تعریف نشانگرهای پایا برای بیماری های پیچیده با استفاده از شبکه های برهمکنش پروتئین

محمد حاتمی[۱]، مسعود رهگذر[۱] ، کاوه کاوسی[۲]

[۱]گروه تحقیقاتی پایگاه داده، قطب علمی کنترل و پردازش هوشمند، دانشکده مهندسی برق و کامپیوتر، دانشگاه تهران،
rahgozar@ut.ac.ir ، mohammadhatami@ut.ac.ir

[۲] آزمایشگاه سیستم های زیستی پیچیده، مرکز تحقیقات بیوشیمی بیوفیزیک دانشگاه تهران ،kkavousi@ut.ac.ir

چکیده

در گذشته بسیاری از محققان به منظور تشخیص بیماری، تلاش می‌کردند با استفاده از داده‌های ریزآرایه[۱]، به یافتن ژن‌های دارای بیشترین میزان تغییر بیان ژنی[۲] بپردازند. اما با توجه به اینکه این ژن‌ها در مجموعه داده‌های دیگر پایا نیستند، یعنی نشانگر[۳] های بدست آمده برای یک مجموعه داده از یک نوع بیماری الزاما نشانگر های مناسبی برای مجموعه داده ای دیگر از همان نوع بیماری نیستند، امروزه روش‌های جدیدتری به وجود آمده است که در آن‌ها سعی می‌شود از اطلاعات موجود در شبکه‌های ژنی نیز استفاده شود.

در این پژوهش روشی مبتنی بر شبکه ارائه شده است که در آن با بهره‌گیری از داده‌های بیان ژنی[۴] و شبکه برهمکنش پروتئین-پروتئین[۵] به رتبه‌بندی ژن‌ها با استفاده از روشی مشابه روش قدم زنی تصادفی[۶] پرداخته شده است. پس از آن با انتخاب ژن های با رتبه بالاتر، زیرشبکه‌هایی متشکل از هر یک از این چند ژن و ژن‌هایی که در شبکه به آن‌ها خیلی نزدیک و با آن‌ها هم‌بیان[۷] هستند، حاصل شده است و ویژگی‌هایی به منظور طبقه‌بندی از روی این زیرشبکه‌ها بدست آمده است، در پایان نشانگرهای پایا معرفی شده اند. الگوریتم ارائه شده بر روی داده های چهار نوع سرطان از نه مجموعه داده مختلف تست شده است و مشاهده شده که در روش ارائه شده ژن‌هایی (و زیرشبکه‌هایی) که برای تشخیص بیماری بکار گرفته می‌شوند به نسبت روش‌های پیشین می‌تواند نتایج بهبود یافته‌تر و پایاتری را ارائه کنند.

.

واژه‌های کلیدی

بیماری، شبکه برهمکنش پروتئین- پروتئین، بیان ژنی، زیرشبکه، نشان‌گر

---

[1] Microarray

[2] Most differential gene expression

[3] Marker

[4] Gene expression

[5] Protein-protein interaction network

[6] Random walker

[7] Co-expression

۱- مقدمه

در گذشته و حالا از امضای داده های بیان ژنی، به عنوان یک نشان‌گر یا هشداردهنده استفاده می‌شد به طوریکه بتوان از طریق آن روش مناسبی برای بهبود تشخیص بیماری و ساخت دارو برای اهداف خاص پیدا کرد. روشی که با آن درمان‌گران این توانایی را پیدا می‌کنند که مجموعه داده های ناهمگون بیماران را بر طبق روش‌های درمانی تقسیم‌بندی کنند و از این طریق به بیشترین بازدهی در تشخیص درست افراد بیمار از افراد سالم برسند[۱]. اما با وجود سال ها تلاش و تحقیق دشوار روی تعداد زیادی سرطان، همچنان دغدغه‌هایی روی نحوه عملکرد این نوع از نشان‌گرها وجود دارد[۲-۳]. همین امر فرصتی برای توسعه روش‌های مختلف را به وجود آورد تا به وسیله آن: ۱) به انتخاب ویژگی‌های مناسب مرتبط با درمان، از اطلاعات داده‌های ریزآرایه بپردازند، و ۲) از معیاری کمی مستخرج از آن ویژگی‌ها استفاده نمایند تا بتوانند مدلی را تعریف کنند که به کمک آن دو گروه افراد سالم و بیمار را متمایز کنند.

