو دسته محاسبه شده، متناسب با تعداد اعضای هر دسته نرمال شده و در نهایت آنتروپی‌های نرمال شده با هم جمع می‌شوند. این نتیجه در متغیر New\_entropy ریخته می‌شود که در قدم بعدی از total\_entropy کم می‌گردد. بنابراین مقدار حاصل شده Information\_Gain است. هر چه میزان این خروجی بزرگتر باشد، ویژگی ما امتیاز بالاتری گرفته‌است.

علاوه بر پیاده‌سازی روش انتخاب ویژگی Information Gain، نتایج انتخاب ویژگی با استفاده از معیار Mutual information نیز به دست آمده است. Mutual information مفهومی بسیار نزدیک به Information gain می‌باشد. در رابطه با ارتباط این دو مفهوم با یکدیگر می‌توان گفت، هر چه تفاوت بیشتری بین توزیع‌های توام و marginal وجود داشته باشد (mutual information)، information gain بزرگتر خواهد بود.

برای پیاده‌سازی انتخاب ویژگی با استفاده از معیار mutual information از تابع sklearn.feature\_selection.SelectKbest با score\_func = mutual\_info\_classif استفاده شده است. خروجی تابع براساس امتیاز داده شده به ستون‌ها مرتب شده و ۱۰ ویژگی برتر که بیشترین امتیاز را به دست آورده بودند بر این اساس مشخص شده‌اند.

## روش انتخاب ویژگی مربع Chi

روش Chi2 یک روش انتخاب ویژگی دیگر است، که ویژگی‌های مختلف را نسبت به کلاس‌ها ارزیابی می‌کند. این روش یک روش آماری برای مشخص کردن وابستگی یک ویژگی به برجسب کلاس‌ها می‌باشد. بر این اساس می‌توان ویژگی‌هایی که میان آن‌ها و برجسب‌ها وابستگی‌ای وجود ندارد را حذف کرده و ویژگی‌هایی که برای انجام دسته‌بندی مناسب‌تر هستند را انتخاب کرد.

در این روش لازم است تا ویژگی‌های پیوسته با انجام بازه‌بندی به شکل گسسته دربیایند.

$$\chi^2 = \sum_{i=1}^C \sum_{j=1}^I \frac{(A_{ij} - E_{ij})^2}{E_{ij}}$$

که در آن C تعداد کلاس‌ها، I تعداد بازه‌ها،  $E_{ij}$  تعداد نمونه‌های قابل انتظار و  $A_{ij}$  تعداد نمونه‌های کلاس  $C_i$  در بازه  $I_j$  می‌باشد. هرچه مقدار  $\chi^2$  بزرگ‌تر باشد، آن ویژگی اطلاعات بیشتری در مورد کلاس داده‌ها به ما می‌دهد.

برای پیاده‌سازی روش انتخاب ویژگی مربع Chi از تابع `sklearn.feature_selection.chi2` استفاده شده است. این تابع داده‌ها و برجسب‌های آن‌ها را به عنوان ورودی گرفته و برای هر یک از ستون‌ها امتیاز و  $p$  value به دست آمده از روش انتخاب ویژگی Chi2 را خروجی می‌دهد (از ۲۰۷ داده برای انتخاب ویژگی استفاده شده و در مراحل بعدی از ویژگی‌های انتخاب شده با استفاده از این ۲۰۷ داده، برای دسته‌بندی و ارزیابی روی ۱۰۰۰ داده باقی‌مانده استفاده شده است). خروجی تابع بر اساس امتیاز ستون‌ها مرتب شده و به این صورت ۱۰ ستونی که بیشترین امتیاز را به دست آورده بودند مشخص می‌گردد.

**نکته:** با توجه به اینکه ورودی تابع `chi2` نمی‌تواند مقادیر منفی باشد و جنس ورودی این تابع `frequency` است، از داده‌های خام مربوط به ۳۶ miRNA مورد بررسی به جای داده‌های `process` استفاده شده است.

## روش انتخاب ویژگی LASSO

روش LASSO یک روش منظم‌سازی و انتخاب ویژگی برای مدل‌های آماری است. LASSO مجموع مربعات خطاها را در حالی که توسط مجموع قدر مطلق ضرایب مدل رگرسیون محدود می‌شود، کمینه می‌کند.

$$\min_{\beta_0, \beta} \sum_{i=1}^N (y_i - \beta_0 - x_i^T \beta)^2 \text{ s.t. } \sum_{j=1}^p |\beta_j| \leq t$$

N تعداد داده‌ها، p تعداد ویژگی‌ها،  $(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  ضرایب رگرسیون،  $y_i$  خروجی پیش‌بینی شده برای داده  $i$ ام و  $x_i$  مجموعه ویژگی‌های این داده می‌باشد. با تنظیم پارامتر  $t$  ضرایب ویژگی‌های کم اهمیت‌تر به صفر نزدیک شده و ویژگی‌های مهم‌تر قابل شناسایی خواهند بود.

برای پیاده‌سازی روش انتخاب ویژگی `lasso`، ابتدا یک مدل رگرسیون لجستیک با منظم‌ساز `lasso` و روش بهینه‌سازی `liblinear` با استفاده از `sklearn.feature_selection.SelectFromModel` در نظر گرفته شده است. سپس مدل روی ۲۰۷ داده مورد

استفاده برای انتخاب ویژگی آموزش می‌بیند. با استفاده از `_coef_estimator`. می‌توان به ضرایب مدل آموزش دیده دسترسی پیدا کرد. ویژگی‌ها براساس اندازه ضرایب آن‌ها در مدل آموزش دیده مرتب شده و ۱۰ ویژگی برتر بر این اساس انتخاب شده‌اند.

همچنین در ادامه مشابه با روش توضیح داده شده برای پیاده‌سازی انتخاب ویژگی براساس منظم‌ساز `lasso`، انتخاب ویژگی براساس منظم‌سازی با استفاده از `term` تنبیه `Ridge` نیز انجام شده است.

## بررسی `correlation` میان ویژگی‌های برتر انتخاب شده با یکدیگر

با توجه به عدم تطابق کامل ویژگی‌های انتخاب شده با لیست ۱۰ ویژگی برتر به دست آمده در مقاله، برای ۱۵ ویژگی برتر به دست آمده با استفاده از روش `Chi2`، ماتریس `correlation` محاسبه شده است. با توجه به اینکه روش‌های انتخاب ویژگی استفاده شده `univariate` بوده و ویژگی‌ها را به شکل تکی بررسی کرده و به آن‌ها امتیاز می‌دهند، هدف از انجام این کار این است که شاید با محاسبه ماتریس `correlation` بتوانیم بعضی از موارد اضافه‌ای که در میان بهترین ویژگی‌های برتر ما قرار گرفته‌اند اما در لیست مقاله وجود نداشته‌اند را در صورت داشتن `correlation` بالا با ویژگی‌هایی که رنکینگ بهتری نسبت به آن‌ها داشته‌اند، حذف کنیم و به لیستی شبیه‌تر به لیست مقاله برسیم. (در واقع حدس ما این بود که با توجه به `univariate` بودن روش‌های انتخاب ویژگی استفاده شده در مقاله، احتمالاً در کار انجام شده توسط نویسندگان مقاله، مواردی که با یکدیگر `correlation` بالایی داشته‌اند از لیست ویژگی‌های برتر حذف شده‌اند، و به این دلیل شاهد تفاوت در نتایج خودمان با نتایج مقاله هستیم.)

**نکته مهم:** در کار انجام شده توسط نویسندگان مقاله، از تمامی ۱۲۰۷ داده برای انتخاب ۱۰ ویژگی برتر استفاده شده، و سپس مراحل مختلف ارزیابی با استفاده از ویژگی‌های برتر نیز به شکل `10-fold cross-validation` روی همان داده‌ها انجام شده است. انتخاب ویژگی براساس داده‌هایی که قصد داریم ارزیابی را روی خودشان انجام دهیم، به شکل واضحی ایجاد `bias` کرده و می‌تواند باعث `unfair` شدن ارزیابی‌های انجام شده شود. بنابراین در پیاده‌سازی کار مقاله ما از قسمتی از داده‌ها (۲۰۷ داده) برای انتخاب ویژگی‌ها استفاده کرده، و ارزیابی با استفاده از ویژگی‌های انتخاب شده را روی ۱۰۰۰ داده دیگر باقی‌مانده به شکل `10-fold cross-validation` انجام دادیم. به نظر می‌رسد دلیل اصلی تفاوت لیست به دست آمده ما با لیست مقاله، انتخاب ویژگی بر مبنای تعداد محدودی از داده‌ها باشد.

## ارزیابی با حالت‌های مختلف در نظر گرفتن ویژگی‌های برتر

در این بخش هدف ساخت مجدد جدول ۲ مقاله بوده است.

در پیاده‌سازی انجام شده از توابع `RandomForestClassifier` از `sklearn.ensemble` و `SVC` از `sklearn.svm` استفاده شده است.

تمامی ارزیابی‌ها روی ۱۰۰۰ داده استفاده نشده در مرحله انتخاب ویژگی، و به شکل 10 fold cross-validation انجام شده است. برای هر classifier حالت‌های مختلف در نظر گرفتن ویژگی‌ها، که در ادامه قابل مشاهده می‌باشد، بررسی گردیده است.

