Iranian Quarterly Journal of Breast Disease. 2019; 12(1):39- 47.

Original Article

Diagnosis of Breast Cancer Subtypes using the Selection of Effective Genes from Microarray Data

Tabatabaei A[1], Derhami V[1*], Sheikhpour R[2], Pajoohan M-R[1]

[1] Computer Engineering Department, Faculty of Engineering, Yazd University, Yazd, Iran
[2] Department of Computer Engineering, Faculty of Engineering, Ardakan University, Ardakan, Iran

Receive: 2019/2/13
Accepted: 2019/4/14

[*]**Corresponding Author:**
Vali Derhami
vderhami@yazd.ac.ir

Ethics Approval:

Abstract

Introduction: Early diagnosis of breast cancer and the identification of effective genes are important issues in the treatment and survival of the patients. Gene expression data obtained using DNA microarray in combination with machine learning algorithms can provide new and intelligent methods for diagnosis of breast cancer.

Methods: Data on the expression of 9216 genes from 84 patients across 5 different types of cancer was obtained using microarray technology. In this study, we proposed a feature selection method based on the correlation between abnormal expression of genes and cancer for diagnosis of breast cancer. Then, we used K-nearest neighbor (KNN), support vector machine (SVM), and naive Bayesian (NB) classifiers to evaluate the performance of the proposed method in the selection of relevant genes.

Results: The proposed feature selection method coupled with the KNN classifier predicted all types of cancer with 100% accuracy and using 38 of the 9216 genes. The proposed method could also identify the genes associated with each class. Moreover, the proposed feature selection method coupled with NB and SVM classifiers achieved accuracy rates of 90% and 96.67% using 17 and 22 genes, respectively.

Conclusion: The results of this study demonstrated that the proposed feature selection method has better performance compared with other methods. The proposed method is able to distinguish the genes involved in each cancer class and detect overexpression or underexpression of selected genes, which can be used by physicians and researchers in the field of health care.

Keywords: Breast Cancer, Feature Selection, Microarray Data, Classification

فصلنامه بیماری های پستان ایران
۱۳۹۸؛ ۱۲(۱): ۳۹-۴۷

مقاله پژوهشی

تشخیص نوع سرطان پستان با استفاده از انتخاب ژن‌های موثر از داده‌های ریزآرایه

سیدابوالفضل طباطبایی[۱]، ولی درهمی[۱*]، راضیه شیخ‌پور[۲]، محمدرضا پژوهان[۱]

[۱]گروه مهندسی کامپیوتر، پردیس فنی و مهندسی، دانشگاه یزد، یزد، ایران
[۲]گروه مهندسی کامپیوتر، دانشکده فنی و مهندسی، دانشگاه اردکان، اردکان، ایران

تاریخ ارسال: ۹۷/۱۱/۲۴
تاریخ پذیرش: ۹۸/۱/۲۵

نشانی نویسنده مسئول:
ولی درهمی
vderhami@yazd.ac.ir

چکیده

مقدمه: تشخیص زودهنگام سرطان پستان و ژن‌های موثر در آن نقش بسیار کلیدی در درمان و حیات بیمار ایفا می‌کند. با استفاده ازداده‌های بیان ژن استخراج شده از فناوری ریزآرایه و الگوریتم‌های یادگیری ماشین می‌توان روش‌های نوین و هوشمندی درنظام سلامت و درمان ارایه داد که با دقت بالایی قادر به تشخیص سرطان پستان باشند.

روش بررسی: داده‌های استفاده شده در این پژوهش، شامل داده‌های بیان۹۲۱۶ ژن مربوط به ۸۴ بیمار در ۵ نوع مختلف سرطان است که با استفاده از فناوری ریزآرایه به دست آمده است. در این مطالعه برای افزایش کارایی سیستم‌های تشخیص سرطان پستان، روش انتخاب ویژگی مبتنی بر ارتباط بین بیان غیرنرمال ژن و سرطان ارایه شده است. سپس از سه دسته‌بند پرکاربرد و رایج K- نزدیک‌ترین همسایه (KNN)، ماشین بردار پشتیبان (SVM) و بیزی ساده (NB) برای سنجش کارایی ژن‌های انتخاب شده استفاده شد.

