ژنتیک نوین
دوره شانزدهم، شماره ۱، تابستان ۱۴۰۰
صفحه ۱۱۲ - ۱۰۳

ارائه‌ی روشی جدید برای کشف نشانگرهای زیستی پیش آگاهی دهنده در سرطان ریه

Developing a novel algorithm to identify diagnostic biomarkers in lung cancer

مرتضی کوهسار[۱]، یوسف مسعودی سبحان زاده[۲]، علی مسعودی نژاد*[۳]

۱- محقق پسا دکتری، دانشگاه صنعتی شریف، تهران، ایران
۲- استادیار، مرکز تحقیقات ریزفناوری داروئی، پژوهشکده زیست پزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران
۳- دانشیار، سیستم بیولوژی و بیوانفورماتیک، دانشگاه تهران، تهران، ایران

Kouhsar M[1], Masoudi-Sobhanzadeh Y[2], Masoudi-Nejad A[*3]

1- Post Doc. Researcher, Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran
2- Assistant Professor, Research Center for Pharmaceutical Nanotechnology, Biomedicine Institute, Tabriz University of Medical Sciences, Tabriz, Iran
3- Associate Professor, Laboratory of Systems Biology and Bioinformatics (LBB), Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran

* نویسنده مسئول مکاتبات، پست الکترونیکی: amasoudin@ut.ac.ir

(تاریخ دریافت: ۹۹/۱۱/۲۱ - تاریخ پذیرش: ۰۰/۰۲/۱۹)

چکیده

امروزه رویکردهای یادگیری ماشین به‌طور گسترده‌ای در تجزیه و تحلیل داده‌های حجیم استفاده می‌شود. با توجه به فناوری جدید و تولید داده‌های با بازده بالا در زیست شناسی (مانند داده‌های تعیین توالی نسل جدید)، استفاده از روش یادگیری ماشین بر روی داده‌های بزرگ بیولوژیکی می‌تواند به درک مکانیسم بیماری پیچیده مانند سرطان کمک کند. استخراج ژن‌های کاندیدا به‌عنوان یک هدف درمانی یا نشانگرهای زیستی از داده‌های بیولوژیکی حجیم مانند داده‌های بیان ژن را می‌توان به‌عنوان اولین مرحله در درمان سرطان در نظر گرفت. بنابراین، توسعه یک رویکرد کارآمد برای تجزیه و تحلیل چنین داده‌هایی نقشی اساسی در بیوانفورماتیک و زیست‌شناسی محاسباتی دارد. در این مقاله، ما با اعمال الگوریتم رقابت جهانی و ماشین بردار پشتیبان بر روی داده‌های بیان ژن مربوط به سرطان ریه تلاش کرده‌ایم ژن‌های مرتبط با سرطان ریه را به‌عنوان نشانگرهای زیستی بالقوه کشف کنیم. داده‌های مورد استفاده، داده‌های به‌دست آمده از تکنولوژی RNA-Seq و مربوط به نمونه‌های سرطان ریه و همین‌طور نمونه‌های بافت سالم هستند که از پایگاه داده‌ی TCGA دریافت شده‌اند. این داده‌های شامل بیان ژن‌های mRNA در نمونه‌های بافت سرطانی و سالم می‌باشند. نتایج بررسی منجر به کشف ژن‌های با اهمیتی شد که با توجه به مقالات قبلی منتشر شده نقش مهمی در شکل‌گیری سرطان دارند و نقش آن‌ها در شکل‌گیری سرطان ریه را نیز می‌تواند در مطالعات آینده مورد بررسی قرار داد. همچنین نتایج اعتبار سنجی روش پیشنهادی نشان دهنده‌ی قدرت روش‌های مبتنی بر یادگیری ماشین در تحلیل داده‌هایی بیان ژن هستند.

واژه‌های کلیدی

داده‌های بیان ژن
سرطان ریه
الگوریتم رقابت جهانی
ماشین بردار پشتبان
نشانگر زیستی
یادگیری ماشین

مقدمه

یادگیری ماشینی یکی از مهم‌ترین زمینه‌های مطالعه است که به‌طور گسترده در حل بسیاری از مشکلات چالش برانگیز استفاده می‌شود. داده‌کاوی و بهینه‌سازی دو هدف اصلی رویکردهای یادگیری ماشین است. بسیاری از الگوریتم‌های ریاضی، ابتکاری، فراتحصیلی، الهام گرفته از زیست و سایر انواع براساس این دو هدف ساخته شده‌اند. با توجه به رشد پر سرعت داده‌های تولید شده در علم، استخراج داده‌های بزرگ و استخراج اطلاعات مفید مهم‌ترین مسئله در علم داده است.

در سال‌های اخیر، توسعه فن‌آوری توان بالا مانند توالی نسل بعدی (NGS) باعث تولید داده‌های بیولوژیکی بزرگ به‌عنوان مثال داده‌های RNA-Seq می‌شود. تجزیه و تحلیل و استخراج این نوع داده‌ها از طریق روش‌های محاسباتی در مطالعات بیوانفورماتیک در مورد بیماری‌های پیچیده مانند سرطان بسیار مهم است. چنین مطالعاتی به درک مکانیسم بیماری به‌منظور شناسایی و معرفی اهداف درمانی یا نشانگرهای زیستی پیش آگهی و تشخیص کمک می‌کند. در دهه گذشته، رویکردهای یادگیری ماشین به‌طور گسترده‌ای برای تجزیه و تحلیل داده‌های بیولوژیکی به‌ویژه در تحقیقات سرطان استفاده شده است (Kourou et al. 2015; Levine et al. 2019).