برای هر روش انتخاب ویژگی حالت استفاده از ۳، ۵، یا ۱۰ ویژگی برتر انتخاب شده براساس آن روش در نظر گرفته شده است. علاوه بر این حالت‌ها ۳ حالت استفاده از تمامی miRNAهای غیرصفر (۱۶۰۴ مورد)، استفاده از miRNAهای غیرصفر غیر کلینیکال (۱۵۶۸ مورد) و استفاده از تمامی ۳۶ miRNA تایید شده در کلینیک، در دسته‌بندی به وسیله هر یک از ۳ مدل SVM خطی، SVM با کرنل RBF و Random Forest مورد بررسی قرار گرفته است.

ابتدا با استفاده از تابع cross\_val\_predict از sklearn.model\_selection براساس 10 fold cross validation برای داده‌های validation\_data (۱۰۰۰ داده آخر پس از shuffle شدن مجموعه داده‌های اولیه) و برجسب‌های validation\_label پیش‌بینی انجام شده و سپس با استفاده از تابع measure، معیارهای مورد نیاز (Sensitivity, Accuracy, Specificity و AUC) محاسبه می‌گردد. (در تابع measure ابتدا با استفاده از تابع confusion\_matrix از sklearn.metrics، با مقایسه برجسب‌های واقعی و برجسب‌های پیش‌بینی شده، تعداد TP، TN، FP و FN محاسبه شده و سپس براساس این مقادیر، sensitivity و specificity محاسبه می‌گردد؛ همچنین برای محاسبه AUC از تابع roc\_curve از sklearn.metrics استفاده شده است).

خروجی تابع measure برای هر یک از حالت‌های انتخاب ۳، ۵، یا ۱۰ ویژگی برتر انتخاب شده توسط یک روش انتخاب ویژگی مشخص و در حالت استفاده از یکی از ۳ Classifier، در لیستی با نام مشخص‌کننده نام دسته‌بند و روش انتخاب ویژگی استفاده شده ذخیره می‌گردد. (مثلا در IG-Linear نتایج تابع measure به ترتیب برای حالت استفاده از ۱۰، ۵، و ۳ ویژگی برتر به دست آمده از روش Information Gain و دسته‌بندی با مدل SVM خطی ذخیره شده است).

همچنین خروجی تابع measure برای هر یک از سه حالت کلی در نظر گرفتن تمامی miRNAهای غیر صفر، در نظر گرفتن miRNAهای غیر صفر غیر کلینیکال، و در نظر گرفتن تمام ۳۶ miRNA تایید شده در کلینیک، به عنوان ویژگی؛ در لیست‌هایی با پیشوند All و نام دسته‌بند استفاده شده ذخیره شده است. (برای مثال در All\_RF نتیجه دسته‌بندی با مدل random forest برای ۳ حالت کلی در نظر گرفتن ویژگی‌ها ذخیره شده است).

سپس لیست مربوط به نتایج حالت کلی در نظر گرفتن ویژگی‌ها، لیست مربوط به نتایج ۳ حالت در نظر گرفتن (۳، ۵، ۱۰) ویژگی‌های برتر به دست آمده از روش انتخاب ویژگی IG، لیست مربوط به نتایج ۳ حالت در نظر گرفتن (۳، ۵، ۱۰) ویژگی‌های برتر به دست آمده از روش انتخاب ویژگی CHI2، و لیست مربوط به نتایج ۳ حالت در نظر گرفتن (۳، ۵، ۱۰) ویژگی‌های برتر به دست آمده از روش انتخاب ویژگی LASSO، در لیستی با نام دسته‌بند مربوطه ذخیره می‌شوند. (برای مثال RF شامل All\_RF، IG\_RF، CHI2\_RF و LASSO\_RF خواهد بود).

در مرحله بعد یک لیست به نام func ساخته شده است که شامل RBF، RF، (مربوط به SVM با کرنل RBF) و Linear (مربوط به SVM خطی) می‌باشد.

در نهایت با استفاده از تابع make\_table مقادیر ذخیره شده در لیست سه بعدی func در قالب یک جدول print می‌شود.

## رسم نمودار **Specificity** نسبت به شماره زیرمجموعه‌های ۳ عضوی ۱۰ ویژگی برتر

برای رسم این نمودار از دو تابع `cal_specificity` و `specificity_plot` استفاده می‌شود.

**نکته:** در پیاده‌سازی انجام شده به دلیل به استفاده از تعداد محدودی از داده‌ها برای انتخاب ویژگی، برخلاف نتیجه به دست آمده در مقاله در ۱۰ ویژگی برتر به دست آمده از هر یک از روش‌ها با یکدیگر تطابق کامل نداشته‌اند. بنابراین در این قسمت برای ویژگی‌های برتر به دست آمده از هر یک از روش‌های استفاده شده به شکل جداگانه نمودار **Specificity** نسبت به شماره زیرمجموعه‌های ۳ عضوی ۱۰ ویژگی برتر را رسم می‌کنیم.

تابع `cal_specificity` دسته‌بند مورد نظر (SVM با کرنل RBF یا Random Forest)، لیست ۱۰ miRNA برتر به دست آمده از یکی از روش‌های انتخاب ویژگی، برچسب واقعی داده‌های `validation` و مجموعه ویژگی‌های داده‌های `validation` را به عنوان ورودی دریافت می‌کند. سپس با استفاده از تابع `cross_val_predict` از `sklearn.model_selection` به صورت `10 fold cross validation` برچسب داده‌های `validation` را به ازای ۸ حالت مختلف در نظر گرفتن ۳ ویژگی (ویژگی‌های برتر ۱ تا ۳ از لیست ۱۰ ویژگی برتر، ویژگی‌های برتر ۲ تا ۴ از لیست ۱۰ ویژگی برتر، و...)، با استفاده از مدل در نظر گرفته شده پیش‌بینی کرده، و سپس هربار یکی از ۸ مجموعه برچسب پیش‌بینی شده با برچسب‌های واقعی به عنوان به عنوان ورودی تابع `measure` در نظر گرفته شده و خروجی اول این تابع یعنی `specificity` در یک لیست ذخیره می‌شود. نهایتاً این لیست که به ترتیب اعضای آن `specificity` محاسبه شده برای در نظر گرفتن ۸ زیرمجموعه ۳ تایی ۱۰ ویژگی برتر (مقدار اول مربوط به در نظر گرفتن ویژگی‌های ۱ تا ۳، مقدار دوم مربوط به در نظر گرفتن ویژگی‌های ۲ تا ۴، و...) یک روش انتخاب ویژگی مشخص هستند، خروجی تابع `cal_specificity` خواهد بود.

خروجی تابع `cal_specificity` برای حالت‌های مختلف در نظر گرفتن دسته‌بند (SVM با کرنل RBF و Random Forest) در متغیرهایی با نام به شکل `Specificity_Classifier_FeatureSelectionMethod` (برای مثال `Specificity_RF_IG`) ذخیره شده است. سپس از تابع `specificity_plot` برای رسم نمودارهای مربوط به استفاده از دو دسته‌بند برای یک روش انتخاب ویژگی مشخص در یک نمودار استفاده می‌شود. (برای مثال `Specificity_RF_IG` و `Specificity_SVM_IG` را به عنوان ورودی گرفته و دو منحنی `specificity` نسبت به شماره زیرمجموعه را در یک نمودار رسم می‌کند).

## تفاوت‌های پیاده‌سازی انجام‌شده با کار مقاله

- مهم‌ترین تفاوت پیاده‌سازی انجام شده نسبت به کار مقاله، استفاده از تعداد محدودی از داده‌ها برای انتخاب ویژگی و سپس استفاده از ویژگی‌های انتخاب شده براساس این داده‌ها، برای انجام ارزیابی روی داده‌هایی دیگر است. به نظر می‌رسد در کار اجرا شده توسط نویسندگان مقاله، از همه ۱۲۰۷ داده برای انتخاب ویژگی استفاده شده و سپس از ویژگی‌های انتخاب شده براساس این داده‌ها برای دسته‌بندی خودشان براساس `10 fold cross-validation` استفاده شده باشد. این کار به وضوح `bias` داشته و ارزیابی انجام شده را `unfair` می‌کند. (به خصوص در رسم نمودار `specificity` نسبت به شماره

زیرمجموعه‌های سه‌تایی، طبیعی است که هر چه سه ویژگی مهم‌تری را براساس داده‌هایی که قصد دسته‌بندی‌شان را داریم انتخاب کرده باشیم، به نتیجه بهتری روی همان داده‌ها می‌رسیم!

بنابراین برای اینکه در پیاده‌سازی انجام شده ارزیابی درستی از اهمیت miRNAهای مورد بررسی داشته باشیم، ابتدا ۲۰۷ داده از میان ۱۲۰۷ داده برای انتخاب ویژگی استفاده شده، و سپس مراحل ارزیابی مبتنی بر 10 fold cross-validation روی ۱۰۰۰ داده باقی‌مانده، با ویژگی‌های برتر انتخاب شده بر اساس این ۲۰۷ داده انجام شده است.