شناسایی یک امضای بیان ژنی می‌تواند به عنوان یک روال دو مرحله‌ای مورد توجه قرار بگیرد. در مرحله اول، ویژگی‌های حاوی اطلاعات شناسایی می‌شوند. برای مثال با رتبه‌بندی همه ویژگی‌های بالقوه که به کمک آن، ویژگی‌های دارای بهترین رتبه بتوانند بین دو گروه مورد مطالعه تمایز ایجاد کنند. در مرحله دوم فرایند ساخت امضاء، ویژگی‌های دارای برترین رتبه برای طبقه‌بندی[8] انتخاب می‌شوند. بیشتر روش‌هایی که تا کنون برای مدل کردن امضاء بیان ژنی مورد استفاده قرار گرفته‌اند را می‌توان در دو گروه عمده قرار داد[۴]. اولین گروه، روش‌های موسوم به تک- ژن[9] است، که در آن هیچ گونه اطلاعات خارجی برای آنالیز مسئله وجود ندارد. ویژگی‌های مورد استفاده در این روش‌ها ژن‌های اختصاصی هستند که به عنوان مثال با آنالیز اختلاف در بیان ژنی شناسایی شده‌اند. در مرحله طبقه‌بندی روش تک- ژن، طبقه‌بندی ساخته می‌شود که مقادیر بیان ژنی این ژن‌های حاوی اطلاعات را به عنوان ورودی گرفته، و نتیجه پیش‌بینی کلاس نمونه[10] را به عنوان خروجی برمی‌گرداند. گروه دوم از این روش‌ها، با نام روش مجموعه- ژن[11]، شامل گروه‌بندی ژن‌ها در مجموعه‌هایی می‌شود، که از آن‌ها بتوان به عنوان ویژگی برای طبقه‌بندی استفاده کرد. این ژن‌ها نوعا در مسیرها[12]، با هم همگروه هستند. در روش‌های مجموعه-ژن معمولا خود مجموعه ژن‌هایی که به عنوان تشخیص دهنده دو گروه (سالم و بیمار) مورد توجه قرار گرفته‌اند، به عنوان ویژگی طبقه‌بندی شناخته می‌شوند روش‌های مجموعه- ژن، در روال انتخاب ویژگی، اطلاعات یال‌های شبکه را دخالت نمی‌دهند، از این رو در سال‌های اخیر، روش‌های مبتنی بر شبکه[13] در آنالیز بیان ژنی، به خاطر ظرفیت بالا در بدست آوردن و توصیف خصوصیاتی مثل ماژولاریتی[14]، از محبوبیت بالایی برخوردار شده‌اند[۵]. یک شبکه، یا یک گراف، همانطور که در مفاهیم ریاضی شناخته می‌شود شامل مجموعه‌ای از گره‌ها و مجموعه‌ای از یال‌ها، بین گره‌هاست. مفهوم شبکه به آسانی می‌تواند به یک مفهوم بیولوژیکی تبدیل شود، به طوریکه ژن‌ها به عنوان گره‌ها و یال‌ها به عنوان رابطه بین ژن‌ها (تعاملات) شناخته شوند. در شبکه برهمکنش پروتئینی گره‌ها متناظر با ژن‌های کد شده با پروتئین(نشان‌دهنده اطلاعاتmRNA)، و وجود یک یال بین دو ژن نشان‌دهنده این است که پروتئین های کد شده برای آن ژن‌ها به صورت فیزیکی با یکدیگر برهمکنش دارند.

روش‌های زیادی وجود دارند که از مزایای مرتبط کردن شبکه تعاملی پروتئین با امضای بیان ژنی سود می‌برند[۵]. خود روش‌های شبکه‌ای را می‌توان به دو گروه تقسیم کرد. روش‌هایی مانند روش ارائه شده در [۶] که روی خصوصیات شبکه‌ای ژن‌های اختصاصی تمرکز می‌کنند. این ژن‌های منحصر به فرد به عنوان ویژگی برای طبقه‌بندی استفاده می‌شوند. این در حالیست که، به منظور انتخاب ویژگی برای تشخیص ژن‌های حاوی اطلاعات، از شبکه تعاملی پروتئین-پروتئین استفاده می‌گردد. این امر، در مقابل روش تک-ژن (بدون شبکه) که پیش‌تر معرفی شد، می‌باشد. زیرا علیرغم اینکه در آن روش هم ژن‌های منحصر به فرد به عنوان ویژگی طبقه‌بندی استفاده می‌گردند، اما هیچ بهره‌ای از اطلاعات موجود در شبکه برده نمی‌شود. دسته دوم روش‌های مبتنی بر شبکه، روی زیرشبکه‌های حاوی اطلاعات (به جای ژن‌های منحصر به فرد) تمرکز می‌کنند. این روش‌ها تا کنون به طور گسترده‌ای مورد توجه قرار گرفته‌اند[۷-۸]. این روش شامل ادغام شبکه‌های برهمکنش پروتئین با اطلاعات بیان ژنی می‌باشند که با این کار به شناسایی رفتار شبکه‌های اختلافی[15] مانند تغییر در ارتباط بین پروفایل‌های بیان ژنی جفت ژن‌های تعامل‌کننده، می‌پردازند. در این روش، فرایند انتخاب ویژگی شامل تشخیص زیرشبکه‌های حاوی اطلاعات می‌شود. زیرشبکه‌هایی که ساختار آن‌ها بین دو گروه مورد مطالعه(سالم و بیمار) تفاوت ایجاد می‌کند[۹-۱۰].