به‌عنوان مثال در مطالعه‌ای (Huang et al. 2018)، از یک الگوریتم یادگیری ماشین برای پیش‌بینی پاسخ بیماران سرطانی به داروهای شیمی درمانی بر اساس داده‌های بیان ژن استفاده شده است. همچنین در مطالعه‌ای دیگر برای پیش‌بینی احتمال سرطان پروستات از شبکه عصبی مصنوعی (ANN) استفاده شده است (Jovic et al. 2017). در یک مطالعه‌ی دیگر ترکیبی از الگوریتم پشتیبانی برداری ماشین (SVM) و استاندارد بازگشت ویژگی بازگشتی (RFE) بر روی داده‌های بیان ژن برای پیش‌بینی پاسخ‌های دارویی شخصی در سرطان انجام شد(Huang et al. 2017). استفاده از الگوریتم‌های یادگیری ماشین چندگانه را در پایگاه داده ۹۰۰ بیمار سرطانی اعمال کرده‌اند. محققین سپس نتایج روش پیشنهادی برای پیش‌بینی بقای سرطان پستان مقایسه کردند (Montazeri et al. 2016). آن‌ها نتیجه گرفتند که مدل جنگل‌های تصادفی درختان (TRF) که یک مدل طبقه‌بندی مبتنی بر قاعده است، با بالاترین سطح دقت در پیش‌بینی بقا بهترین است. تشخیص به موقع نقش مهمی در درمان سرطان دارد. بنابراین، یافتن ویژگی‌های بیولوژیکی به‌عنوان نشانگرهای زیستی که قادر به تشخیص بیماری در مراحل اولیه باشند، مسئله بسیار مهمی است (این ویژگی‌ها می‌توانند مجموعه‌ای از ژن‌ها و الگوهای بیان آن‌ها باشند) (Pournoor Elmi et al 2019). رویکردهای یادگیری ماشین نقش محوری در حل این مشکل دارند (Kourou et al. 2015; Huang et al. 2018; Ko et al. 2018) به‌عنوان مثال در یک کار تحقیقاتی، نشانگرهای زیستی را در سرطان سینه با استفاده از انتخاب ویژگی به روش Chi-square و ماشین بردار پشتیبان کشف کردند (Tabl et al. 2019). در مطالعه دیگری محققین با استفاده از شبکه‌های عصبی عمیق و روش t-SNE داده‌های بیان ژن مربوط به سرطان سارکوما را به‌منظور کشف نشانگرهای زیستی و اهداف دارویی، مورد بررسی قرار داده‌اند (van IJzendoorn et al. 2019). شبکه‌های عصبی عمیق برای کشف نشانگرهای زیستی پیش آگاهی‌دهنده در سرطان سینه با استفاده از تصاویر CT نیز استفاده شده‌اند (Wang et al. 2019). سرطان ریه یکی از شایع‌ترین سرطان‌ها در مردان و زنان است. تخمین زده می‌شود که این سرطان حدود ۲ میلیون مورد جدید ابتلا در سال ۲۱۸ داشته است همچنین این سرطان منجر به مرگ ۱/۸ میلیون بیمار در مردان مبتلا در سال ۲۰۱۸ شده است (Bray et al. 2018) ریسک فاکتورهای مختلفی مثل مصرف مشروبات الکلی و سیگار در ابتلای افراد به این نوع سرطان مؤثر است. علی‌رغم مطالعات بسیار انجام شده روی این نوع سرطان، همچنان مکانیزم مولکولی آن ناشناخته مانده است. در این پژوهش سعی کرده‌ایم با اعمال یک روش بر مبنای یادگیری ماشین روی داده‌های بیان ژن نمونه‌های سرطان ریه و بافت سالم، ژن‌های مرتبط با این سرطان که به‌صورت بالقوه می‌توانند نقش نشانگرهای زیستی پیش آگاهی‌دهنده را داشته باشند، کشف کنیم. در سال‌های اخیر، به‌دلیل افزایش تولید داده‌های بیولوژیکی و اهمیت سرطان درمانی در سلامت انسان، مطالعات زیادی در زمینه تجزیه و تحلیل داده‌های بیولوژیکی بر اساس رویکردهای یادگیری ماشین انجام شده است. این مطالعات را می‌توان بر اساس اهداف و منابع داده آن‌ها در گروه‌های زیر دسته‌بندی کرد:

۱- تحلیل و دسته‌بندی تصاویر رادیوگرافی
۲- دسته‌بندی بر اساس ویژگی‌های داده‌های اومیک

تصاویر رادیوگرافی حاصل از روش‌های تصویربرداری مانند اسکن توموگرافی محاسباتی (CT) یا تصاویر MRI منابع مفیدی را برای پیش‌بینی درجه و مراحل تومور فراهم می‌کنند. این تصاویر را می‌توان با استفاده از الگوریتم یادگیری ماشین به‌منظور تقسیم‌بندی و تشخیص تومورهای موجود در یک بافت یا طبقه‌بندی برای جدا کردن سرطان از بافت‌های طبیعی استفاده کرد به‌عنوان مثال می‌توان به‌کمک یک شبکه‌ی عصبی عمیق تصاویر MRI مربوط به بیماران مبتلا به سرطان پروستات را دسته‌بندی کردند (Karimi et al. 2018). استخراج ویژگی از تصاویر پزشکی به‌عنوان نشانگرهای زیستی برای پیش‌بینی بقا، جهش‌ها، عود بیماری و غیره، یکی دیگر از چالش‌های جالب توجه در مطالعات اخیر است.