یافته‌ها: بررسی‌های انجام شده نشان دادند که با روش انتخاب ویژگی پیشنهادی و با استفاده از دسته‌بند KNN می‌توان فقط با انتخاب ۳۸ ژن از میان ۹۲۱۶ ژن مربوط به داده‌های بیماران سرطانی که با استفاده از فناوری ریزآرایه به دست آمده است، انواع سرطان پستان بیماران در داده‌های آزمایش را با دقت ۱۰۰٪ تشخیص داد و ژن‌های مرتبط با هر کلاس را نیز تفکیک کرد. همچنین با استفاده از دسته‌بند NB، دقت ۹۰٪ با ۱۷ ژن و با دسته‌بند SVM، دقت ۹۶/۶۷٪ با ۲۲ ژن انتخابی به‌دست آمد.

نتیجه‌گیری: نتایج این مطالعه نشان دادند که روش انتخاب ویژگی پیشنهادی، با در نظر گرفتن همزمان دو معیار دقت و تعداد ژن انتخابی، عملکرد مناسبی نسبت به سایر روش‌ها دارد. توانایی تفکیک ژن‌های موثر در هر کلاس سرطان، به‌علاوه تشخیص بیان بیشتر از حد یا کمتر از حد ژن‌های انتخابی، از خصوصیات ویژه روش پیشنهادی است که می‌تواند مورد استفاده متخصصین و پژوهشگران حوزه درمان و مراقبت قرار گیرد.

واژه‌های کلیدی: سرطان پستان، انتخاب ویژگی، داده‌های ریزآرایه، دسته‌بندی

مقدمه

سرطان پستان شایع‌ترین سرطان و دومین عامل مرگ و میر در زنان آمریکایی است (۱). در ایران سالیانه بیش از ۷۰۰۰ مورد به مبتلایان سرطان پستان اضافه می‌شود (۲). در این بیماری سلول‌های بدخیم (دارای رشد غیر قابل کنترل) در بافت پستان شروع به فعالیت می‌کنند (۳، ۴). علاوه بر پارامترهای تشخیص سنتی بیماری مثل اندازه و درجه (گرید) تومور، معمولاً از وضعیت مثبت یا منفی بودن نشان‌گرهای ایمونوهیستوشیمی IHC[1] نظیر گیرنده‌های ER[2] و PR[3] و HER2[4] و نشان‌گر Ki67 برای تشخیص و مراقبت از بیماران سرطانی استفاده می‌شود (۵، ۶). از دیدگاه مولکولی، انواع تومورهای پستان به پنج دسته اصلی با علایم بالینی متمایز تقسیم می‌شوند (۷، ۸). این دسته‌ها عبارتند از: Luminal A, Luminal B, HER2 over-expression, Basal-like, Normal-like. غیر از گروه Normal-like، سایر گروه‌ها از نشان‌گرهای ایمونوهیستوشیمی متمایزی برخوردار هستند (۹). به عنوان مثال درگروه Basal-like، نشان‌گرهای HER2، PR و ER هر سه منفی هستند.