در این پژوهش ابتدا تلاش شده است که با در دست داشتن پایگاه داده های مختلف برهمکنش پروتئین- پروتئین و ترکیب و اصلاح آنها، شبکه برهمکنش پروتئینی مناسبتر و قابل اعتمادتری حاصل شود، سپس با توجه به داده های بیان ژنی، ژن‌هایی پیدا شود که حاوی اطلاعات مفیدی برای تشخیص بیماری هستند و با توجه به ساختار شبکه برهمکنش پروتئینی اصلاح شده، به رتبه بندی مجدد این ژن ها پرداخته شود و زیرگراف هایی با توجه به ژن های برتر بدست آمده، تشکیل داده شود. این زیرگراف‌ها مبنای اصلی کار در ساخت بردارهای ویژگی می‌باشند. با استفاده از بردار های ویژگی بدست آمده به ساخت طبقه بندی مناسب پرداخته شده و ژن های مؤثر و پایا در بیماری ها بدست می آیند.

---

[8] Classification

[9] Single-gene

[10] Sample

[11] Gene-set

[12] Pathway

[13] Subnetwork-based

[14] Modularity

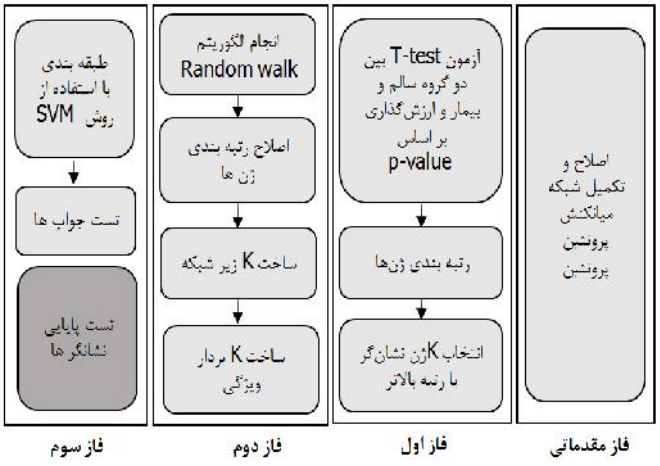
[15] Differential networks

۲- کارهای مرتبط

در این بخش مروری بر کارهای پیشین مبتنی بر شبکه ارائه می شود. اکثر روش‌های مبتنی بر شبکه سعی کرده اند با استفاده از آنالیزهای آماری از جمله خوشه‌بندی[۱۶]، روی تخمین اطلاعات شبکه از داده‌های بیان ژنی تمرکز کنند[۱۱]. روش آنالیز شبکه هم‌بیانی ژن‌های وزن‌دار[۱۷] (WGCNA) [۱۲] اطلاعات متناظر با بقا بیمار را با شبکه مشتق شده ترکیب کرده است. روش‌های دیگری نیز وجود دارند که بر مبنای شبکه انجام شده‌اند. از جمله این روش‌ها می‌توان به روشی که در [۱۳] شرح داده شده است اشاره کرد. این روش بر مبنای گام‌های تصادفی است و از اطلاعات موجود در اهمیت نسبی هر ژن بهره می‌برد[۱۱]. مثال دیگری از این روش‌ها الگوریتم CIPHER می‌باشد که توسط [۱٤] ارائه شده است. این الگوریتم ژن‌های بیماری را با مورد توجه قرار دادن بیماری‌های مشابه (از نظر فنوتیپی) و ایجاد یک لیست کامل از ارتباط بین فنوتیپ و ژن بیماری، تشخیص می‌دهد. مثال دیگر از این گونه روش‌ها، کاری است که در [۱۵] با هدف اندازه‌گیری اهمیت ژن‌ها به کمک بیشینه‌سازی تابع احتمال و تشخیص ژن‌هایی که بیشترین اتصال را دارند، انجام شده است. روش شبکه‌ای دیگری که روی تشخیص زیرشبکه‌های اختلافی تمرکز می‌کند، روشی است که توسط [۱٦] ارائه شده است. این روش مبتنی بر تجزیه طیفی[۱۸] پروفایل‌های بیان ژنی می‌باشد. به طور مشابه، [۱۷] یک روش مبتنی بر بی‌نظمی[۱۹] را، با تمرکز روی اندازه‌گیری تاثیر تصادفی بودن تک-ژن‌ها شرح داده است. در حالیکه [۱۸] روشی مبتنی بر بی‌نظمی با تمرکز بر روی آنالیز یک ماتریس آماری[۲۰] هسته حرارتی[۲۱]، ارائه داده است. از روش های دیگر تشخیص بیماری و ژن های بیمار می توان به روش ارائه شده در [۲۸] اشاره کرد. در این روش برای تشخیص ژن های بیمار از بررسی تغییرات ژن ها استفاده شده است.