- علاوه بر روش‌های انتخاب ویژگی به کار رفته در مقاله، در پیاده‌سازی انجام شده، از روش‌های Ridge و Mutual Information نیز استفاده شده است.

- در جدول ۲ مقاله ارزیابی با استفاده از مدل‌های مختلف (SVM خطی، SVM با کرنل RBF و Random forest) براساس دسته‌بندی‌های مختلف برترین ویژگی‌های به دست آمده از سه روش انتخاب ویژگی مورد استفاده در مقاله قابل مشاهده است. ردیف اول بخش مربوط به هر یک از سه دسته‌بند استفاده شده نیز، عملکرد مدل را هنگام استفاده از تمامی miRNAهای غیرصفر (۱۶۰۴ مورد) نشان می‌دهد. در پیاده‌سازی انجام شده دو ردیف دیگر نیز به بخش مربوط به هر دسته‌بند اضافه شده است. یک ردیف مربوط به حالت استفاده از miRNAهای غیر صفر تایید نشده در کلینیک (۳۶-۱۶۰۴ مورد) به عنوان ویژگی‌ها، و ردیف دیگر مربوط به استفاده از تمامی ۳۶ miRNA تاییدشده در کلینیک به عنوان ویژگی می‌باشد.

بررسی عملکرد مدل‌ها در دسته‌بندی داده‌ها در گروه سرطانی و غیرسرطانی براساس miRNAهایی که پیش از این در کلینیک شواهدی برای اهمیت آن‌ها در ایجاد سرطان به دست نیامده است، از اهمیت فراوانی برخوردار است. در صورتی که مدل‌ها با استفاده از این ویژگی‌ها بتوانند عملکرد خوبی را نشان دهند، می‌توان به این موضوع پی برد که در میان مواردی که هنوز شواهد کلینیکی‌ای برایشان به دست نیامده است، اطلاعات ارزشمندی وجود دارد که درآینده می‌توانند در شناسایی سرطان به ما کمک کنند، و یا در صورت استفاده از روش‌های انتخاب ویژگی روی این دسته از miRNAها شاید بتوان به موارد مهمی رسید که در آینده بتوان به شکل تجربی نیز اثر آن‌ها در ایجاد سرطان را تایید کرد.

## نتیجه آزمایشات

### انتخاب ویژگی

نتیجه انتخاب ویژگی با استفاده از روش‌های انتخاب ویژگی مختلف به شکل زیر بوده است.

در ستون اول نام ۱۰ miRNA برتر به شکل مرتب شده قابل مشاهده است (از بالا به پایین اهمیت miRNAها کاهش پیدا می کند). در ستون دوم نیز امتیاز به دست آمده برای هر miRNA قابل مشاهده می باشد.

همچنین زیر لیست ۱۰ ویژگی به دست آمده از هر روش، تعداد miRNAهای مشترک لیست به دست آمده با لیست نتایج مقاله برای همان روش، قابل مشاهده است. (از تابع `numpy.intersect1` استفاده شده است. برای روش های Ridge و Mutual Information که در مقاله از آن ها استفاده نشده، مقایسه با لیست ویژگی های برتر به دست آمده از روش Chi2 مقاله انجام گردیده است).

باید در نظر داشت که به دلیل انتخاب ویژگی براساس تعداد محدودی از داده ها در پیاده سازی انجام شده، میان نتایج به دست آمده و نتایج مقاله تطابق کاملی دیده نمی شود.

### روش Information Gain

```
Information Gain:
      0      1
0   hsa-mir-21  0.131212
1   hsa-mir-10b 0.122451
2   hsa-mir-125b-1 0.120766
3   hsa-mir-125b-2 0.119103
4   hsa-mir-145  0.112653
5   hsa-mir-335  0.112653
6   hsa-let-7c   0.095003
7   hsa-mir-200c 0.091428
8   hsa-mir-206  0.079341
9   hsa-mir-155  0.064306

Intersection of our result and article is: 6
```

در مجموع ۶ ویژگی از میان ۱۰ ویژگی به دست آمده با لیست نتایج مقاله مشترک بوده است.

### روش Chi2

```
CHI2:
      0      1
0   hsa-mir-10b 6330413.864923
1   hsa-mir-21  2550588.624113
2   hsa-let-7c  229541.265906
3   hsa-mir-145 227837.826828
4   hsa-mir-205 224308.135053
5   hsa-mir-200c 97625.041442
6   hsa-mir-203a 78434.942676
7   hsa-mir-125b-2 50252.423414
8   hsa-mir-125b-1 49628.343037
9   hsa-mir-126  20534.655540

Intersection of our result and article is: 6
```

۶ ویژگی از میان ۱۰ ویژگی به دست آمده با استفاده از این روش نیز با لیست ویژگی های به دست آمده از روش Chi2 مقاله مشترک بوده اند. با مقایسه نتایج به دست آمده از روش Chi2 و روش IG می توان دید که ۸ مورد در این دو لیست مشترک بوده و ترتیب کلی miRNAها نیز دارای تشابه می باشد.



## روش Lasso

```
Lasso:
      0      1
0 hsa-mir-10b 1.600401
1 hsa-let-7c 1.211345
2 hsa-mir-21 1.199510
3 hsa-mir-335 0.930871
4 hsa-mir-145 0.541928
5 hsa-mir-206 0.229977
6 hsa-mir-205 0.106715
7 hsa-mir-27a 0.091864
8 hsa-mir-155 0.066880
9 hsa-mir-181a-1 0.049360

Intersection of our result and article is: 5
```

۵ ویژگی از میان ۱۰ ویژگی به دست آمده با استفاده از این روش با لیست ویژگی‌های به دست آمده از روش Lasso مقاله مشترک بوده است.

## روش Mutual Information

```
Mutual Info:
      0      1
0 hsa-let-7c 0.227136
1 hsa-mir-10b 0.206914
2 hsa-mir-21 0.202149
3 hsa-mir-145 0.176228
4 hsa-mir-125b-1 0.154451
5 hsa-mir-125b-2 0.150559
6 hsa-mir-335 0.141670
7 hsa-mir-200c 0.103017
8 hsa-mir-206 0.064673
9 hsa-mir-205 0.055993

Intersection of our result and article is: 6
```

۶ ویژگی از میان ۱۰ ویژگی به دست آمده با استفاده از این روش با لیست ویژگی‌های به دست آمده از روش Chi2 مقاله مشترک بوده است.

همچنین ۹ ویژگی به دست آمده از این روش با لیست به دست آمده از روش IG مشترک می‌باشد.

## روش Ridge

```
Ridge:
      0      1
0 hsa-mir-10b 1.051396
1 hsa-mir-335 1.020453
2 hsa-mir-21 0.973538
3 hsa-let-7c 0.830659
4 hsa-mir-145 0.739337
5 hsa-mir-205 0.598987
6 hsa-mir-125b-2 0.559932
7 hsa-mir-181a-1 0.539149
8 hsa-mir-125b-1 0.524103
9 hsa-mir-206 0.487635

Intersection of our result and article is: 6
```

۶ ویژگی از میان ۱۰ ویژگی به دست آمده با استفاده از این روش با لیست ویژگی‌های به دست آمده از روش Chi2 مقاله مشترک بوده است.

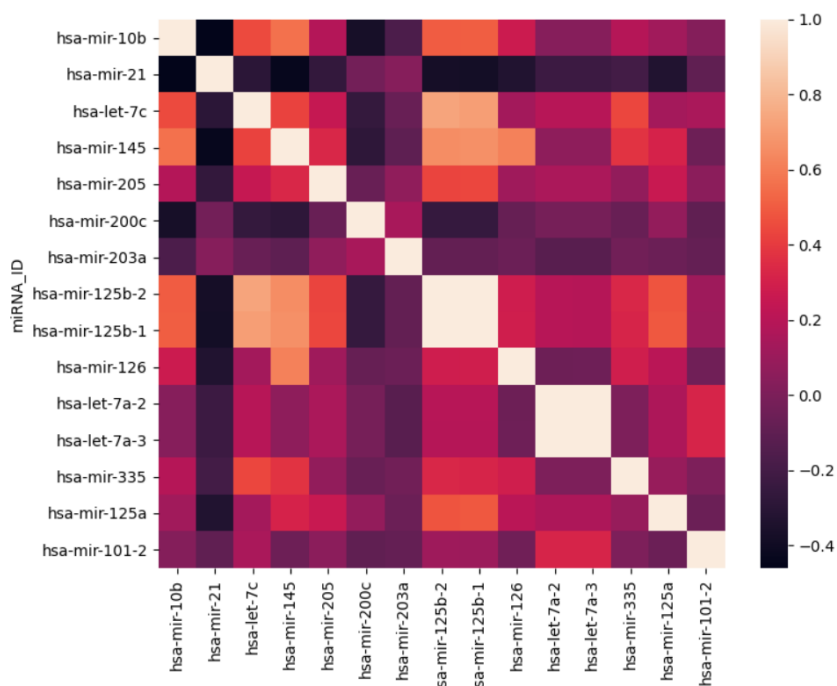
باز هم غالب ویژگی‌های به دست آمده از این روش با لیست به دست آمده از روش‌های قبلی مشترک بوده است.