یکی از راه‌های تشخیص انواع سرطان، استفاده از داده‌های بیان ژن بیماران است. پیشرفت فناوری در حوزه بیوانفورماتیک، به خصوص فناوری ریزآرایه[5] باعث شده که داده‌های بیان ژنی[6] هزاران ژن مربوط به یک نمونه مبتلا به سرطان به طور همزمان استخراج شود (۱۰). البته این استخراج داده مستلزم هزینه‌های بالا است، به همین دلیل تعداد نمونه‌ها در مقایسه با تعداد ویژگی‌های[7] استخراج‌شده بسیار کم است. چنین داده‌هایی به‌اختصار HDLSS[8] نامیده می‌شود (۱۱) که در آن معمولاً تعداد نمونه‌ها کمتر از ۱۰۰ و تعداد ویژگی‌ها بین ۶۰۰۰ تا ۶۰۰۰۰ است (۱۲). تشخیص ژن‌های مؤثر در بیماری از داده‌های ریزآرایه، یک موضوع مهم در رابطه با ارتقا دانش در مورد مکانیزم بیماری و بهبود روش‌های مواجهه با بیماری است. مطالعات و روش‌های سنتی، تعداد زیادی از ژن‌ها را به عنوان ژن‌های مؤثر در بیماری در نظر می‌گیرند (۱۳). به‌دلیل این‌که روش‌های تجربی برای تشخیص ژن‌های مؤثر از میان تعداد زیاد ژن‌های نامزد بسیار هزینه‌براست، به‌کارگیری روش‌های انتخاب ویژگی (ژن) توصیه می‌شود. به‌علاوه وجود نویز[9] و تعداد بسیار بالای اطلاعات ژنی، تحلیل داده‌های ریزآرایه را به یک دامنه مهیج تبدل کرده است (۱۴). مطالعات متعددی نشان داده‌اند که تعداد بسیاری از داده‌های بیان ژن ریزآرایه DNA (ویژگی‌ها) از نظر معیارهای دسته‌بندی دارای بار اطلاعاتی نیستند (داده‌ی غیرمرتبط) (۱۵). بنابراین انتخاب ویژگی (ژن) و فرآیند تشخیص و حذف ویژگی‌های غیرمرتبط، نقشی حیاتی در تحلیل داده‌های ریزآرایه استخراج شده از DNA ایفا می‌کنند.

یافته‌های علمی به خصوص طی دو دهه اخیر ارتباط بین بیان غیرنرمال ژن و انواع سرطان‌ها را نشان می‌دهد (۱۶-۱۸). جهش[10]، بیان بیش از حد[11] و بیان کمتر از حد[12] ژن مصادیق بیان غیرنرمال ژن هستند. مارتین کورنا[13] و همکاران در (۱۹) ارتباط بین جهش سلولی با برخی سرطان‌ها را نشان دادند. البته باید توجه داشت که هر جهش منجر به سرطان نمی‌شود. زین العابدین و همکاران در (۲۰) دسته‌ای از ژن‌های با بیان کمتر از حد و دسته‌ای از ژن‌ها با بیان بیش از حد را برای تشخیص نوعی سرطان پوست به نام ملانوما[14] معرفی کردند. به نظر می‌رسد می‌توان از ارتباط بین بیان ژن و سرطان به عنوان دانش و به صورت مهندسی معکوس در مسئله انتخاب ژن در داده‌های ریزآرایه استفاده نمود که منجر به کاهش پیچیدگی مسئله خواهد شد.

مواد و روش‌ها

مطالعه حاضر توصیفی و داده محور است که به ارایه روشی برای تشخیص انواع سرطان پستان از دید مولکولی با استفاده از داده‌های بیان ژن بیماران سرطانی پرداخته است.

[1] Immunohistochemistry
[2] Estrogen receptor
[3] Progesterone receptor
[4] Human epidermal growth factor receptor 2
[5] Micro Array Data set
[6] Gene expression data
[7] Feature
[8] High Dimension Low Sample Size
[9] Noise
[10] Mutation
[11] Overexpression
[12] Underexpression or misexpression or less expression
[13] Martincorena
[14] Melanoma

توصیف مجموعه داده‌ها: داده‌های استفاده شده در این مطالعه، شامل داده‌های بیان۹۲۱۶ ژن مربوط به ۸۴ بیمار سرطانی است که توسط پراو[15] و همکاران با استفاده از فناوری ریزآرایه به دست آمده است (۷). داده‌ها شامل پنج کلاس (نوع) مختلف سرطان می‌باشد که کلاس ۱ با برچسب Luminal، کلاس۲ با برچسب ERBB2، کلاس۳ با برچسب Basal-like، کلاس۴ با برچسب Normal-like و کلاس۵ با برچسب Cell_lines مشخص شده است. مجموعه داده‌های مذکور به صورت دو دسته آموزش و آزمون در دسترس هستند که در۵۴ نمونه‌ی دسته آموزش به ترتیب ۲۰، ۶، ۷، ۹ و ۱۲ نمونه از هر کلاس وجود دارد و در۳۰ نمونه‌ی مربوط به دسته آزمون، نمونه‌ها به ترتیب ۱۲، ۳، ۳، ۵ و ۷ درکلاس‌های نامبرده قرار دارند.