۳- روش پیشنهادی

در این بخش رویکرد پیشنهادی برای حل مسئله تعریف نشانگرهای پایا برای بیماری های پیچیده با استفاده از شبکه های میانکنش پروتئین ارائه می‌شود. نشانگرهای پایا به مجموعه ژن هایی گویند که برای تمام مجموعه های داده ای از یک نوع بیماری، نشانگر های مؤثر و با دقت بالا برای تشخیص و دسته بندی آن مجموعه داده ای به افراد سالم و بیمار باشند. همانطور که در شکل ۱ مشاهده می کنید روش ارائه شده از ۴ فاز تشکیل شده است که در ادامه به توضیح آنها پرداخته می شود.



شکل ۱:مراحل روش پیش نهادی

- **فاز مقدماتی:** با توجه به وجود منابع متعدد برای بدست آوردن مجموعه داده های شبکه برهمکنش پروتئینی که تعدادی از آنها در بخش (۱-۴) معرفی شده است، استفاده از روشی برای یکپارچه سازی این مجموعه داده ها و بدست آوردن مجموعه ای قابل اطمینان بسیار حائز اهمیت است. از این رو در این پژوهش با استفاده از روش ارائه شده در[۱۹] به یکپارچه سازی این مجموعه ها پرداخته شده است.
- **فازاول**: در داده‌های بیان ژنی از مجموعه داده‌های آموزش موجود در جدول ۲ بین دو گروه سالم و بیمار یک آزمون t-test انجام شده است. سپس p-value های حاصل از این آزمون به ژن های مربوطه اختصاص یافته و بر اساس ارزش های بدست آمده ژن ها رتبه بندی شده اند[۲۰] و K ژن (۱۰ژن) با رتبه بالا برای ادامه ی کار انتخاب شده است.
- **فازدوم**: در این فاز الگوریتم قدم زنی تصادفی که توسط [۲۱-۲۲] ارائه شده است بر روی این ژن‌ها اجرا می شود و با استفاده از کرنل دیفیوژن[۲۲] معرفی شده در[۲۳-۲۴] ژن های بدست آمده، رتبه بندی مجدد می شوند. سپس به ازای k ژن بدست آمده با رتبه ی بالا، k زیر شبکه از ژن های حداکثر تا دو فاصله و هم بیان در گراف برهمکنش پروتئین با ژن مذکور، ساخته و به ازای هر زیر شبکه یک بردار ویژگی بدست می آید [۲۵].
- **فازسوم:** با روش ماشین بردار پشتیبانی[۲۳] (SVM) به انجام طبقه‌بندی نمونه‌ها پرداخته شده است. برای این کار k ویژگی در اختیار است که استفاده همزمان از همه آن‌ها لزوما موجب بدست آمدن پایین‌ترین درصد خطا نمی‌شود. بنابرین ترکیبی از ویژگی‌ها پیدا شده که بالاترین درصد صحت را در طبقه‌بندی می‌دهد. به این منظور روی این k ویژگی یک جستجوی کامل[۲۴] انجام شده است و ترکیب ویژگی‌هایی که بهترین نتایج را می‌دهند انتخاب شده است. در

[16] clustering
[17] weighted gene co-expression network analysis
[18] spectral decomposition
[19] Entropy-based
[20] Stochastic matrix
[21] Heat kernel
[22] Diffusion Kernel
[23] Support vector machine
[24] Exhaustive search

انتها به تست پایایی نتایج که در بخش (۳-۴) توضیح داده شده پرداخته می شود.

٤- آزمایش های انجام شده

در این بخش ابتدا مجموعه‌داده‌های مورد استفاده در تحلیل‌ها شامل شبکه‌های برهمکنش پروتئینی و داده های بیان ژنی معرفی شده است. سپس نتایج بدست آمده از الگوریتم ارائه شده برای تشخیص بیماری و نشانگرهای بیماری بیان شده و با نتایج حاصل از [۲۵] مقایسه و در نهایت به تست پایایی نتایج حاصل پرداخته شده است.