درمجموع می‌توان دید که چند miRNA اول لیست ۱۰ miRNA برتر به دست آمده از تمامی روش‌ها تقریباً با یکدیگر مشترک بوده‌اند. miRNAهای hsa-mir-10b, hsa-let-7c, hsa-mir-145 و hsa-mir-21 مواردی بوده‌اند که در هر ۵ روش استفاده شده در میان ۵ مورد اول لیست قرار گرفته‌اند. ۳ مورد از این ۴ مورد، یعنی hsa-mir-10b, hsa-let-7c, hsa-mir-145 در لیست‌های Chi2 و IG مقاله نیز ۳ ویژگی برتر بوده‌اند. این موضوع نشان می‌دهد که با وجود استفاده از تعداد کمی از داده‌ها برای انتخاب ویژگی‌های برتر، ویژگی‌های برتر به خوبی انتخاب شده‌اند، و با وجود عدم تطابق کامل لیست‌های به دست آمده با لیست‌های مقاله، به هدف اصلی مقاله که شناسایی تعداد کمی miRNAهای پراهمیت در تشخیص نمونه‌های سرطانی از نمونه‌های سالم بوده است، رسیده‌ایم.

### بررسی correlation میان ویژگی‌های برتر انتخاب شده با یکدیگر

در ادامه برای بررسی اینکه ویژگی‌های برتر موجود در لیست مقاله که در لیست ما قرار نگرفته‌اند در لیست ویژگی‌های مرتب شده به دست آمده در چه رنگی قرار گرفته‌اند، تعداد بیشتر از ۱۰ مورد از ویژگی‌های برتر مرتب شده با استفاده از یکی از روش‌های انتخاب ویژگی، یعنی Chi2، را بررسی می‌کنیم.

همچنین به این دلیل که روش‌های انتخاب ویژگی استفاده شده روش‌های univariate بوده‌اند، امکان این وجود دارد که ویژگی‌های مهمی که با هم correlation بالایی دارند به طور همزمان در لیست ویژگی‌های برتر قرار بگیرند. حدس ما این بود که شاید نویسندگان مقاله ویژگی‌هایی که با هم correlation بالایی داشته‌اند را از لیست ۱۰ ویژگی برتر اولیه به دست آمده خود حذف کرده باشند، و این موضوع یکی از عوامل عدم تطابق کامل لیست‌های ما با لیست مقاله باشد (شاید هم بتوانیم با حذف این موارد به نتیجه‌ای شبیه‌تر به لیست مقاله برسیم). بنابراین ماتریس correlation داده‌ها در حالت در نظر گرفتن ۱۵ ویژگی برتر به دست آمده



از روش انتخاب ویژگی Chi2 را با استفاده از تابع `corr()` محاسبه کرده و سپس با استفاده از تابع `heatmap` از `seaborn` نمودار `heat map` آن را رسم می‌کنیم. در ستون سمت چپ `heat map` نام ۱۵ ویژگی برتر به دست آمده با روش انتخاب ویژگی Chi2 قابل مشاهده است. با مقایسه این موارد با لیست مقاله برای روش انتخاب ویژگی Chi2 (این لیست در قسمت مقایسه نتایج به دست آمده با نتایج مقاله قابل مشاهده می‌باشد)، می‌توان دید که ۴ موردی که در لیست مقاله وجود داشته اما در میان ۱۰ ویژگی برتر ما قرار نگرفته‌اند (hsa-let-7a-2, hsa-mir-125a, hsa-mir-335, hsa-let-7a-3)، در رنگ‌های ۱۱ تا ۱۴ قرار دارند.

در heat map رسم شده می‌توان دید که ردیف‌های مربوط به ۴ miRNA‌ی که در لیست مقاله قرار نداشته اما در لیست ما قرار گرفته‌اند (hsa-mir-21, hsa-mir-205, hsa-mir-200c, hsa-mir-203a)، برخلاف تصور قبلی، نسبت به دیگر miRNAها correlation کمی داشته‌اند. به خصوص در رابطه با hsa-mir-21, hsa-mir-200c, و hsa-mir-203a این موضوع کاملاً مشهود است و می‌توان دید که ردیف‌های مربوط به این ۳ مورد با رنگ‌های تیره کاملاً از دیگر ردیف‌ها متمایز شده‌اند. همچنین می‌توان دید ۴ miRNA، که هر ۴ مورد در لیست مقاله وجود داشته‌اند، به صورت دو به دو correlation بسیار بالایی با یکدیگر داشته‌اند (hsa-mir-125b-2 و hsa-mir-125b-2, hsa-let-7a-2 و hsa-let-7a-3) اما با این وجود این correlation بالا هر چهار مورد در لیست ۱۰ ویژگی برتر مقاله قرار گرفته‌اند.

بنابراین با توجه به مشاهدات انجام شده ممکن است نویسندگان مقاله نیز به مواردی مانند hsa-mir-21, hsa-mir-200c, و hsa-mir-203a رسیده باشند اما با منطقی کاملاً متفاوت از آنچه ما داشتیم، این موارد را حذف کرده و مواردی که با یکدیگر در correlation بالاتری بوده‌اند را حفظ کرده باشند.

درنهایت با توجه به نتایج به دست‌آمده هیچ یک از موارد از لیست ۱۰ ویژگی برتر حذف نشده و براساس همین ۱۰ ویژگی برتر به دست آمده پیاده‌سازی مراحل بعدی انجام شده است.

### نتایج ارزیابی با حالت‌های مختلف در نظر گرفتن ویژگی‌های برتر

***** Evaluation (Table 2) *****					
Classifier	Method	Accuracy	Sensitivity	Specificity	AUC
RF	Non Zero	0.988	0.9881081081081081	0.9866666666666667	0.9873873873873875
	Non Clinical Non Zero	0.985	0.9859611231101512	0.972972972972973	0.9794670480415621
	All Clinical	0.985	0.9880694143167028	0.9487179487179487	0.9683936815173256
	IG - 10	0.989	0.9923747276688453	0.9512195121951219	0.9717971199319836
	IG - 5	0.985	0.9912663755458515	0.9166666666666666	0.953966521106259
	IG - 3	0.986	0.9902067464635473	0.9382716049382716	0.9642391757009094
	CHI2 - 10	0.988	0.9912948857453754	0.9506172839506173	0.9709560848479963
	CHI2 - 5	0.988	0.9912948857453754	0.9506172839506173	0.9709560848479963
	CHI2 - 3	0.987	0.9912854030501089	0.9390243902439024	0.9651548966470056
	Lasso - 10	0.986	0.98914223669924	0.9493670886075949	0.9692546626534174
	Lasso - 5	0.988	0.9923664122137404	0.9397590361445783	0.9660627241791594
	Lasso - 3	0.987	0.99235807860262	0.9285714285714286	0.9604647535870243
SVM-RBF	Non Zero	0.991	0.9934640522875817	0.9634146341463414	0.9784393432169616
	Non Clinical Non Zero	0.99	0.9923830250272034	0.9629629629629629	0.9776729939950832
	All Clinical	0.988	0.9912948857453754	0.9506172839506173	0.9709560848479963
	IG - 10	0.989	0.9923747276688453	0.9512195121951219	0.9717971199319836
	IG - 5	0.987	0.9934354485776805	0.9186046511627907	0.9560200498702357
	IG - 3	0.988	0.9945235487404163	0.9195402298850575	0.9570318893127369
	CHI2 - 10	0.991	0.9945414847161572	0.9523809523809523	0.9734612185485548
	CHI2 - 5	0.99	0.994535519125683	0.9411764705882353	0.9678559948569592
	CHI2 - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169
	Lasso - 10	0.989	0.9923747276688453	0.9512195121951219	0.9717971199319836
	Lasso - 5	0.99	0.994535519125683	0.9411764705882353	0.9678559948569592
	Lasso - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169
SVM	Non Zero	0.993	0.9967177242888403	0.9534883720930233	0.9751030481909319
	Non Clinical Non Zero	0.993	0.9967177242888403	0.9534883720930233	0.9751030481909319
	All Clinical	0.987	0.9934354485776805	0.9186046511627907	0.9560200498702357
	IG - 10	0.989	0.9945295404814004	0.9302325581395349	0.9623810493104676
	IG - 5	0.989	0.9945295404814004	0.9302325581395349	0.9623810493104676
	IG - 3	0.989	0.9945295404814004	0.9302325581395349	0.9623810493104676
	CHI2 - 10	0.992	0.9956284153005465	0.9529411764705882	0.9742847958855673
	CHI2 - 5	0.992	0.9956284153005465	0.9529411764705882	0.9742847958855673
	CHI2 - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169
	Lasso - 10	0.988	0.9945235487404163	0.9195402298850575	0.9570318893127369
	Lasso - 5	0.991	0.9956236323851203	0.9418604651162791	0.9687420487506997
	Lasso - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169