روش پیشنهادی: در این بخش، از ارتباط بین بیان غیرنرمال ژن و سرطان به عنوان دانش، برای شناسایی و انتخاب ژن‌های موثر در تشخیص سرطان پستان استفاده می‌شود و یک روش انتخاب ویژگی برای انتخاب ژن‌های موثر در داده‌های ریزآرایه سرطان پستان پیشنهاد می‌شود که شامل دو بخش امتیازدهی ژن‌ها در هر کلاس و انتخاب ژن‌ها براساس امتیاز به‌دست آمده در هر کلاس است.

امتیازدهی ژن‌ها در هر کلاس: داده‌های بیان ژن ریزآرایه معرف کمیت توده mRNA مرتبط با نمونه هر یک از بیماران است (۱۲)، بنابراین هر تغییر در بیان ژن عامل بیماری باعث می‌شود مقادیر mRNA مرتبط با آن ژن نیز از حالت نرمال خارج شود. در نتیجه، مقادیر آن ژن در نمونه‌های مختلف از یک بیماری به‌هم نزدیک می‌شوند و در کنار هم یک پیوستگی محدود را در کلاس مربوط به آن بیماری تشکیل دهند. در این مطالعه برای کمیت‌سازی این پیوستگی‌ها رابطه زیر پیشنهاد می‌شود:

$$\mathrm{Score}\left(f_i|\mathrm{C}_j\right)=\sum_{k=1}^{\alpha}\frac{\mathrm{L}_k}{|\mathrm{C}_j|}\exp\left(\frac{\mathrm{L}_k}{|\mathrm{C}_j|}\right) \quad (۱)$$

که f_i ویژگی مربوط به بیان ژن iام است. با توجه به تعداد ژن‌ها، مقادیر i از ۱ تا ۹۲۱۶ متغیر است. C_j، کلاس jام سرطان را نشان می‌دهد که با توجه به ۵ کلاس سرطان، مقادیر j از ۱ تا ۵ تغییر می‌کند. $\mathrm{Score}\left(f_i|\mathrm{C}_j\right)$ میزان اهمیت و امتیاز ویژگی (ژن) iام در کلاس jام را نشان می‌دهد. α، تعداد افراز[16] کلاس C_j توسط ویژگی f_i است. منظور از افراز یک مجموعه، تقسیم آن به تعدادی زیرمجموعه است که اشتراک آن‌ها تهی و اجتماع آن‌ها برابر مجموعه اولیه باشد (۲۱). هریک از زیربخش‌های کلاس C_j شامل نمونه‌هایی است که از نظر مقداری کنار هم قرار دارند. L_k، تعداد نمونه‌های زیر بخش kام از زیربخش‌های افرازشده کلاس C_j است. هرچه عدد α کوچک‌تر باشد، کلاس C_j به زیربخش‌های بزرگ‌تر تقسیم می‌شود که به معنی وجود نمونه‌های بیشتر با خصوصیات مشترک است. بنابراین، طبق رابطه (۱) امتیاز بیشتری به ویژگی f_i تعلق می‌گیرد. $|\mathrm{C}_j|$ نیز معرف تعداد نمونه‌های کلاس jام است.

انتخاب ژن‌ها براساس امتیاز به‌دست آمده: بعد از مشخص شدن امتیاز ویژگی‌ها در هر کلاس، نوبت تصمیم‌گیری در مورد ترتیب نهایی ژن‌ها برای استفاده در مدل‌های رایج یادگیری مثل SVM[17] و KNN[18] و NB[19] است. یکی از ساده‌ترین راه حل‌ها این است که ویژگی‌های با ارزش بالاتر از هر کلاس به ترتیب در لیست خروجی قرار گیرند. در این صورت تاثیر ویژگی در سایر کلاس‌ها در نظر گرفته نمی‌شود. راه حل پیشنهادی در این مطالعه این است که با استفاده از رابطه (۲) امتیاز یک ویژگی در سایر کلاس‌ها نیز در محاسبه امتیاز نهایی ویژگی در کلاس مورد نظر لحاظ گردد و سپس ویژگی‌های با ارزش بالاتر از هر کلاس به ترتیب در لیست خروجی قرار گیرند.