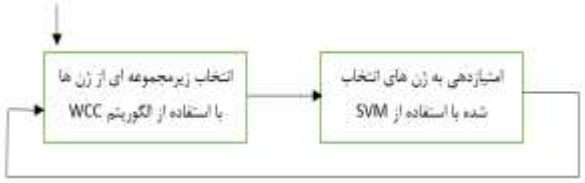
داده‌های چندگانه به‌ویژه داده‌های بیان ژن یکی دیگر از منابع مهم برای رویکردهای یادگیری ماشین در تحقیقات سرطان هستند. به‌طور معمول در این روش‌ها داده‌های ژنومیک (به‌عنوان مثال داده‌های بیان ژن) به‌عنوان ویژگی‌های طبقه‌بندی نمونه‌ها (به‌عنوان مثال طبیعی و سرطانی) در نظر گرفته می‌شوند. به‌طور دقیق‌تر، در داده‌های بیان ژن، الگوی بیان هر ژن یک ویژگی است که شرایط بیولوژیکی نمونه‌ها را نشان می‌دهد. استفاده از چنین ویژگی‌ها و الگوریتم‌های یادگیری ماشین برای شناسایی نشانگرهای زیستی پیش آگهی یا تشخیصی، طبقه‌بندی زیرگروه‌های سرطان، متاستاز یا پیش‌بینی عود بیماری به‌طور گسترده ای در مطالعات سرطان استفاده می‌شوند. نشانگرهای زیستی غیر تهاجمی نشانگرهایی هستند که از مایعات بدن مثل خون به‌دست می‌آیند. چنین نشانگرهایی به‌علت کم خطر بودن برای بیماران در تحقیقات سرطان بسیار با اهمیت هستند. در یک مطالعه محققین موفق شدند به‌کمک روش رگرسیون خطی نشانگرهایی را برای سرطان تخمدان کشف کنند که از طریق خون قابل اندازه‌گیری هستند (Enroth et al. 2019). یادگیری عمیق یکی دیگر از شیوه‌های مرسوم در این دسته از روش‌ها برای تحلیل داده‌های اومیک است که در سال‌های اخیر بسیار مورد توجه بوده است (G et al. 2019; Levine et al. 2019). اگر چه مطالعات جدید نشان می‌دهد که این روش برای تحلیل داده‌های حجیم بسیار قدرتمند است اما چالش‌هایی نیز در این زمینه وجود دارد. به‌عنوان مثال فراهم نمودن مقدار داده‌ی کافی برای آموزش به شبکه‌های عمیق یکی از چالش‌های مطرح است (Hu et al. 2018; Levine et al. 2019).

مواد و روش‌ها

داده‌های مورد استفاده در این پژوهش داده‌های RNA-Seq مربوط به نمونه‌های سرطانی و بافت سالم سرطان ریه هستند که از پایگاه داده‌ی TCGA و پرتال GDC دریافت شده‌اند (Rambow et al. 2016). این داده‌ها شامل ۸۱ نمونه از سرطان ریه از نوع آدنوکارسینوما و ۵۱ نمونه‌ی سالم هستند که هر کدام از آن‌ها شامل بیان ۱۶۴۶۸ ژن کدکننده‌ی پروتئین (mRNA) می‌باشند. ابتدا ژن‌هایی که میانگین بیان آن‌ها در کل نمونه‌ها کمتر از ۱ بود از داده‌ها حذف شدند. سپس با توجه به معیار واریانس ژن‌هایی که واریانس بیان آن‌ها در کل نمونه‌ها از یک مقدار حد آستانه کمتر بود نیز از داده‌ها حذف شدند. این مقدار حد آستانه با در نظر گرفتن میزان واریانس برای همه ی ژن‌ها، چارک اول همه واریانس‌ها در نظر گرفته شد. برای کشف نمونه‌های outlier نیز داده‌ها را پس از نرمال‌سازی با استفاده از روش خوشه‌بندی سلسله مراتبی و فاصله‌ی اقلیدسی دسته‌بندی کردیم و آن دسته از نمونه‌هایی که فاصله‌ی آن‌ها از دو گروه داده بیشتر از سایر نمونه‌ها بودند حذف شدند.

نرمال‌سازی داده‌ها نیز با استفاده از ابزار edgeR و روش TMM انجام شده است (Robinson et al. 2010; Robinson and Oshlack 2010). همچنین پس از نرمال‌سازی از مقادیر بیان ژن‌ها لوگاریتم گرفته شده است.

در این بخش روش پیشنهادی به‌منظور انتخاب مجموعه بهینه‌ای از ژن‌ها از بین ۱۷۰۰۰ ژن موجود تشریح می‌شود. در راه‌کار پیشنهادی، ژن‌های انتخاب شده به‌وسیله ماشین‌های بردار پشتیبان امتیازدهی می‌شوند و آن دسته از ژن‌هایی که می‌توانند سبب افزایش قدرت تفکیک‌پذیری مدل شوند، به‌عنوان نشان‌گرهای زیستی بالقوه معرفی می‌شوند. شماتیک کلی روش پیشنهادی در شکل ۱ ترسیم شده است.



شکل ۱- شماتیک کلی روش پیشنهادی

در شکل ۱ مشاهده می‌شود که چگونه روش پیشنهادی از الگوریتم رقابت جهانی برای انتخاب ژن‌ها استفاده می‌کند(Masoudi-Sobhanzadeh and Motieghader 2016; Masoudi-Sobhanzadeh et al. 2021; Masoudi-Sobhanzadeh et al. 2019; MotieGhader et al. 2017). در ادامه فرموله‌سازی این الگوریتم برای مسئله انتخاب نشانگرهای زیستی بالقوه تشریح می‌شود. همانند دیگر الگوریتم‌های بهینه‌سازی، الگوریتم رقابت جهانی کار خود را با ایجاد جمعیت اولیه‌ای از پاسخ‌های بالقوه که هر کدام از آن‌ها یک تیم نامیده می‌شود آغاز می‌کند. اجزا تشکیل‌دهنده تیم‌ها در این مسئله، تعداد تقریبی ۱۷۰۰۰ ژن موجود می‌باشند. به‌عبارتی دیگر، اندازه هر تیم برابر با ۱۷۰۰۰ بیت می‌باشد که هر یک از آن‌ها نیز نشان‌دهنده یک ژن می‌باشند. در صورتی‌که یک ژن به‌عنوان نشانگر زیستی بالقوه انتخاب شود، بیت متناظر آن یک و در غیر این‌صورت صفر خواهد بود. پس از ایجاد تیم‌ها، میزان شایستگی آن‌ها محاسبه می‌شود. برای محاسبه این‌که مجموعه ژن‌های انتخب شده تا چه اندازه‌ای می‌توانند به‌عنوان نشانگرهای زیستی معرفی شوند، از ماشین‌های بردار پشتیبان استفاده شده است. معیار ارزیابی نیز Accuracy در نظر گرفته شده است که از طریق رابطه (۱) محاسبه می‌شود:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \qquad (۱)$$

در رابطه (۱)، پارامترهای TP، TN، FP و FN به‌شرح زیر می‌باشند:

۱- TP یا True Positive: نمونه داده شده یک نمونه سرطانی بوده است و مدل نیز به‌درستی آن را پیش‌بینی کرده است.