۴-۱- داده های مورد استفاده

مجموعه‌داده‌های مورد استفاده در این پژوهش شامل داده های بیان ژنی چهار نوع مختلف سرطان مطابق جدول۱ است که از پایگاه داده GEO[۲۵] جمع آوری شده اند. منابع متعددی برای بدست آوردن مجموعه داده های شبکه برهمکنش پروتئینی وجود دارد. در جدول۲ تعدادی از محبوب‌ترین این پایگاه‌های داده مشاهده می‌گردد.

سرطان	مجموعه داده	بیمار	نرمال	GEO ACC No
روده[۲۶]	Colon23	15	8	GSE4183
	Colon64	32	32	GSE8671
معده[۲۷]	Gastric24	12	12	GSE19826
	Gastric69	38	31	GSE13911
سینه[۲۸]	Breast58	31	27	GSE10810
	Breast185	42	143	GSE10780
ریه[۲۹]	Lung88	44	44	GSE18842
	Lung120	60	60	GSE19804
	Lung156	91	65	GSE19188

جدول ۱:مجموعه داده های بیان ژنی

نام پایگاه داده	نام پایگاه داده	نام پایگاه داده
MINT[30]	HPRD[31]	BIND[32]
DIP[33]	KEGG[34]	

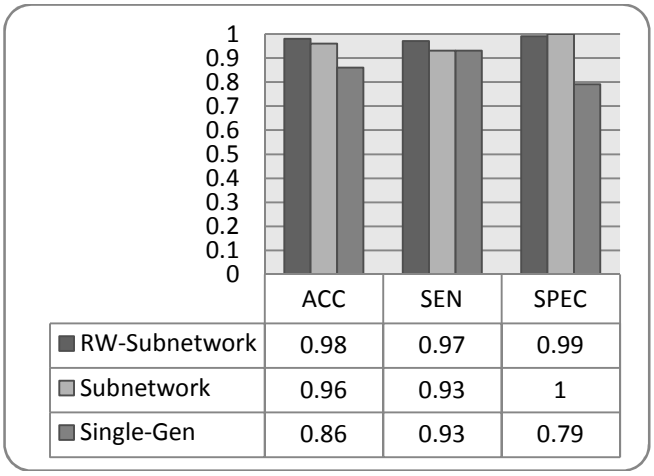
جدول 2:پایگاه داده های شبکه برهم کنش پروتئینی

[25] http://www.ncbi.nlm.nih.gov/geo
[26] Colon
[27] Gastric
[28] Breast
[29] Lung
[30] Molecular INTeraction database
[31] Human Protein Reference Database
[32] Biomolecular Interaction Network Database
[33] Database of Interacting Proteins
[34] Kyoto Encyclopedia of Genes and Genomes

۴-۲- نتایج بدست آمده از آزمایشات

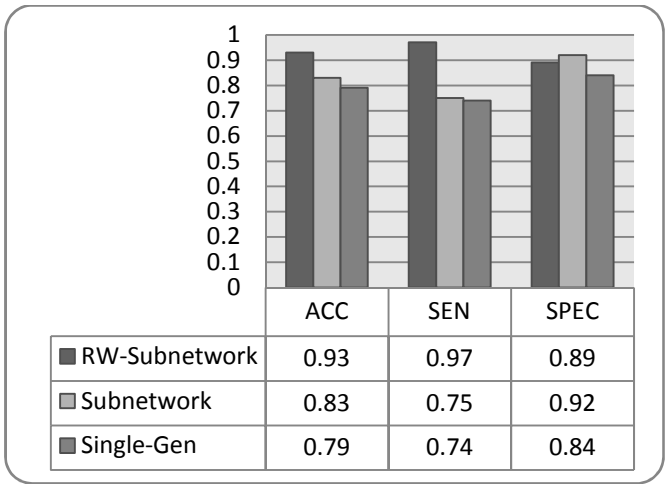
در این قسمت به مقایسه روش ارائه شده با دو روش موجود پیشین، که یکی از دسته روش‌های تک- ژن، و دیگری از گروه روش‌های مبتنی بر شبکه [۲۵] است، پرداخته شده است. در نمودار های این قسمت روش تک- ژن با Single-Gen، روش مبتنی بر شبکه با Subnetwork و روش ارائه شده در این مقاله با RW-Subnetwork نمایش داده شده است. برای سنجش کارایی الگوریتم ارائه شده، از سه معیار صحت[۳۵]، حساسیت[۳۶] و تشخیص[۳۷] [۲۶] استفاده شده است.