$$\mathrm{Score}\left(f_i|\mathrm{C}_j\right)=\mathrm{Score}\left(f_i|\mathrm{C}_j\right)+\beta*\sum_{r\neq j}\mathrm{Score}\left(f_i|\mathrm{C}_r\right) \quad (۲)$$

ضریب β میزان مشارکت سایرکلاس‌ها در امتیازدهی را مشخص می‌کند. مقدار β نباید طوری انتخاب شود که تاثیر و نقش ویژگی در کلاس اصلی کمرنگ شود. از مقادیر منفی β، برای به‌دست آوردن ویژگی‌های متمایز و از مقادیر مثبت آن، برای به‌دست آوردن ویژگی‌های مشترک بین کلاس‌ها استفاده می‌شود. انتخاب مقدار

[15] Perou

[16] Partition
[17] Support Vector Machine
[18] K-nearest neighbor
[19] Naive Bayes

مناسب β، به ماهیت داده‌ها و نوع مدل یادگیری استفاده شده برای دسته‌بندی[۲۰] بستگی دارد.

برای پیاده‌سازی ایده‌های روش پیشنهادی شامل امتیازدهی و رتبه‌بندی نهایی ژن‌ها و استفاده از دسته‌بندهای NB و KNN، از نرم‌افزار MATLAB استفاده شده است. همچنین برای استفاده از دسته‌بند SVM از نرم‌افزار WEKA استفاده شده است. زمان متوسط اجرای بخش اول شامل امتیازدهی به ژن‌ها در هر کلاس و رتبه‌دهی نهایی به آن‌ها روی سیستم با مشخصات سرعت پردازنده ۱/۱ گیگا هرتز و حافظه اصلی ۴ گیگا بایت با سیستم عامل ویندوز ۱۰، برابر ۱/۲۵ ثانیه است. همچنین زمان متوسط اجرای بخش دوم شامل استفاده از دسته‌بندهای NB و KNN و SVM به ترتیب برابر با ۳۶/۷۸ ثانیه، ۲/۲۷ ثانیه و ۲۳/۶۵ ثانیه است.

یافته‌ها

پس از محاسبه امتیاز هر ویژگی با استفاده از روابط (۱) و (۲) و تعیین ترتیب نهایی ویژگی‌ها در لیست خروجی، از مدل‌های یادگیری رایج و پرکاربرد SVM و KNN و NB، برای بررسی کارایی روش پیشنهادی بر اساس معیارهای دقت دسته‌بندی و تعداد ژن‌های انتخابی، استفاده شد و برای محدود کردن تعداد آزمایش‌ها برای انتخاب تعداد ویژگی‌های بهینه، حداکثر تعداد ژن‌های انتخابی برابر ۱۰۰ در نظر گرفته شد. در آزمایش اول، مقادیر ژن ابتدایی لیست (ژن دارای رتبه اول) و در آزمایش دوم، مقادیر دو ژن ابتدایی لیست و در آزمایش k ام (1<=k<=100) ، مقادیر k ژن ابتدایی لیست از ۵۴ نمونه داده‌های آموزش برای آموزش هر یک از دسته‌بندها، استفاده می‌شود و هر بار برای به‌دست آوردن دقت دسته‌بندی، مقادیر ژن‌های متناظر از ۳۰ نمونه‌ی مربوط به داده‌های آزمون به دسته‌بند داده می‌شود. و در پایان حداکثر دقت به‌دست آمده به همراه تعداد ژن‌های متناظر ثبت می‌شود. دقت یک روش دسته‌بندی، درصد نمونه‌های دسته‌بندی‌شده درست را در میان تمام نمونه‌ها نشان می‌دهد. نتایج به‌دست آمده بر اساس مقادیر مختلف β، در جدول‌های (۱) و (۲) نشان داده شده است.