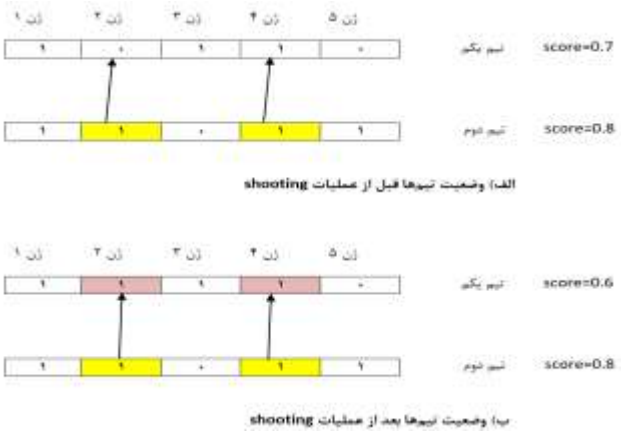
۲- TN یا True Negative: نمونه داده شده یک نمونه سالم بوده است و مدل نیز به‌درستی آن را پیش‌بینی کرده است.

۳- FP یا False Positive: نمونه داده شده یک نمونه سالم بوده است اما مدل به‌صورت اشتباه پیش‌بینی کرده است که نمونه داده شده سرطانی است.

٤- FN یا False Negative: نمونه داده شده یک نمونه سرطانی بوده است اما مدل به‌صورت اشتباه پیش‌بینی کرده است که نمونه داده شده سالم است.

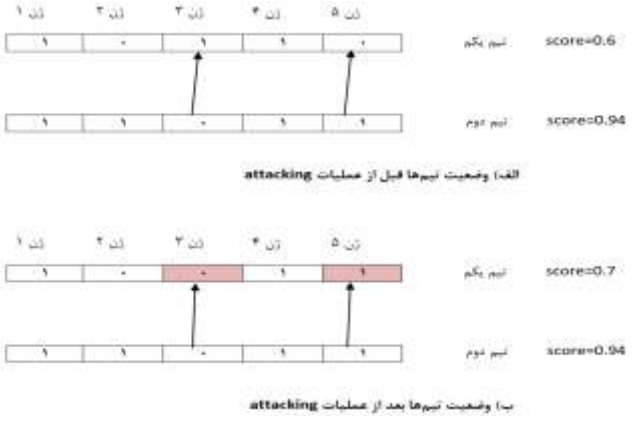
پس از محاسبه برازندگی، تیم‌ها در گروه‌های مختلف قرار می‌گیرند و در آنجا به رقابت می‌پردازند. برای گروه‌بندی تیم‌ها از روش حداکثر اختلاف بین پاسخ‌های بالقوه در این مطالعه استفاده شده است. علت استفاده از این روش نیز این بوده است که تضمین شود در هر گروه جواب‌های اولیه مشابه یکدیگر نیستند و تنوع پاسخ‌ها در هر گروه وجود دارد. پس از تشکیل گروه‌ها، تیم‌ها با یکدیگر به رقابت می‌پردازند که این کار بر اساس عملگرهای الگوریتم رقابت جهانی که صورت می‌گیرد که در ادامه هر یک از آن‌ها تشریح می‌شوند.

عملگر shooting به این ترتیب اعمال می‌شود که تیم جاری یک سری از مقادیر خودش را به‌صورت تصادفی انتخاب می‌کند و به سمت تیم مقابل پرتاب می‌کند. اگر تیم مقابل شایستگی مناسب‌تری یا score بهتری پیدا کند مقادیر دریافت شده را ثبت می‌کند ولی اگر مقدار score آن کمتر از مقدار قبلی باشد، مقادیر جدید دریافت شده را نادیده می‌گیرد. این عمل را می‌توان این‌گونه تفسیر کرد که تیمی که عمل shooting را انجام می‌دهد بخشی از مقادیر متغیرهای خود را (ژن‌های انتخاب شده) به تیم مقابل ارسال می‌کند. به‌عنوان مثال در شکل ۲ تیم دوم دو نطقه تصادفی از تیم شماره یک را انتخاب می‌کند و مقادیر خود را به آن انتساب می‌دهد که با انجام این کار، بازیکنان شماره دو و چهار از تیم دوم، مقادیر خود را به سمت بازیکنان با شماره‌های دو و چهار تیم یک پرتاب می‌کنند. تیم یک نیز با دریافت مقادیر جدید امتیاز خود را محاسبه می‌کند و در صورتی‌که مقادیر جدید باعث بهبود امتیاز آن شده است، آن‌ها را می‌پذیرد. در صورتی‌که مقادیر جدید امتیاز تیم یکم را بهبود نمی‌دادند، تیم یک آن‌ها را نادیده می‌گرفت.



شکل ۲- مثالی از عملگر Shooting

در عملگر attacking هر تیم یک سری از مقادیر جدید را به‌صورت تصادفی ایجاد می‌کند و به سمت تیم مقابل می فرستد. اگر تیم مقابل شایستگی مناسب‌تری پیدا کند مقادیر دریافت شده را ثبت می‌کند ولی اگر مقدار score ان کمتر باشد مقادیر جدید دریافت شده را نادیده می‌گیرد. تفاوتی که بین عملیات shooting و attacking در الگوریتم WCC وجود دارد آن است که در عملیات shooting، یک تیم مقادیر خود را به تیم دیگر می‌فرستد اما در عملیات attacking یک تیم مقادیر تصادفی جدیدا ایجاد شده را به تیم مقابل می‌فرستد این کار باعث آن می‌شود که الگوریتم رقابت‌های جهانی در بهینه‌های محلی گیر نکند. چگونگی انجام عملیات attacking بر روی پنج ژن فرضی یک گراف در شکل ۳ مشاهده می‌شود. در عملیات انجام شده مقدار امتیاز تیم شماره یک به ۰/۷ افزایش یافته است بنابراین، این تیم مقادیر دریافت شده را تثبیت می‌گیرد. تیم شماره یک در صورتی مقادیر دریافتی را به‌عنوان نشانگرهای بالقوه زیستی در نظر می‌گیرد که این مقادیر باعث بهبود تابع امتیاز گردند. عملگر دیگری که در الگوریتم WCC وجود دارد عملیات passing است. در عملیات passing نیز دو نقطه به‌صورت تصادفی انتخاب می‌شوند و مقادیر آن‌ها با یکدیگر جابه‌جا می‌شوند. نمونه‌ای از عملیات passing در شکل ۵ مشاهده می‌شود که در آن تعویض دو نقطه از یک تیم سبب انتخاب و عدم انتخاب ژن‌های مربوطه به‌عنوان نشانگرهای زیستی بالقوه شده است.