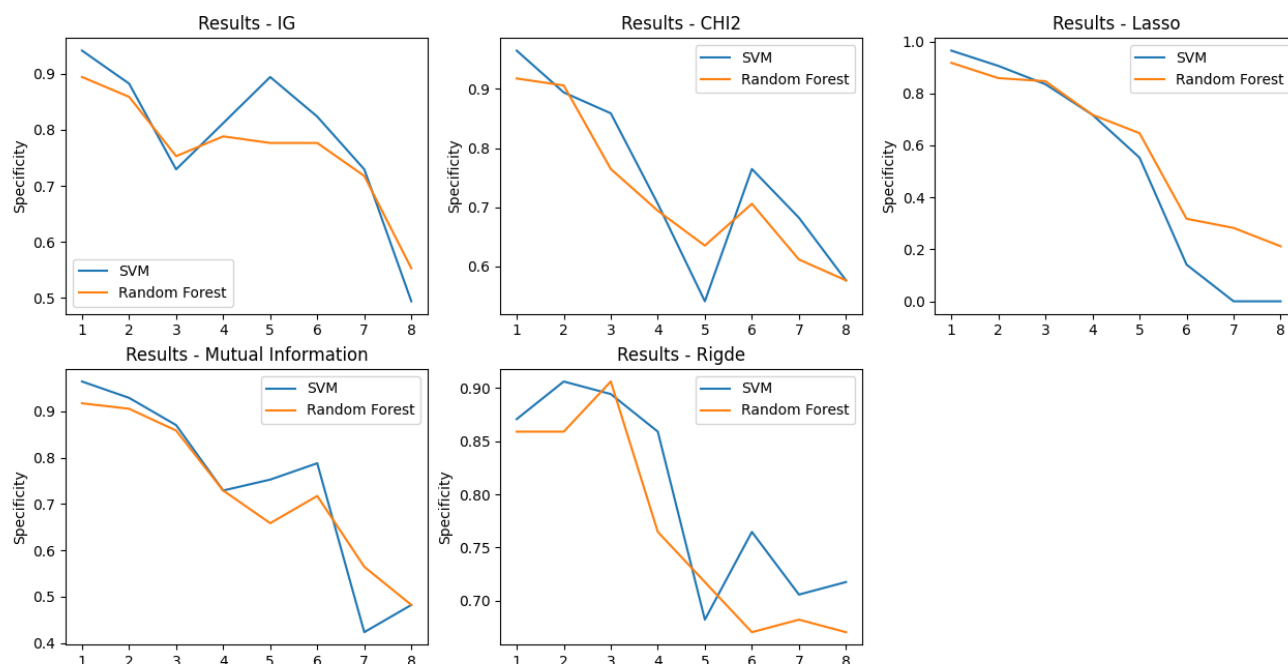
در جدول فوق نتایج به دست آمده از ارزیابی به شکل 10 fold cross validation براساس حالت های مختلف در نظر گرفتن ویژگی ها و استفاده از دسته بندی های Random Forest، SVM با کرنل RBF، و SVM خطی قابل مشاهده است. جدول براساس نوع classifier به کار رفته به سه بخش تقسیم می شود. در مقابل هر classifier ۱۰ حالت در نظر گرفتن ویژگی ها قابل مشاهده بوده و در چهار ستون بعدی در مقابل هر حالت در نظر گرفتن دسته بندی، و ویژگی ها، ۴ معیار به دست آمده از پیش بینی مدل، یعنی Accuracy، Sensitivity، Specificity و AUC قابل مشاهده می باشد. در این جدول علاوه بر حالت های مختلف در نظر گرفتن ۳، ۵، و ۱۰ ویژگی برتر به دست آمده از هر یک از روش های انتخاب ویژگی IG، Chi2، و Lasso، و حالت در نظر گرفتن تمامی ۱۶۰۴ miRNA که count غیر صفر (Non Zero) داشته اند (مقاله تنها این موارد را پوشش داده است)، حالت های در نظر گرفتن تمامی miRNA های غیر صفر غیر کلینیکال (Non Clinical Non Zero، ۱۵۶۸ مورد) و حالت در نظر گرفتن تمامی ۳۶ miRNA تایید شده در کلینیک (All Clinical) نیز بررسی شده است. (به این دلیل که در مقاله منبع تنها از این سه روش برای انتخاب ویژگی استفاده شده بود، نتایج مربوط به استفاده از ویژگی های به دست آمده از روش های Ridge و Mutual Information در این جدول قرار داده نشده اند).

در تمامی حالت ها Accuracy به دست آمده بالای ۰٫۹۸۵ بوده است. با مقایسه مقادیر موجود در ستون های مختلف Accuracy، Sensitivity، Specificity و AUC می توان دید که بیشترین تغییرات در ستون مربوط به Specificity دیده شده و به طور کلی مقادیر این معیار کمتر از معیارهای دیگر بوده است، که البته با توجه به imbalance بودن داده ها و تعداد بسیار کمتر نمونه های سالم این موضوع طبیعی می باشد. در مقایسه روش های مختلف انتخاب ویژگی می توان گفت که به طور کلی عملکرد مدل ها در حالت استفاده از ویژگی های به دست آمده از روش Chi2 نسبت به دو روش دیگر کمی بهتر بوده است. همچنین به طور کلی عملکرد مدل SVM با کرنل RBF کمی بهتر از دو مدل دیگر به نظر می رسد.

همچنین نزدیک بودن نتایج مربوط به حالت در نظر گرفتن ۳، ۵، و ۱۰ ویژگی برتر در هر یک از ترکیب های ممکن برای در نظر گرفتن classifier و روش انتخاب ویژگی، نشان می دهد که می توان براساس تعداد کمی از miRNA ها با دقت خوبی در رابطه با سرطانی بودن یا نبودن نمونه تصمیم گیری کرد.

نکته بسیار مهم عملکرد خوب مدل ها در حالت استفاده از ویژگی های غیر صفر غیر کلینیکال می باشد. در هر سه مدل استفاده شده نتایج ارزیابی برای حالت استفاده از این ۱۵۶۸ miRNA بسیار مناسب بوده، و حتی پیش بینی انجام شده Specificity بسیار بالایی داشته است (۰٫۹۷۳، ۰٫۹۶۳، و ۰٫۹۵۳۵). این موضوع (که در مقاله بررسی نشده است) نشان می دهد که miRNA های اندازه گیری شده ای که هنوز شواهد کلینیکالی برای اثبات اثرگذاری آن ها در ایجاد سرطان وجود ندارد، می توانند اطلاعات مفیدی برای شناسایی سرطان پستان در اختیارمان قرار دهند. در نتیجه با انجام بررسی های بیشتر ممکن است بتوانیم miRNA هایی را شناسایی کنیم که کاندیدهای مناسبی برای بررسی به صورت تجربی بوده و احتمال دارد بتوانیم اهمیت آن ها در ایجاد سرطان را در آزمایشگاه اثبات کنیم.

## نمودار Specificity نسبت به شماره زیرمجموعه‌های ۳ عضوی ۱۰ ویژگی برتر



با توجه به حساسیت مدل‌ها نسبت به معیار Specificity برای لیست ۱۰ ویژگی برتر به دست آمده از هر یک از ۵ روش انتخاب ویژگی، زیرمجموعه‌های سه‌تایی به شکل ویژگی‌های برتر ۱ تا ۳ (زیرمجموعه اول)، ویژگی‌های برتر ۲ تا ۴ (زیرمجموعه دوم)، ...، ویژگی‌های برتر ۸ تا ۱۰ (زیرمجموعه هشتم)، در نظر گرفته شده، مدل‌های SVM با کرنل RBF و Random Forest براساس این ویژگی‌ها روی ۱۰۰۰ داده validation پیش‌بینی انجام داده و برای هر مورد Specificity محاسبه شده است. نمودارهای Specificity نسبت به شماره زیرمجموعه برای ۵ روش انتخاب ویژگی رسم شده‌اند. مطابق انتظار با بالا رفتن شماره زیرمجموعه Specificity کاهش یافته و شکل کلی همه نمودارها نزولی بوده است، وجود این فرم نزولی نشان می‌دهد که miRNAهایی که اهمیت آن‌ها در ایجاد سرطان پستان در کلینیک تایید شده‌است، از نظر اهمیت با یکدیگر تفاوت دارند.

می‌توان دید که با وجود نزولی بودن فرم کلی نمودارها، در بعضی از زیرمجموعه‌ها Specificity به دست آمده نسبت به زیرمجموعه قبلی بیشتر بوده است (به خصوص در رابطه با دسته‌بند SVM با کرنل RBF این نوسانات شدیدتر بوده‌اند)، دلیل این موضوع این است که ویژگی‌های برتر به دست آمده از هر روش براساس تعداد محدودی از داده‌ها (۲۰۷ داده) انتخاب شده‌اند، اما ارزیابی روی ۱۰۰۰ داده دیگر انجام شده است، بنابراین ویژگی‌های برتر به دست آمده روی این تعداد محدود از داده‌ها لزوماً بهترین ویژگی‌های ممکن برای دسته‌بندی تمامی ۱۲۰۷ داده نبوده‌اند. اما درنهایت شروع شدن تمامی نمودارها از Specificity بالای ۹۰ درصد (به غیر از Ridge، با شروع از حدود ۸۵ درصد) و همچنین فرم نزولی نمودارها نشان می‌دهد که توانسته‌ایم براساس تعداد محدودی از داده‌ها miRNAهای مختلف را از نظر اهمیت در دسته‌بندی نمونه‌های سرطانی و سالم مرتب کرده و با استفاده از تعداد کم ۳ miRNA به مقدار مناسبی از معیار Specificity برسیم. این نتایج نشان می‌دهد که می‌توان با استفاده از تنها ۳ miRNA برتر به دست آمده از روش‌های انتخاب ویژگی در این مطالعه، با دقت خوبی نمونه‌های سرطانی و سالم را از هم تشخیص داد، و در نتیجه این موارد دارای این پتانسیل هستند که در آینده در تشخیص سرطان پستان به عنوان biomarker مورد استفاده قرارگیرند.