جدول۱: عملکرد روش پیشنهادی با توجه به معیار دقت

دسته‌بند	$\beta = 0$	$\beta = 0.01$	$\beta = 1$	$\beta = -0.2$
NB	۸۶/۶۷	۹۰	۹۰	۸۳/۳۳
SVM	۹۶/۶۷	۹۶/۶۷	۸۶/۶۷	۹۳/۳۳
KNN	۹۶/۶۷	۱۰۰	۸۶/۶۷	۸۳/۳۳

جدول۲: عملکرد روش پیشنهادی با توجه به معیار تعداد ژن

دسته‌بند	$\beta = 0$	$\beta = 0.01$	$\beta = 1$	$\beta = -0.2$
NB	۶۰	۳۹	۱۷	۲۶
SVM	۴۵	۲۲	۳۰	۴۱
KNN	۴۳	۳۸	۵۱	۳۰

توجه جدول‌های (۱) و (۲)، در مجموع مدل‌های یادگیری SVM و KNN روی ژن‌های انتخابی عملکرد بهتری نسبت به مدل یادگیری NB دارند. در بهترین وضعیت با استفاده از مدل یادگیری KNN و با تعداد ۳۸ ژن انتخابی، نوع سرطان در کلیه نمونه‌های آزمایش به درستی تشخیص داده شد. همچنین دقت قابل قبول ۹۶/۶۷ با تعداد ۲۲ ژن در مدل SVM به‌دست آمد. جدول‌های (۳) و (۴) کارایی روش پیشنهادی را در مقایسه با سه روش انتخاب ویژگی معروف Relief (۲۲) و ARCO (۲۳) و FAST (۲۴) در شرایط یکسان (ماکزیمم ۱۰۰ ژن انتخابی و دسته‌بندهای مشابه) نشان می‌دهد. مقایسه نتایج بر اساس میانگین خطاها در هر کلاس صورت گرفته است.

با توجه به جدول (۳)، تنها روش ARCO با دسته‌بند SVM و روش پیشنهادی در این مطالعه با دسته‌بند KNN دارای خطای صفر یا دقت ۱۰۰٪ هستند، با این تفاوت که با توجه به جدول (۴) تعداد ژن‌های انتخابی روش پیشنهادی با ۳۸ ژن کمتر از ۵۲ ژن روش ARCO است.

نکته مهم در مورد روش پیشنهادی نسبت به سایر روش‌ها، توانایی تفکیک ژن‌های موثر در هر کلاس است که می‌تواند کمک شایانی به فعالیت‌های آزمایشگاهی و داروسازی کند. جدول (۵) چگونگی توزیع ۳۸ ژن انتخابی در کلاس‌های مختلف سرطان پستان را نشان می‌دهد.

[20] Classification

جدول۳: حداقل میانگین خطا در حداکثر ۱۰۰ ژن انتخابی

دسته‌بند	Proposed method	ReliefF*	ARCO*	FAST*
NB	۰/۶۶	۰/۱۶	۰/۰۱۴	۰/۱۴۴
SVM	۰/۰۴	۰/۱۴۳	۰	۰/۰۵۲
KNN	۰	۰/۰۷۴	۰/۰۱۹	۰/۰۸۱

* نتایج گزارش شده در (۲۵)

جدول۴: تعداد ژن‌های انتخابی هر روش با حداقل میانگین خطا

دسته‌بند	Proposed method	ReliefF*	ARCO*	FAST*
NB	۱۷	۸۲	۶۸	۶۳
SVM	۲۲	۵۷	۵۲	۷۹
KNN	۳۸	۹۷	۵۵	۶۲

* نتایج گزارش شده در (۲۵)

جدول ۵: توزیع ۳۸ ژن انتخابی با استفاده از روش پیشنهادی در کلاس‌های مختلف سرطان پستان