شکل ۳- مثالی از عملگر Attacking

در عملیات crossing دو نقطه از یک تیم به‌صورت تصادفی انتخاب می‌شوند و مقادیر آن‌ها با استفاده از عملگر شیفت با یکدیگر جابه‌جا می‌گردند. همانند عملیات attacking و shooting یک تیم در صورتی مقادیر جدید را می‌پذیرد که باعث بهبود امتیاز یک تیم شوند در غیر این‌صورت مقادیر دریافت شده را نادیده می‌گیرد. این عملیات بر خلاف دو عملگر قبلی تک تیمی می‌باشد. به‌عبارتی دیگر یک تیم در خود تغییرات را ایجاد می‌کند و تیم ثانی در این تغییرات نقشی ندارد. مثالی از عملیات مربوط به عملگر crossing در شکل ٤ نشان داده شده است.

شکل ٤- مثالی از عملگر Attacking

برخلاف شکل ٤ که در آن عملیات انجام شده مطلوبیت پاسخ را افزایش داده است، در عملیات انجام شده در شکل ۵ مطلوبیت پاسخ کاهش پیدا کرده است بنابراین تغییرات انجام شده نادیده گرفته خواهند شد.

شکل ۵- مثالی از عملگر Passing

پس از اعمال عملگرهای الگوریتم و خاتمه مسابقات گروهی، مسابقات حذفی برگزار می‌شوند که نحوه اعمال تغییرات در آن‌ها همانند آن چیزی است که برای مسابقات گروهی توسط عملگرها اعمال می‌شوند. در انتهای مسابقات تنها یک تیم باقی می‌ماند که ژن‌های آن به‌عنوان نشانگرهای بالقوه سرطان ریه معرفی می‌شوند. شرط خاتمه الگوریتم نیز تعداد از پیش تعریف شده تکرار مراحل الگوریتم (۳۰ تکرار) در نظر گرفته شده است. همچنین تعداد اولیه تیم‌ها برابر با ۱۰۰ تیم در نظر گرفته شده‌اند که در ۸ گروه به رقابت پرداخته‌اند. دیگر پارامترهای الگوریتم به‌صورت خودکار و توسط خود الگوریتم تغییر می‌یابند.

نتایج و بحث

روش پیشنهادی با دو الگوریتم دیگر به نام کولونی مورچه‌ها (ACO) و الوریتم ژنتیک (GA) مقایسه شده است. برای اجرای الگوریتم‌ها از ابزار Feature Select (Masoudi-Sobhanzadeh et al. 2019) استفاده شده است و پارامترهای الگوریتم‌ها به‌صورت پیش فرض در نظر گرفته شده‌اند. برای ارزیابی روش‌ها از شیوه‌ی 5-fold cross validation استفاده شده است. جدول ۱ نتایج اجرای سه روش مورد نظر بر روی داده‌ها را نشان می‌دهد. در این جدول به‌ترتیب از چپ به راست از معیارهای sensitivity، specificity، precision، false positive rate و accuracy استفاده شده است. همان‌طور که در این جدول مشاهده می‌شود، به غیر از معیار specificity در سایر معیارها الگوریتم رقابت جهانی عملکرد بهتری دارد.

جدول ۱- نتایج ارزیابی الگوریتم‌هابه شیوه‌ی 5-fold cross validation بر روی داده‌های سرطان ریه.

نام الگوریتم	SEN	SPC	PRE	FPR	ACC
WCC	0.98	0.96	0.99	0.03	0.98
GA	0.96	0.96	0.94	0.04	0.95
ACO	0.94	0.96	0.93	0.04	0.95

جهت بررسی بیشتر کارایی الگوریتم‌ها و میزان خطای آن‌ها نرخ خطای الگوریتم‌ها برای انتخاب چهار ویژگی (ژن) و بهترین نتیجه‌ی به‌دست آمده توسط الگوریتم‌ها مقایسه شده است. نتایج این مقایسه در جدول ۲ قابل مشاهده است. ستون‌های این جدول به‌ترتیب از چپ به راست شامل موارد زیر است:

NOF: Number of Features این پارمتر نشان می‌دهد هر کدام از الگوریتم‌ها از چه تعداد ویژگی برای طبقه‌بندی داده‌ها استفاده می‌کنند.

ET: این پارامتر زمان مورد نیاز برای اجرای هر الگوریتم را بر حسب ثانیه بیان می‌کند.

SC_STD: در راه‌کار پیشنهادی SC نشان‌دهنده مقدار امتیاز یک پاسخ بالقوه می‌باشد.

SC_CI_x: این پارامتر نشان‌دهنده حد پائین و بالای بازه قابل اطمینان برای امتیاز می‌باشد که از طریق سی بار اجرای مختلف الگوریتم‌ها سنجیده می‌شود

SC_P: این پارامتر آماری نیز مخفف P-value برای امتیازهای به‌دست آمده می‌باشد و نشان می‌دهد که تا چه اندازه‌ای نتایج الگوریتم‌ها به‌صورت تصادفی حاصل شده‌اند.

SC_TS: بر اساس این پارامتر آماری معنادار بودن نتایج مورد ارزیابی قرار می‌گیرد. این پارامتر با پارامتر مقدار P رابطه معکوس دارد.