## مقایسه نتایج به دست آمده با نتایج مقاله

### نتایج انتخاب ویژگی

۱۰ ویژگی برتر به دست آمده از هر یک از روش های انتخاب ویژگی Chi2، Information Gain و Lasso به شکل زیر است.

Info Gain	CHI2	Lasso
hsa-mir-10b	hsa-mir-10b	hsa-let-7a-3
hsa-let-7c	hsa-let-7c	hsa-let-7c
hsa-mir-145	hsa-mir-145	hsa-let-7d
hsa-mir-125b-1	hsa-mir-125b-2	hsa-mir-101-1
hsa-mir-125b-2	hsa-mir-125b-1	hsa-mir-10b
hsa-mir-335	hsa-mir-335	hsa-mir-125b-2
hsa-mir-126	hsa-mir-126	hsa-mir-145
hsa-mir-125a	hsa-mir-125a	hsa-mir-206
hsa-let-7a-2	hsa-let-7a-2	hsa-mir-27b
hsa-let-7a-3	hsa-let-7a-3	hsa-mir-335

ویژگی های به دست آمده از روش های IG و Chi2 تنها نسبت به یکدیگر یک جابه جایی دارند. همچنین نتیجه به دست آمده از روش Lasso با نتیجه Chi2 و IG دارای ۶ اشتراک می باشد.

```
Information Gain:
0      1
0      hsa-mir-21 0.131212
1      hsa-mir-10b 0.122451
2      hsa-mir-125b-1 0.120766
3      hsa-mir-125b-2 0.119103
4      hsa-mir-145 0.112653
5      hsa-mir-335 0.112653
6      hsa-let-7c 0.095003
7      hsa-mir-200c 0.091428
8      hsa-mir-206 0.079341
9      hsa-mir-155 0.064306

Intersection of our result and article is: 6
```

```
CHI2:
0      1
0      hsa-mir-10b 6330413.864923
1      hsa-mir-21 2550588.624113
2      hsa-let-7c 229541.265906
3      hsa-mir-145 227837.826828
4      hsa-mir-205 224308.135053
5      hsa-mir-200c 97625.041442
6      hsa-mir-203a 78434.942676
7      hsa-mir-125b-2 50252.423414
8      hsa-mir-125b-1 49628.343037
9      hsa-mir-126 20534.655540

Intersection of our result and article is: 6
```

```
Lasso:
0      1
0      hsa-mir-10b 1.600401
1      hsa-let-7c 1.211345
2      hsa-mir-21 1.199510
3      hsa-mir-335 0.930871
4      hsa-mir-145 0.541928
5      hsa-mir-206 0.229977
6      hsa-mir-205 0.106715
7      hsa-mir-27a 0.091864
8      hsa-mir-155 0.066880
9      hsa-mir-181a-1 0.049360

Intersection of our result and article is: 5
```

```
Mutual Info:
0      1
0      hsa-let-7c 0.227136
1      hsa-mir-10b 0.206914
2      hsa-mir-21 0.202149
3      hsa-mir-145 0.176228
4      hsa-mir-125b-1 0.154451
5      hsa-mir-125b-2 0.150559
6      hsa-mir-335 0.141670
7      hsa-mir-200c 0.103017
8      hsa-mir-206 0.064673
9      hsa-mir-205 0.055993

Intersection of our result and article is: 6
```

```
Ridge:
0      1
0      hsa-mir-10b 1.051396
1      hsa-mir-335 1.020453
2      hsa-mir-21 0.973538
3      hsa-let-7c 0.830659
4      hsa-mir-145 0.739337
5      hsa-mir-205 0.598987
6      hsa-mir-125b-2 0.559932
7      hsa-mir-181a-1 0.539149
8      hsa-mir-125b-1 0.524103
9      hsa-mir-206 0.487635

Intersection of our result and article is: 6
```

تعداد ویژگی های مشترک به دست آمده از روش های IG، Chi2 و Lasso در پیاده سازی انجام شده با لیست های مقاله به ترتیب، ۶، ۶ و ۵ مورد می باشد (برای روش های Ridge و Mutual Information که در مقاله بررسی نشده اند این مقایسه با لیست Chi2 مقاله انجام شده و تعداد miRNA های مشترک برای هر دو روش ۶ مورد بوده است). در انتخاب ویژگی انجام شده در مقاله به نظر می رسد که ویژگی ها براساس تمامی مجموعه ۱۲۰۷ داده انتخاب شده و نهایتاً از ویژگی های به دست آمده برای ارزیابی همین داده ها



استفاده شده باشد. اما استفاده از بهترین ویژگی‌های به دست آمده از یک مجموعه داده برای دسته‌بندی خود این داده‌ها unfair بوده و می‌تواند اعتبار evaluation انجام شده را زیر سوال ببرد. بنابراین در پیاده‌سازی انجام شده ابتدا ۲۰۷ داده از میان ۱۲۰۷ مجموعه داده انتخاب شده، انتخاب ویژگی براساس این داده‌ها صورت گرفته، و سپس ارزیابی بر روی ۱۰۰۰ داده باقی‌مانده انجام شده است. دلیل عدم تطابق کامل نتایج به دست آمده با نتایج مقاله احتمالا انتخاب ویژگی‌ها بر مبنای تعداد محدودی از داده‌ها در پیاده‌سازی انجام شده می‌باشد.

اما در نهایت چند miRNA اول لیست ۱۰ miRNA برتر به دست آمده از تمامی روش‌ها تقریبا با یکدیگر مشترک بوده‌اند. miRNAهای hsa-mir-145، hsa-let-7c، hsa-mir-10b و hsa-mir-21 مواردی بوده‌اند که در هر ۵ روش استفاده شده در میان ۵ مورد اول لیست قرار گرفته‌اند. ۳ مورد از این ۴ مورد، یعنی hsa-mir-10b، hsa-let-7c، hsa-mir-145 در لیست‌های Chi2 و IG مقاله نیز ۳ ویژگی برتر بوده‌اند. این موضوع نشان می‌دهد که با وجود استفاده از تعداد کمی از داده‌ها برای انتخاب ویژگی‌های برتر، ویژگی‌های برتر به خوبی انتخاب شده‌اند، و با وجود عدم تطابق کامل لیست‌های به دست آمده با لیست‌های مقاله، به هدف اصلی مقاله که شناسایی تعداد کمی miRNAهای پراهمیت در تشخیص نمونه‌های سرطانی از نمونه‌های سالم بوده است، رسیده‌ایم.

### نتایج ارزیابی با حالت‌های مختلف در نظر گرفتن ویژگی‌های برتر

Classifier	Method	Accuracy	Sensitivity	Specificity	AUC
RF		0.996	1.000	0.952	0.999
	IG-10	0.995	0.998	0.962	0.996
	IG-5	0.996	0.997	0.977	0.998
	IG-3	0.997	0.997	0.990	0.999
	CHI2-10	0.995	0.999	0.952	0.995
	CHI2-5	0.996	0.999	0.979	0.996
	CHI2-3	0.996	0.997	0.981	0.999
	LASS-10	0.996	0.998	0.971	0.997
	LASS-5	0.995	0.997	0.965	0.998
	LASS-3	0.994	0.997	0.962	0.999
SVM-RBF		0.989	1.000	0.875	0.938
	IG-10	0.994	0.998	0.952	0.995
	IG-5	0.996	1.000	0.990	0.985
	IG-3	0.998	0.998	0.990	0.980
	CHI2-10	0.994	0.999	0.951	0.995
	CHI2-5	0.996	0.998	0.983	0.993
	CHI2-3	0.998	0.999	0.990	0.980
	LASS-10	0.995	0.998	0.962	0.996
	LASS-5	0.995	0.999	0.974	0.985
	LASS-3	0.996	0.999	0.962	0.980
SVM		0.997	0.999	0.971	0.985
	IG-10	0.997	0.999	0.971	0.997
	IG-5	0.997	0.999	0.985	0.989
	IG-3	0.998	0.999	0.990	0.981
	CHI2-10	0.997	0.999	0.971	0.997
	CHI2-5	0.996	1.000	0.988	0.987
	CHI2-3	0.998	0.999	0.990	0.991
	LASS-10	0.994	0.997	0.962	0.996
	LASS-5	0.995	0.999	0.956	0.993
	LASS-3	0.997	1.000	0.962	0.981

جدول دوی مقاله به شکل مقابل بوده است. همچنین در ابتدای صفحه بعد نتایج پیاده‌سازی انجام شده قابل مشاهده است.