کلاس سرطان	شماره ژن							
Luminal	G9111	G7797	G868	G7961	G6797	G8506	G3150	G4450
ERBB2	G8921	G886	G7133	G1054	G6595	G1379	G2188	
Basal-like	G6802	G4380	G6623	G2452	G4211	G6410	G6243	G4805
Normal-like	G1280	G74	G1756	G322	G5342	G6606	G2951	G6554
Cell-lines	G4919	G8949	G6385	G5651	G8137	G2030	G7625	

همان‌گونه که در جدول (۵) مشخص است، روش پیشنهادی توانسته است ژن‌های موثر در هر کلاس را به صورت جداگانه شناسایی کند. علاوه براین، با بررسی بیشتر ژن‌های موثر در هر کلاس، ژن‌های با بیان بیشتر یا کمتر از حد نرمال در هر کلاس قابل تشخیص است. به عنوان مثال ژن G9111 در ۸۵٪ نمونه‌های آموزش مربوط به سرطان نوع Luminal یعنی ۱۷ نمونه از ۲۰ نمونه، بیان بیش از حد داشته است.

در مورد ژن G1054، در ۵ نمونه از ۶ نمونه مربوط به سرطان نوع ERBB، بیان ژنی بیشتر نسبت به سایر نمونه‌ها مشاهده گردید. در سرطان نوع Basal-like، ۶ نمونه از ۷ نمونه داده‌های آموزش در ژن G2452، مقادیر بیان ژنی بالاتر نسبت به سایر نمونه‌های انواع دیگر سرطان داشتند. ژن‌های G1280 و G74 در کلیه نمونه‌های کلاس نوع Normal-like بیان بیش از حد نرمال داشته‌اند. در مورد کلاس Cell-lines، کلیه‌ی ژن‌های مشخص شده، بیان ژن کمتر از حد نرمال در ۱۲ نمونه مربوطه داشتند.

بحث

در روش پیشنهادی در این مطالعه، تاثیر هر ژن در هریک از کلاس‌های سرطان پستان به طور جداگانه بررسی شد. امتیاز هر ژن در هر کلاس بر اساس چگونگی افراز اعضای آن کلاس توسط ژن مورد نظر، از طریق رابطه (۱) محاسبه شد. امتیاز به‌دست آمده برای در نظر گرفتن تاثیر ژن‌های مشترک، توسط رابطه (۲) اصلاح شد و در نهایت، ژن‌های موثر در هر کلاس به ترتیب امتیاز در لیست خروجی ژن‌های انتخابی قرار گرفتند. سپس از سه دسته‌بند SVM و KNN و NB، برای بررسی کیفیت ژن‌های انتخابی در تشخیص انواع سرطان پستان، استفاده شد. نتایج مدل‌سازی نشان داد که با استفاده از دسته‌بند KNN و فقط ۳۸ ژن از ۹۲۱۶ ژن اولیه، کلیه‌ی نمونه‌های سرطان پستان از مجموعه داده‌های آزمایش به درستی تشخیص داده شد. در این بخش کارایی روش پیشنهادی بر اساس معیارهای دقت دسته‌بندی و تعداد ژن‌های انتخابی، با تعدادی از روش‌های مرتبط که آن‌ها نیز روی مجموعه داده‌های مورد بحث این مقاله اعمال شده‌اند، مقایسه می‌شود.

در سال ۲۰۰۵، تان[21] و همکاران در (۲۶) روش‌های HC-TSP و HC-K-TSP را بر مبنای مقایسه نسبی مقادیر جفت ژن‌ها ارایه کردند که توسعه یافته روش TSP[22] (۲۷) برای استفاده در دسته‌بندی‌های چندکلاسه است. آن‌ها با استفاده از روش‌های فوق به ترتیب با انتخاب ۸ و ۲۴ ژن، به دقت ۶۶/۶۷٪ دست یافتند که هرچند از نظر تعداد ژن عملکرد بهتری را نسبت به ۳۸ ژن انتخابی روش پیشنهادی نشان می‌دهد ولی از نظر دقت، فاصله زیادی تا دقت ۱۰۰٪ دارد. تان و همکاران همچنین با نرم‌افزار PAM[23] (۲۸) که یک ابزار تحلیل آماری رایج برای کار روی داده‌های ریزآرایه است، با ۴۸۲۲ ژن انتخابی، انواع سرطان سینه را با دقت ۹۳/۳۳ درصد پیش‌بینی کردند که با وجود بهبود دقت دسته‌بندی، با تعداد ۴۸۲۲ ژن انتخابی عملکرد مناسبی ندارد.