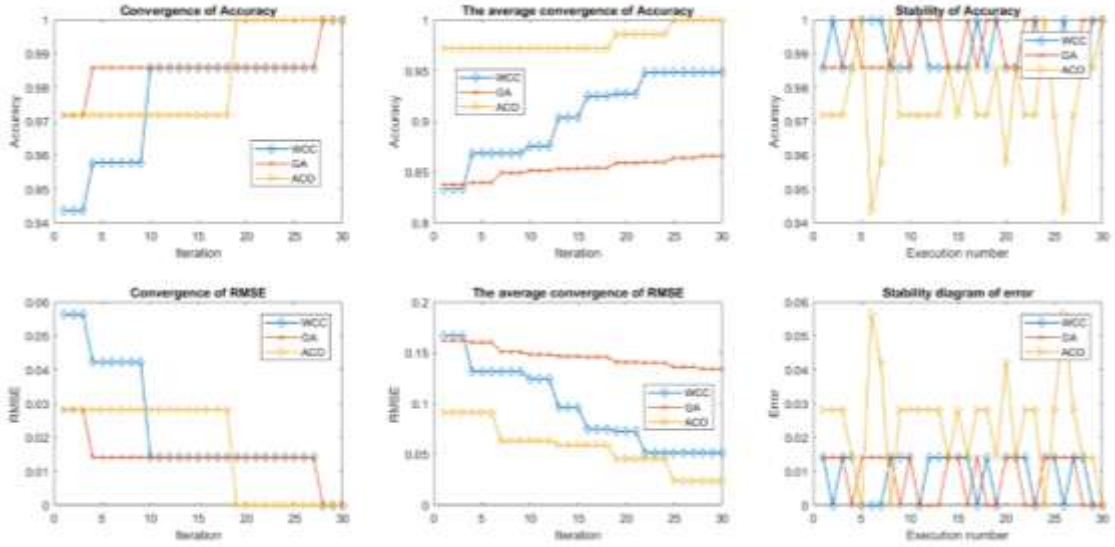
تفسیر نتایج برای دیگر پارامترها همانند موارد مطرح شده می‌باشد با این تفاوت که معیار ارزیابی مقدار حاصل شده از خطا (ER) به‌جای امتیاز (SC) می‌باشد. معیارها با استفاده از ابزار Feature Select محاسبه شده و نحوه‌ی محاسبه در منبع (Masoudi-Sobhanzadeh Motieghader et al. 2019) قابل دسترس است.

همچنین نرخ همگرایی معیارهای مختلف بر حسب تعداد تکرارهای الگوریتم در شکل ٦ قابل مشاهده است.

ژن‌های انتخاب شده توسط الگوریتم‌ها در جدول ۳ قابل مشاهده هستند. بررسی مقالات و متون منتشر شده نشان می دهد که بیان ژن *AP1G1* در سرطان سر و گردن نقش دارد (Tao et al. 2017). پروتئینی که این ژن کد می‌کند یک انتقال‌دهنده اسید آمینه در سلول است که نقش بسیار پررنگی در جذب گلوتامین در سلول‌های سرطانی دارد. این پروتئین همچنین ارتباط نزدیکی با مسیر پیام رسانی PI3K/AKT دارد.

جدول ۲- مقایسه الگوریتم‌ها از نظر پارامترهای آماری

الگوریتم	NOF	ET	SC	SC_STD	SC_CI_1	SC_CI_2	SC_P	SC_TS	ER	ER_STD	ER_CI_1	ER_CI_2	ER_P	ER_TS
WCC	4	24.27	1	0/006	0/988	0/993	2/504	786/3	0	0/006	0/006	0/011	8/716	7/077
GA	4	3.09	1	0/007	0/969	0/974	5/473	765/4	0	0/007	0/007	0/013	1/032	6/158
ACO	4	15.66	1	0/014	0/950	0/961	7/032	371/3	0	0/014	0/018	0/029	5/242	9/109



شکل ٦- نمودار همگرائی الگوریتم‌ها

جدول ۳- ژن‌های کشف شده توسط الگوریتم پیشنهادی و دو الگوریتم ACO و GA.

الگوریتم	ژن‌های انتخاب شده
WCC	AP1G1, AP3M1, FIZ1, SPATS1
GA	SOCS4, GCA, C19orf57, ASNS
ACO	CASZ1, SLC9A9, TPT1, ZNF362

ژن *HCP5* با غیر فعال کردن AP1G1 و از طریق مسیر پیام رسانی PI3K/AKT روند تومورزایی در سرطان روده بزرگ را افزایش می‌دهد (Yun et al. 2019) همچنین در سرطان کبد مهار این ژن توسط lncRNA بسیار مهم MEG3 و تأثیر آن در مسیر پیام رسانی PI3K/AKT منجر به افزایش تهاجم سلول‌های سرطانی می‌شود (Sun Cao et al. 2019). نقش این ژن در سرطان ریه می‌تواند مورد بررسی قرار بگیرد. ژن FIZ1 یک پروتئین zinc finger را کد می‌کند که این پروتئین با پروتئین Flt3 که یک receptor tyrosin kinase است برهمکنش دارد. پروتئین Flt3 در بقا و تقسیم سلولس برخی سلول‌های خونساز در بدن نقش دارد (Wolf and Rohrschneider 1999) از این رو نقش ژن کشف شده در سرطان می‌تواند قابل بررسی باشد زیرا فرآیند تقسیم سلولی ارتباط مستقیم با ایجاد سلول‌های سرطانی دارد. ژن SOCS4 یک سرکوب‌گر تومور مهم در سرطان معده است. هایپر متیله شدن این ژن و توقف‌بان آن منجر به ایجاد تومور در آن سرطان بسیار شایع می‌شود (Kobayashi et al. 2012). بنابراین متیله شدن این ژن می‌تواند نشانگر زیستی پیش آگهی دهنده برای سرطان معده باشد. همچنین اخیراً در این پژوهش نقش این ژن در سرطان تخمدان نیز به اثبات رسیده است (Yang et al. 2020). علاوه بر آن در یک کار تحقیقاتی دیگر محققین نشان داده‌اند که miR-1290 از طریق مهار این ژن تکثیر و مهاجم بودن سلول‌های سرطان آدنوکارسینومای ریه (مورد مطالعه در پژوهش حاضر) را تحریک می‌کند (Xiao et al. 2018). علاوه بر این‌ها بیان بالای این ژن و سایر هم‌خانواده‌های آن در شکل‌گیری تومور و روند پاسخ به درمان سرطان پستان نیز نقش دارد (Sasi et al. 2010) ژن *ASNS* یکی از ژن‌های بسیار مهم در سرطان به‌ویژه سرطان خون است (Krall et al. 2016; Chiu et al. 2020). این ژن کد کننده پروتئین آسپاراژین است. مقدار این پروتئین در سلول برایی رشد سلول‌ها سرطان بسار پر اهمیت است (Krall et al. 2016). همچنین در سرطان ریه نیز شواهدی مبنی بر نقش این ژن دیده شده است (Xu et al. 2016). ژن *CASZ1* یک سرکوب‌گر تومور در سرطان مغز است (Liu et al. 2011). این ژن همچنین عامل متاستاز در سرطان تخمدان است (Wu et al. 2016). ژن *SLC9A9* نیز یک ژن مؤثر در مکانیزم سرطان روده‌ی بزرگ است (Ueda et al. 2017). ژن *TPT1* که ارتباط نزدیکی با ژن با اهمیت p53 دارد، محرک بقای سلول است و شواهدی مبنی بر ارتباط آن با سرطان ریه وجود دارد (Chen et al. 2013).