در هر دو جدول Accuracyهای به دست آمده در همه حالت‌ها بالا بوده و بیشترین تغییرات در حالت‌های مختلف در نظر گرفتن classifier و روش انتخاب ویژگی مربوط به معیار Specificity بوده است. بنابراین با توجه به بالانس نبودن تعداد داده‌های دو کلاس، معیار Accuracy نمی‌تواند در تصمیم‌گیری موثر واقع شود و بنابراین باید بیشتر براساس معیار Specificity حالت‌های مختلف در نظر گرفتن classifier، روش انتخاب ویژگی و تعداد ویژگی‌های برتر انتخاب شده را مقایسه کرد.

***** Evaluation (Table 2) *****					
Classifier	Method	Accuracy	Sensitivity	Specificity	AUC
RF	Non Zero	0.988	0.9881081081081081	0.9866666666666667	0.9873873873873875
	Non Clinical Non Zero	0.985	0.9859611231101512	0.972972972972973	0.9794670480415621
	All Clinical	0.985	0.9880694143167028	0.9487179487179487	0.9683936815173256
	IG - 10	0.989	0.9923747276688453	0.9512195121951219	0.9717971199319836
	IG - 5	0.985	0.9912663755458515	0.9166666666666666	0.953966521106259
	IG - 3	0.986	0.9902067464635473	0.9382716049382716	0.9642391757009094
	CHI2 - 10	0.988	0.9912948857453754	0.9506172839506173	0.9709560848479963
	CHI2 - 5	0.988	0.9912948857453754	0.9506172839506173	0.9709560848479963
	CHI2 - 3	0.987	0.9912854030501889	0.9390243902439024	0.9651548966470056
	Lasso - 10	0.986	0.98914223669924	0.9493670886075949	0.9692546626534174
	Lasso - 5	0.988	0.9923664122137404	0.9397590361445783	0.9660627241791594
	Lasso - 3	0.987	0.99235807860262	0.9285714285714286	0.9604647535870243
SVM-RBF	Non Zero	0.991	0.9934640522875817	0.9634146341463414	0.9784393432169616
	Non Clinical Non Zero	0.99	0.9923830250272034	0.9629629629629629	0.9776729939958932
	All Clinical	0.988	0.9912948857453754	0.9506172839506173	0.9709560848479963
	IG - 10	0.989	0.9923747276688453	0.9512195121951219	0.9717971199319836
	IG - 5	0.987	0.9934354485776805	0.9186046511627907	0.9560200498702357
	IG - 3	0.988	0.9945235487404163	0.9195402298850575	0.9570318893127369
	CHI2 - 10	0.991	0.9945414847161572	0.9523809523809523	0.9734612185485548
	CHI2 - 5	0.99	0.994535519125683	0.9411764705882353	0.9678559948569592
	CHI2 - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169
	Lasso - 10	0.989	0.9923747276688453	0.9512195121951219	0.9717971199319836
	Lasso - 5	0.99	0.994535519125683	0.9411764705882353	0.9678559948569592
	Lasso - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169
SVM	Non Zero	0.993	0.9967177242888403	0.9534883720930233	0.9751030481909319
	Non Clinical Non Zero	0.993	0.9967177242888403	0.9534883720930233	0.9751030481909319
	All Clinical	0.987	0.9934354485776805	0.9186046511627907	0.9560200498702357
	IG - 10	0.989	0.9945295404814004	0.9302325581395349	0.9623810493104676
	IG - 5	0.989	0.9945295404814004	0.9302325581395349	0.9623810493104676
	IG - 3	0.989	0.9945295404814004	0.9302325581395349	0.9623810493104676
	CHI2 - 10	0.992	0.9956284153005465	0.9529411764705882	0.9742847958855673
	CHI2 - 5	0.992	0.9956284153005465	0.9529411764705882	0.9742847958855673
	CHI2 - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169
	Lasso - 10	0.988	0.9945235487404163	0.9195402298850575	0.9570318893127369
	Lasso - 5	0.991	0.9956236323851203	0.9418604651162791	0.9687420487506997
	Lasso - 3	0.992	0.9967141292442497	0.9425287356321839	0.9696214324382169

به طور کلی با وجود نزدیک بودن نتایج به یک دیگر، نتایج به دست آمده توسط مقاله کمی بهتر از نتایج به دست آمده در پیاده سازی ما بوده است (در order نیم درصد تفاوت دیده می شود)، که این تفاوت می تواند ناشی از این باشد که مقاله از ویژگی های برتر به دست آمده از تمامی داده ها برای ارزیابی آن ها استفاده کرده است، اما انتخاب ویژگی در پیاده سازی ما بر مبنای تعداد محدود و مجموعه مجزایی از داده ها بوده است. همچنین در نتایج به دست آمده در پیاده سازی انجام شده عملکرد مدل SVM با کرنل RBF کمی بهتر از دیگر مدل ها به نظر می رسد، اما این موضوع در نتایج به شکل مشهودی دیده نمی شود.

نتایج مربوط به حالت در نظر گرفتن ۳، ۵، و ۱۰ ویژگی برتر در هر یک از ترکیب های ممکن برای در نظر گرفتن classifier و روش انتخاب ویژگی، هم در نتایج مقاله و هم در نتایج ما نزدیک به هم بوده اند. این موضوع نشان می دهد که می توان بر اساس تعداد کمی از miRNA ها با دقت خوبی در رابطه با سرطانی بودن یا نبودن نمونه تصمیم گیری کرد.

در پیاده سازی انجام شده حالت های در نظر گرفتن تمامی miRNA های غیر صفر غیر کلینیکی (Non Clinical Non Zero)، ۱۵۶۸ مورد) و حالت در نظر گرفتن تمامی ۳۶ miRNA تایید شده در کلینیک (All Clinical) نیز علاوه بر حالت های در نظر گرفته شده در مقاله مورد بررسی قرار گرفته اند. نکته بسیار مهم عملکرد خوب مدل ها در حالت استفاده از ویژگی های غیر صفر غیر کلینیکی می باشد. در هر سه مدل استفاده شده نتایج ارزیابی برای حالت استفاده از این ۱۵۶۸ miRNA بسیار مناسب بوده، و حتی پیش بینی انجام شده Specificity بسیار بالایی داشته است (۰،۹۷۳، ۰،۹۶۳، و ۰،۹۵۳۵). این موضوع نشان می دهد که miRNA های



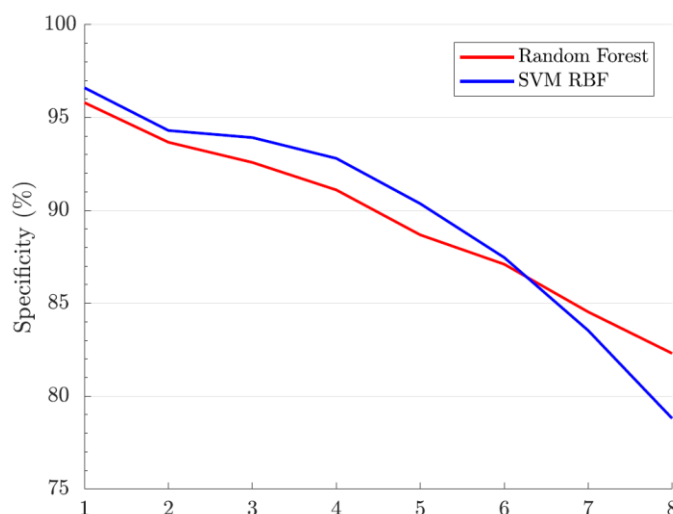
اندازه‌گیری‌ای شده‌ای که هنوز شواهد کلینیکالی برای اثبات اثرگذاری آن‌ها در ایجاد سرطان وجود ندارد، می‌توانند اطلاعات مفیدی برای شناسایی سرطان در اختیارمان قرار دهند. در نتیجه با انجام بررسی‌های بیشتر ممکن است بتوانیم miRNAهایی را شناسایی کنیم که کاندیدهای مناسبی برای بررسی به صورت تجربی بوده و احتمال دارد بتوانیم اهمیت آن‌ها در ایجاد سرطان را در آزمایشگاه اثبات کنیم.

### نمودار Specificity نسبت به شماره زیرمجموعه‌های ۳ عضوی ۱۰ ویژگی برتر

در مقاله منبع زیرمجموعه‌های ۳ عضوی براساس ویژگی‌های برتر به دست‌آمده از روش‌های Chi2 و IG به شکل زیر تعریف شده‌اند.