- صحت: تعداد پیش‌بینی‌های درست تقسیم بر تعداد کل نمونه‌ها
- حساسیت: تعداد پیش‌بینی‌های درست نمونه‌های مثبت (بیمار) تقسیم بر کل نمونه‌های مثبت
- تشخیص: تعداد پیش‌بینی‌های درست نمونه‌های منفی (سالم) تقسیم بر کل نمونه‌های منفی



	ACC	SEN	SPEC
RW-Subnetwork	0.98	0.97	0.99
Subnetwork	0.96	0.93	1
Single-Gen	0.86	0.93	0.79

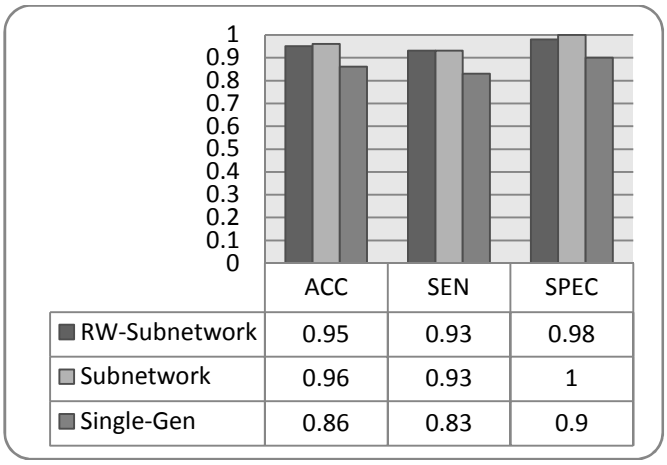
شکل 2:مقدار سه معیار ارزیابی برای سرطان ریه



	ACC	SEN	SPEC
RW-Subnetwork	0.93	0.97	0.89
Subnetwork	0.83	0.75	0.92
Single-Gen	0.79	0.74	0.84

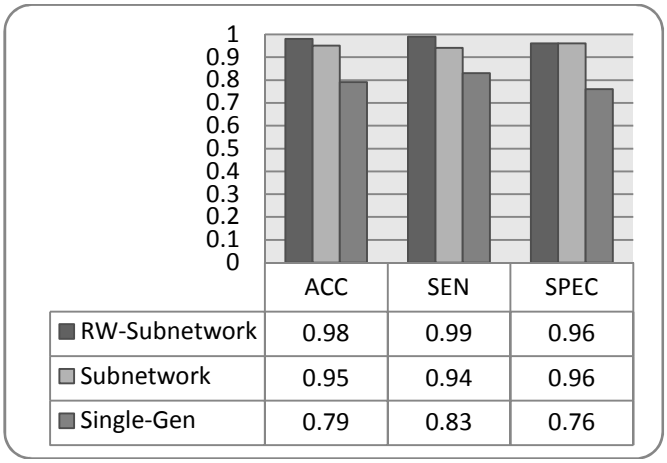
شکل ۳:مقدار سه معیار ارزیابی برای سرطان معده

[35] Accuracy(ACC)
[36] Sensitivity(SEN)
[37] Specificity(SPEC)



	ACC	SEN	SPEC
RW-Subnetwork	0.95	0.93	0.98
Subnetwork	0.96	0.93	1
Single-Gen	0.86	0.83	0.9

شکل ۴:مقدار سه معیار ارزیابی برای سرطان روده



	ACC	SEN	SPEC
RW-Subnetwork	0.98	0.99	0.96
Subnetwork	0.95	0.94	0.96
Single-Gen	0.79	0.83	0.76

شکل ۵:مقدار سه معیار ارزیابی برای سرطان سینه

در شکل های بالا نتاج بدست آمده برای ۴ نوع سرطان توسط سه روش مختلف نمایش داده شده است. همانطور که مشاهده می شود روش پیشنهادی به غیر از سرطان روده در بقیه موارد از بقیه روش ها بهتر عمل کرده است.

۳-۴- بررسی پایا بودن نشانگرها

با توجه به عدم پایایی نتایج در دو روش تک- ژن و مجموعه ژن، روش های شبکه ای به منظور حل این مشکل مورد توجه قرار گرفته اند. از این رو به منظور بررسی پایایی طبقه‌بندهای انجام شده، نتایج کار روی سرطان ریه، روی یک مجموعه داده دیگر (Lung88) تست شده است. به این منظور چهار ترکیب ویژگی که برترین نتایج را روی مجموعه داده Lung156 داشتند، تست شده و مشاهده شده که پیش‌بینی‌های خوبی حاصل شده است. ترکیب ژن‌های "SMIM5"، "SPOCK3"، "SPOCK2" و "NCKAP5" (و البته بقیه ژن‌های اطراف در زیرشبکه بدست آمده از طریق الگوریتم شبکه) در مجموعه داده Lung156 ٪۹۸ نمونه‌ها را به درستی پیش‌بینی کرده است. همان ترکیب ژن‌ها باعث٪۹۵.۸ پیش‌بینی درست در مجموعه داده Lung88 شده است.