نتایج حاصل از بررسی مقالات نشان می‌دهد که روش‌های اعمال شده به‌طور هدفمند قادر بوده‌اند ژن‌های مرتبط با سرطان را از بین هزاران ژن موجود در داده‌ها کشف کنند. در نتیجه می‌توان سایر ژن‌های موجود در جدول ۳ را جهت بررسی بیشتر در مطالعات آتی پیشنهاد کرد.

نتیجه‌گیری کلی

در این مقاله سعی شد با استفاده از یک رویکرد انتخاب ویژگی بر مبنای یادگیری ماشین ژن‌های مرتبط با سرطان ریه کشف شود. انتخاب ویژگی با توجه به برچسب داده‌ها که شامل نمونه‌های سرطانی و سالم هستند، منجر به کشف ژن‌هایی می‌شود که الگوی بیان آن‌ها در گروه از نمونه‌ها متفاوت است. این تفاوت اهمیت ژن‌های کشف شده در تحقیقات مربوط به کشف نشانگرهای زیستی پیش آگاهی دهنده را بالا می‌برد. هر چند که نتایج حاصل نیازمند تایید آزمایشگاهی بر روی داده‌های مختلف است. نتایج ارزیابی روش پیشنهادی و سایر روش‌های مورد آزمایش نشان می‌دهد که علی‌رغم حجم بالای تعداد ویژگی‌ها در داده‌ی مورد استفاده، روش‌های انتخاب ویژگی قادرند داده‌ها را به‌درستی تحلیل نموده و ویژگی‌های مناسب را انتخاب کنند.

به‌عنوان پیشنهاد برای مطالعات آینده بررسی روش‌های یادگیری عمیق و استخراج ویژگی با استفاده از شیوه‌هایی مثل autoencoderها میتواند مناسب باشد. البته استفاده از چنین روش‌هایی مستلزم در اختیار داشتن تعداد زیادی داده به‌همراه برچسب آن‌ها است. پایگاه داده‌ی TCGA و داده‌های سرطانی می‌توانند منبع خوبی برای این هدف باشند.

علاوه بر بیماری سرطان چنین رویکردهایی میتوانند برای بررسی و کشف مکانیزم سایر بیماری‌های پیچیده مثل بیماری‌های خود ایمنی نیز مناسب باشد و به‌عنوان پیشنهاد برای مطالعات آینده می‌تواند مد نظر قرار بگیرد.

منابع

Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, and Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394-424.

Chen W, Wang H, Tao S, Zheng Y, Wu W, Lian F, Jaramillo M, Fang D and Zhang DD (2013) Tumor protein translationally controlled 1 is a p53 target gene that promotes cell survival. Cell Cycle 12:2321-2328.

Chiu, M., Taurino, G., Bianchi, M. G., Kilberg, M. S., and Bussolati, O. (2020). Asparagine synthetase in cancer: beyond acute lymphoblastic leukemia. Frontiers in oncology 9:1480.

Enroth S, Berggrund M, Lycke M, Broberg J, Lundberg M, Assarsson E, Olovsson M, Stalberg K, Sundfeldt K, and Gyllensten U (2019) High throughput proteomics identifies a high-accuracy 11 plasma protein biomarker signature for ovarian cancer. Commun Biol 2:221.

G ST, Biswas M, O GK, Tiwari A, H SS, Turk M, Laird JR, Asare CK, A AA, N NK, B KM, Saba L, and Suri JS (2019) A Review on a Deep Learning Perspective in Brain Cancer Classification. Cancers (Basel) 11.

Hu Z, Tang J, Wang Z, Zhang K, Zhang L and Sun Q (2018) Deep learning for image-based cancer detection and diagnosis− A survey. Pattern Recognition 83:134-149.

Huang C, Mezencev R, McDonald JF and Vannberg F (2017) Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. PLoS One 12:e0186906.

Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y and Xu W (2018) Applications of Support Vector Machine (SVM) Learning in Cancer Genomics. Cancer Genomics Proteomics 15:41-51.

Jovic S, Miljkovic M, Ivanovic M, Saranovic M and Arsic M (2017) Prostate Cancer Probability Prediction By Machine Learning Technique. Cancer Invest 35:647-651.

Karimi D, Samei G, Kesch C, Nir G and Salcudean SE (2018) Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models. Int J Comput Assist Radiol Surg 13:1211-1219.

Ko J, Baldassano SN, Loh PL, Kording K, Litt B and Issadore D (2018) Machine learning to detect signatures of disease in liquid biopsies - a user's guide. Lab Chip 18:395-405.

Kobayashi D, Nomoto S, Kodera Y, Fujiwara M, Koike M, Nakayama G, Ohashi N and Nakao A (2012) Suppressor of cytokine signaling 4 detected as a novel gastric cancer suppressor gene using double combination array analysis. World journal of surgery 36:362-372.

Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV and Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8-17.

Krall AS, Xu S, Graeber TG, Braas D and Christofk HR (2016) Asparagine promotes cancer cell proliferation through use as an amino acid exchange factor. Nature communications 7:1-13.

Levine AB, Schlosser C, Grewal J, Coope R, Jones SJM, and Yip S (2019) Rise of the Machines: Advances in Deep Learning for Cancer Diagnosis. Trends Cancer 5:157-169.

Liu Z, Yang X, Li Z, McMahon C, Sizer C, Barenboim-Stapleton L, Bliskovsky V, Mock B, Ried T and London W (2011) CASZ1, a candidate tumor-suppressor gene, suppresses neuroblastoma tumor growth through reprogramming gene expression. Cell Death and Differentiation 18:1174-1183.

Masoudi-Sobhanzadeh Y and Motieghader H (2016) World Competitive Contests (WCC) algorithm: A novel intelligent optimization algorithm for biological and non-biological problems. Informatics in Medicine Unlocked 3:15-28.

Masoudi-Sobhanzadeh Y, Motieghader H and Masoudi-Nejad A (2019) FeatureSelect: a software for feature selection based on machine learning approaches. BMC Bioinformatics 20:170.

Masoudi-Sobhanzadeh Y, Motieghader H, Omidi Y and Masoudi-Nejad A (2021) A machine learning method based on the genetic and world competitive contests algorithms for selecting genes or features in biological applications. Scientific Reports 11:1-19.

Masoudi-Sobhanzadeh Y, Omidi Y, Amanlou M and Masoudi-Nejad A (2019) DrugR+: a comprehensive relational database for drug repurposing, combination therapy, and replacement therapy. Computers in biology and medicine 109:254-262.

Montazeri M, Montazeri M, Montazeri M and Beigzadeh A (2016) Machine learning models in breast cancer survival prediction. Technol Health Care 24:31-42.

MotieGhader H, Gharaghani S, Masoudi-Sobhanzadeh Y and Masoudi-Nejad A (2017) Sequential and mixed genetic algorithm and learning automata (SGALA, MGALA) for feature selection in QSAR. Iranian journal of pharmaceutical research: IJPR 16:533.

Pournoor E, Elmi N, Masoudi-Sobhanzadeh Y and Masoudi-Nejad A (2019) Disease global behavior: a systematic study of the human interactome network reveals conserved topological features among categories of diseases. Informatics in Medicine Unlocked 17:100249.

Rambow F, Bechadergue A, Luciani F, Gros G, Domingues M, Bonaventure J, Meurice G, Marine JC and Larue L (2016) Regulation of melanoma progression through the TCF4/miR-125b/NEDD9 cascade. Journal of Investigative Dermatology 136:1229-1237.

Robinson MD, McCarthy DJ and Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, 26:139-140.

Robinson MD and Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome biology 11:1-9.

Sasi W, Jiang WG, Sharma A and Mokbel K (2010) Higher expression levels of SOCS 1, 3, 4, 7 are associated with earlier tumour stage and better clinical outcome in human breast cancer. BMC cancer 10:1-13.

Sun Y, Cao F, Qu L, Wang Z and Liu X (2019) MEG3 promotes liver cancer by activating PI3K/AKT pathway through regulating AP1G1. Eur Rev Med Pharmacol Sci 23:1459-1467.

Tabl AA, Alkhateeb A, ElMaraghy W, Rueda L and Ngom A (2019) A Machine Learning Approach for Identifying Gene Biomarkers Guiding the Treatment of Breast Cancer. Front Genet 10:256.

Tao X, Lu Y, Qiu S, Wang Y, Qin J and Fan Z (2017) AP1G1 is involved in cetuximab-mediated downregulation of ASCT2-EGFR complex and sensitization of human head and neck squamous cell carcinoma cells to ROS-induced apoptosis. Cancer letters 408:33-42.

Ueda M, Iguchi T, Masuda T, Komatsu H, Nambara S, Sakimura S, Hirata H, Uchi R, Eguchi H and Ito S (2017) Up-regulation of SLC9A9 promotes cancer progression and is involved in poor prognosis in colorectal cancer. Anticancer research 37:2255-2263.

van IJzendoorn DG, Szuhai K, Briaire-de Bruijn IH, Kostine M, Kuijjer ML and Bovée JV (2019) Machine learning analysis of gene expression data reveals novel diagnostic and prognostic biomarkers and identifies therapeutic targets for soft tissue sarcomas. PLoS computational biology 15:e1006826.

Wang S, Liu Z, Rong Y, Zhou B, Bai Y, Wei W, Wei W, Wang M, Guo Y and Tian J (2019) Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. Radiother Oncol 132:171-177.

Wolf I and Rohrschneider LR (1999) Fiz1, a novel zinc finger protein interacting with the receptor tyrosine kinase Flt3. Journal of Biological Chemistry 274:21478-21484.

Wu YY, Chang CL, Chuang YJ, Wu JE, Tung CH, Chen YC, Chen YL, Hong TM and Hsu KF (2016) CASZ1 is a novel promoter of metastasis in ovarian cancer. American journal of cancer research 6:1253.

Xiao X, Yang D, Gong X, Mo D, Pan S and Xu J (2018) miR-1290 promotes lung adenocarcinoma cell proliferation and invasion by targeting SOCS4. Oncotarget 9:11977.

Xu Y, Lv F, Zhu X, Wu Y and Shen X (2016) Loss of asparagine synthetase suppresses the growth of human lung cancer cells by arresting cell cycle at G0/G1 phase. Cancer gene therapy 23:287-294.

Yang M, Chen H, Zhou L, Huang X, Su F and Wang P (2020) Identification of SOCS family members with prognostic values in human ovarian cancer. American Journal of Translational Research 12:1824.

Yun W, Hu Y, Zhao C, Yu D and Tang J (2019) HCP5 promotes colon cancer development by activating AP1G1 via PI3K/AKT pathway. Eur Rev Med Pharmacol Sci 23:2786-2793.