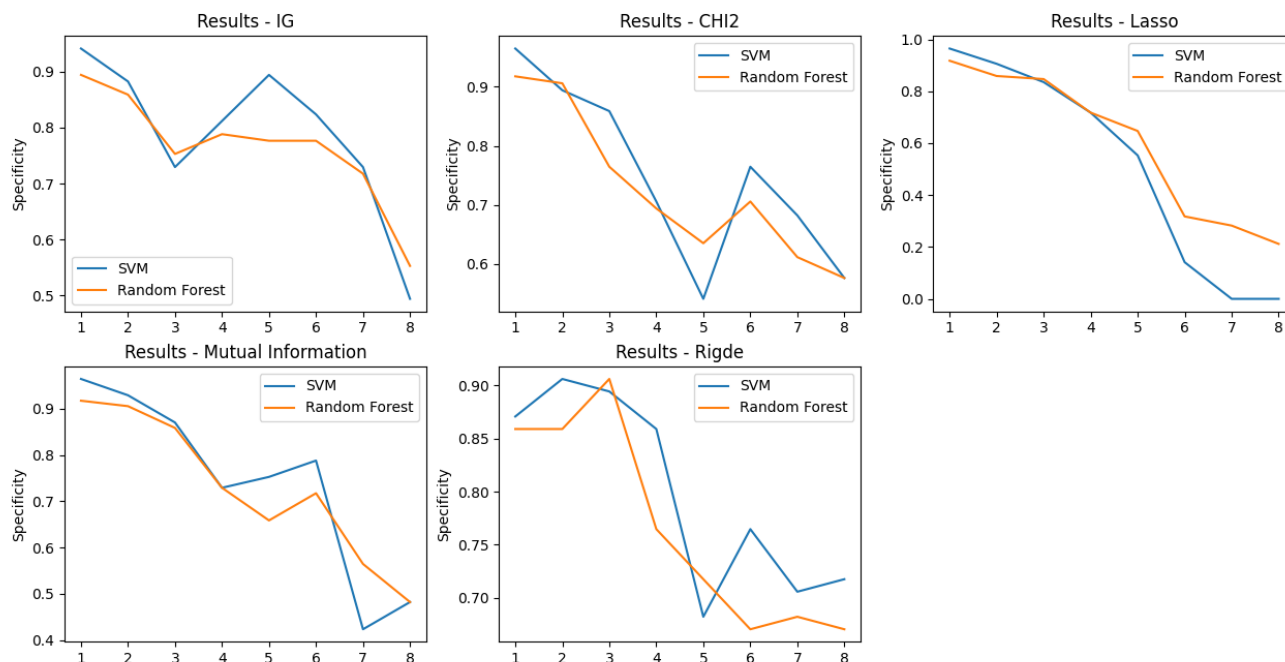
Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8
hsa-mir-10b hsa-let-7c hsa-mir-145	hsa-let-7c hsa-mir-145 hsa-mir-125b-1	hsa-mir-145 hsa-mir-125b-1 hsa-mir-125b-2	hsa-mir-125b-1 hsa-mir-125b-2 hsa-mir-335	hsa-mir-125b-2 hsa-mir-335 hsa-mir-126	hsa-mir-335 hsa-mir-126 hsa-mir-125a	hsa-mir-126 hsa-mir-125a hsa-let-7a-2	hsa-mir-125a hsa-let-7a-2 hsa-let-7a-3

نمودار Specificity نسبت به شماره زیرمجموعه در نظر گرفته‌شده به عنوان ویژگی به شکل زیر بوده است.



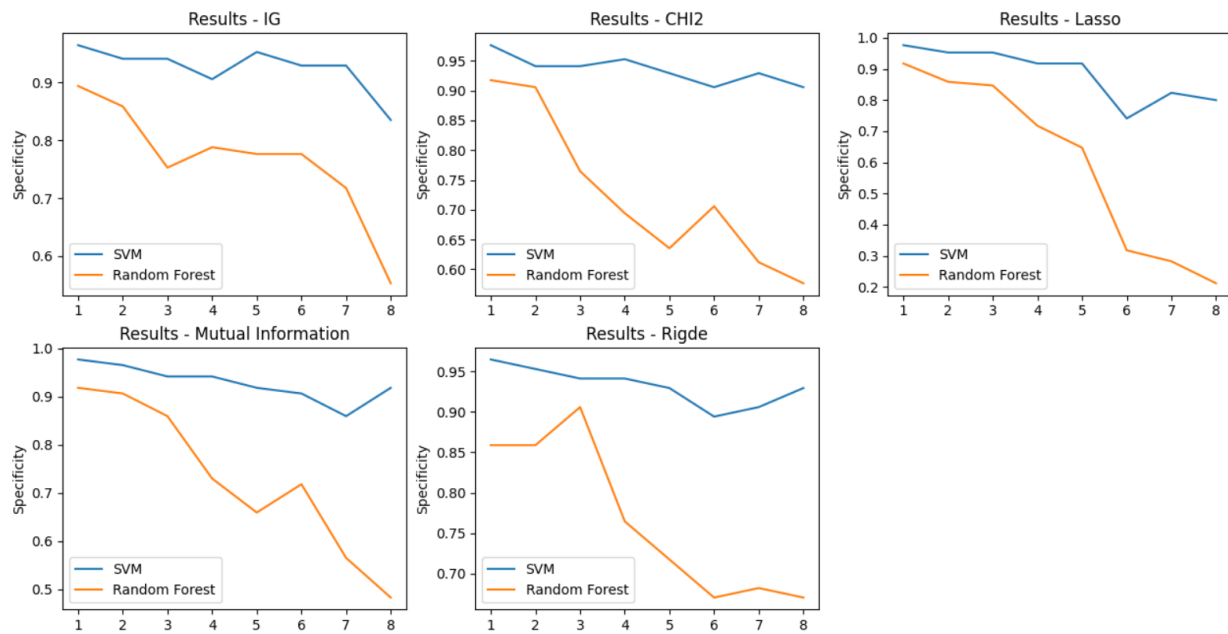
Specificity به دست‌آمده در حالت استفاده از ۳ ویژگی برتر در حالت استفاده از هر دو دسته‌بند SVM با کرنل RBF و Random Forest در حدود ۹۵ درصد بوده است. همچنین با بزرگ شدن شماره زیرمجموعه انتخاب شده به عنوان ویژگی Specificity کاهش می‌یابد. وجود این فرم نزولی نشان می‌دهد که miRNAهایی که اهمیت آن‌ها در ایجاد سرطان پستان در کلینیک تایید شده‌است، از نظر اهمیت با یکدیگر تفاوت دارند.

در ابتدای صفحه بعد نتایج به دست‌آمده در پیاده‌سازی انجام شده قابل مشاهده است. با توجه به عدم تطابق کامل ویژگی‌های برتر به دست‌آمده از ۵ روش انتخاب ویژگی بررسی شده، براساس لیست ۱۰ ویژگی برتر به دست‌آمده از هر یک از این روش‌ها به طور مجزا ۸ زیرمجموعه سه‌تایی در نظر گرفته شده و نمودارهای Specificity نسبت به شماره زیرمجموعه‌های سه‌تایی برای هر روش انتخاب ویژگی به شکل جداگانه رسم شده است.



در نتایج به دست آمده از پیاده سازی انجام شده، فرم نزولی کلی در تمامی نمودارها قابل مشاهده است. با وجود نزولی بودن فرم کلی نمودارها، در بعضی از زیر مجموعه ها **Specificity** به دست آمده نسبت به زیرمجموعه قبلی بیشتر بوده (به خصوص در رابطه با دسته بند **SVM** با کرنل **RBF** این نوسانات شدیدتر بوده اند)، و نتایج به دست آمده به تمیزی نتایج مقاله نمی باشد. دلیل این موضوع این است که در پیاده سازی انجام شده ویژگی های برتر به دست آمده از هر روش براساس تعداد محدودی از داده ها (۲۰۷ داده) انتخاب شده اند، اما ارزیابی روی ۱۰۰۰ داده دیگر انجام شده است، بنابراین ویژگی های برتر به دست آمده روی این تعداد محدود از داده ها لزوماً بهترین ویژگی های ممکن برای دسته بندی تمامی ۱۲۰۷ داده نبوده اند. اما به نظر می رسد که مقاله از بهترین ویژگی های انتخاب شده بر مبنای ۱۲۰۷ داده برای ارزیابی همین داده ها استفاده کرده باشد. در نهایت اما شروع شدن تمامی نمودارها از **Specificity** بالای ۹۰ درصد (به غیر از **Ridge**، با شروع از حدود ۸۵ درصد) و همچنین فرم نزولی نمودارها نشان می دهد که توانسته ایم براساس تعداد محدودی از داده ها **miRNA** های مختلف را از نظر اهمیت در دسته بندی نمونه های سرطانی و سالم مرتب کرده و با استفاده از تعداد کم ۳ **miRNA** به مقدار مناسبی از معیار **Specificity** برسیم. این نتایج نشان می دهد که می توان با استفاده از تنها ۳ **miRNA** برتر به دست آمده از روش های انتخاب ویژگی در این مطالعه، با دقت خوبی نمونه های سرطانی و سالم را از هم تشخیص داد، و در نتیجه این موارد دارای این پتانسیل هستند که در آینده در تشخیص سرطان پستان به عنوان **biomarker** مورد استفاده قرار گیرند.

یکی دیگر از عوامل موثر در دیده شدن نوسان نمودار و افت شدیدتر نمودارهای مربوط به نتایج ما، می تواند تنظیم پارامترهای **classifier** باشد. برای مثال در صورت در نظر گرفتن **class\_weight='balanced'** (در این حالت نسبت برچسب های دو کلاس در تصمیم گیری دسته بند اثر داده می شود). در تابع **SVC** نمودارها به شکل زیر خواهند بود.



می‌توان دید که عملکرد مربوط به دسته‌بند SVM با کرنل RBF در تمامی نمودارها به شکل چشمگیری بهبود یافته است، اما همچنان شکل نزولی نمودارها حفظ شده و می‌توان کاهش معیار **Specificity** با دسته‌بندی براساس ویژگی‌های کم‌اهمیت‌تر را مشاهده کرد.