یکی از موارد مؤثر در پایایی نتایج در نظر گرفتن کل شبکه برهمکنش پروتئینی در بدست آوردن k ژن اولیه است. با توجه به آنکه روش های شبکه ای پیشین فقط قسمتی از شبکه را مورد توجه قرار می دادند[۲۵]، نشانگر های بدست آمده از پایایی قابل قبولی برخوردار نیستند و یا ژن هایی وجود دارند که علی رغم مؤثر بودن در بیماری، در نظر گرفته نمی شوند، از این رو در روش ارائه شده در فاز دوم با استفاده از روش قدم زنی تصادفی کل شبکه برهم کنش پروتئینی جستجو شده و ژن های بدست آمده رتبه بندی مجدد می شوند. این امر علاوه بر اینکه در پایایی نتایج بسیار مؤثر است، با توجه به آنکه کل شبکه را مورد توجه قرار می دهد به معرفی نشانگر های جدید بیماری کمک می کند.

رتبه	ترکیب ژن‌ها	صحت پیش‌بینی
۱	NCKAP5, SPOCK2, SPOCK3,SMIM5	95.87%
۲	SPOCK2, SLC4A4, SPOCK3	94.87%
۳	SPOCK2, PLXNB3, SPOCK3	94.87%
۴	SPOCK2, SLC4A4, PLXNB3, SPOCK3	94.87%
۵	NCKAP5, SPOCK2, SLC4A4, SPOCK3	93.58%
۶	SLC4A4, PLXNB3, SPOCK3	93.58%
۷	PLXNB3, SPOCK3	92.94%
۸	NCKAP5, SPOCK2, PLXNB3, SPOCK3	92.30%
۹	SPOCK2, SLC4A1, SPOCK3	91.66%
۱۰	SLC4A1, SPOCK3	91.02%

جدول۳: ۱۰ ترکیب های برتر از ژن‌های پایا بطوریکه خودشان و زیرگراف بدست آمده از آن‌ها منجر به بالاترین میزان صحت در پیش‌بینی سرطان ریه شده‌اند.

۵- نتیجه گیری و پیشنهادات

در گذشته بسیاری از محققان به منظور تشخیص بیماری، تلاش می‌کردند با استفاده از داده‌های ریزآرایه، به یافتن ژن‌های دارای بیشترین میزان تغییر بیان ژنی در بیماران بپردازند. اما با توجه به اینکه این ژن‌ها در مجموعه پایگاه داده‌های دیگر مربوط به همان بیماری پایا نیستند، روش های مبتنی بر شبکه جایگاه خاصی یافته اند، از این رو در این پژوهش یک روش جدید مبتنی بر شبکه ارائه شده است. همانطور که در نمودارها نشان داده شد، روش ارائه شده از روش های پیشین بهتر عمل کرده است. در این کار در دو مرحله از اطلاعات موجود در شبکه استفاده شده است. بار اول با استفاده از شبکه برهمکنش پروتئینی به رتبه‌بندی ژن‌ها (با استفاده از

الگوریتم Random Walker) پرداخته شده و بار دوم با ساختن زیرگراف مربوط به ژن‌های برتر محاسبه بردارهای ویژگی برای طبقه‌بندی نمونه‌ها انجام شده است.

براساس پژوهش انجام شده و نتایج به‌دست آمده، موارد زیر برای ادامه کار و پژوهش‌های آتی پیشنهاد می‌گردند:

- شبکه‌ای که برای اجرای الگوریتم Random Walker مورد استفاده قرار گرفته است، می‌تواند با شبکه‌های دیگر جایگزین شود. برای مثال مقاله [۲۷] دو شبکه دیگر را نیز مورد بررسی قرار داده است: شبکه هستی‌شناسی ژن[۳۸] و شبکه ضریب همبستگی و شبکه ترکیبی.
- از دیگر موارد قابل بررسی، سایز نمونه‌های داده‌های آموزش می‌باشد. سایز نمونه‌های مورد استفاده می‌تواند نقش مهمی در پایایی نشان‌گرها در مجموعه داده‌های مختلف داشته باشد. لازم است بررسی گردد برای هر نوع بیماری برحسب پیچیدگی آن چه سایزی از نمونه‌ها می‌تواند به ساخت نشان‌گرهایی با قابلیت پایایی بالا ، کمک نماید

مراجع

[1] Weigelt, Britta, Frederick L. Baehner, and Jorge S. Reis-Filho. "The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade." *The Journal of pathology* 220.2 (2010): 263-280.

[2] Tímár, József, Balázs Győrffy, and Erzsébet Rásó. "Gene signature of the metastatic potential of cutaneous melanoma: too much for too little?."*Clinical & experimental metastasis* 27.6 (2010): 371-387.

[3] Harbeck, Nadia, et al. "Molecular and protein markers for clinical decision making in breast cancer: today and tomorrow." *Cancer treatment reviews*40.3 (2014): 434-444.

[4] Emmert-Streib, Frank, Shailesh Tripathi, and Ricardo de Matos Simoes. "Harnessing the complexity of gene expression data from cancer: from single gene to structural pathway methods." *Biology direct* 7.1 (2012): 1.

[5] Ideker, Trey, and Nevan J. Krogan. "Differential network biology." *Molecular systems biology* 8.1 (2012): 565.

[6] Winter, Christof, et al. "Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes." *PLoS Comput Biol* 8.5 (2012): e1002511.

[7] Taylor, Ian W., et al. "Dynamic modularity in protein interaction networks predicts breast cancer outcome." *Nature biotechnology* 27.2 (2009): 199-204.

[8] Schramm, Sarah-Jane, et al. "Disturbed protein–protein interaction networks in metastatic melanoma are associated with worse prognosis and increased functional mutation burden." *Pigment cell & melanoma research* 26.5 (2013): 708-722.

[9] Luo, Jiawei, and Shiyu Liang. "Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data." *Journal of biomedical informatics* 53 (2015): 229-236.

[10] Zhu, Jie, et al. "Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles." *BMC bioinformatics* 14.5 (2013): 1.

[11] Barter, Rebecca L., et al. "Network-based biomarkers enhance classical approaches to prognostic gene expression signatures." *BMC systems biology* 8.Suppl 4 (2014): S5.

[12] Zhang, Bin, and Steve Horvath. "A general framework for weighted gene co-expression network analysis." *Statistical applications in genetics and molecular biology* 4.1 (2005): 1128.

[13] Shi, Mingguang, R. Daniel Beauchamp, and Bing Zhang. "A network-based gene expression signature informs prognosis and treatment for colorectal cancer patients." *PloS one* 7.7 (2012): e41292.

[14] Wu, Xuebing, et al. "Network-based global inference of human disease genes." *Molecular systems biology* 4.1 (2008): 189.

[15] Ma, Shuangge, et al. "Incorporating gene co-expression network in identification of cancer prognosis markers." *BMC bioinformatics* 11.1 (2010): 1.

[16] Rapaport, Franck, et al. "Classification of microarray data using gene networks." *BMC bioinformatics* 8.1 (2007): 35.

[17] Teschendorff, Andrew E., and Simone Severini. "Increased entropy of signal transduction in the cancer metastasis phenotype." *BMC systems biology* 4.1 (2010): 1.

[18] Ideker, Trey, and Nevan J. Krogan. "Differential network biology." *Molecular systems biology* 8.1 (2012): 565.

[19] Deane, Charlotte M., et al. "Protein interactions two methods for assessment of the reliability of high throughput observations." *Molecular & Cellular Proteomics* 1.5 (2002): 349-356.

[20] Callow, Matthew J., et al. "Microarray expression profiling identifies genes with altered expression in HDL-deficient mice." *Genome research* 10.12 (2000): 2022-2029.

[21] Köhler, Sebastian, et al. "Walking the interactome for prioritization of candidate disease genes." *The American Journal of Human Genetics* 82.4 (2008): 949-958.

[22] Zhu, Jie, et al. "Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles." *BMC bioinformatics* 14.5 (2013): 1.

[23] Kondor, Risi Imre, and John Lafferty. "Diffusion kernels on graphs and other discrete input spaces." *ICML*. Vol. 2. 2002.

[24] Zhao, Zhi-Qin, et al. "Laplacian normalization and random walk on heterogeneous networks for disease-gene prioritization." *Computational biology and chemistry* 57 (2015): 21-28.

[25] Zhang, Lin, et al. "Extracting a few functionally reproducible biomarkers to build robust subnetwork-based classifiers for the diagnosis of cancer." *Gene*526.2 (2013): 232-238.

[26] Metz, Charles E. "Basic principles of ROC analysis." *Seminars in nuclear medicine*. Vol. 8. No. 4. WB Saunders, 1978.

[27] Morrison, Julie L., et al. "GeneRank: using search engine technology for the analysis of microarray experiments." *BMC bioinformatics* 6.1 (2005): 1.

[28] Zhou, Hongyi, and Jeffrey Skolnick. "A knowledge-based approach for predicting gene-disease associations." *Bioinformatics* (2016):btw358.

[38] Gene ontologies network