ونگ[24] و همکاران در (۲۹)، الگوریتم TSG[25] را ارایه دادند که بر اساس انتخاب مجموعه‌هایی از k زوج ویژگی با امتیاز بالاتر براساس معیار chi-square و ارزیابی هر یک از مجموعه ویژگی‌ها بر پایه‌ی LOOCV[26]، به دقت ۸۶/۶۷٪ با تعداد ۶۳ ژن دست یافتند که در هر دو پارامتر دقت دسته‌بندی و تعداد ژن انتخابی، عملکرد ضعیف‌تری نسبت به روش پیشنهادی دارد.

لی سان[27] و همکاران (۲۵) در سال ۲۰۱۷ یک روش انتخاب ویژگی به نام AVC را بر اساس منحنی ROC[28] و خاصیت مکمل‌پذیری ارایه کردند. آن‌ها از چهار دسته‌بند SVM و C4.5 و KNN و NB برای ارزیابی ۷۲ ژن انتخاب شده توسط AVC استفاده کردند. نتایج گزارش شده توسط آن‌ها با استفاده از چهار دسته‌بند ذکرشده، به ترتیب خطای میانگین ۰/۰۰۶ و ۰/۱۳۶ و ۰/۰۱۹ و ۰/۰۷۳ درصد را در تشخیص انواع سرطان پستان نشان داد که با وجود دستیابی به دقت نزدیک ۱۰۰ درصد، تعداد ۷۲ ژن انتخابی بیشتر از ۳۸ ژن روش پیشنهادی است.

ژو و همکاران در (۳۰)، با بهره‌گیری از مفهوم آنتروپی و مبانی تئوری اطلاعات، روشی به نام IK-TSP را که توسعه یافته روش K-TSP (۲۶) برای استفاده در مسائل دسته‌بندی چند کلاسه است، ارایه کردند که با ۵۶ ژن انتخابی، انواع سرطان سینه را با دقت ۸۳/۳۳٪ پیش‌بینی کردند.

چن[29] و همکاران در (۳۱) ، روشی با نام RS-based[30] DC ارایه دادند که در آن ابتدا مسئله دسته‌بندی چند کلاسه را بر اساس روش OVR[31]، به تعدادی مسئله باینری تبدیل کردند. سپس معیارهایی بر اساس مبانی تئوری اطلاعات و آنتروپی برای ارزیابی افقی و عمودی مقادیر یک ژن نسبت به سایر ژن‌ها ارایه دادند که بر اساس آن‌ها در کنار معیاری برای ارزیابی نسبی مقادیر درون ژنی، امتیاز ژن‌ها را تعیین کردند. چن و همکاران با این روش با تعداد ۱۵ ژن انتخابی به دقت ۹۳/۳۳٪ در تشخیص درست سرطان پستان دست یافتند که از نظر تعداد ژن انتخابی نسبت به روش پیشنهادی عملکرد بهتری دارد ولی در مقایسه با دقت ۱۰۰ درصدی روش پیشنهادی، کارایی پایین‌تری دارد.

ونگ[32] و وی[33] (۳۲) در سال ۲۰۱۷ دو روش انتخاب ویژگی DSCFS و CAM را بر اساس اندازه‌گیری توانایی یک ویژگی در دسته‌بندی زیر مسایل دو کلاسه و مفهوم مکمل بودن ویژگی‌ها ارایه کردند. آن‌ها با ۶ ژن انتخاب شده توسط روش DSCFS و با استفاده از سه دسته‌بند NB و SVM و KNN به ترتیب به دقت‌های ۹۰ و ۹۶/۶۷ و ۹۶/۶۷ درصد دست یافتند. آن‌ها همچنین با ۱۵۲ ژن انتخاب شده توسط روش CAM و با استفاده از سه دسته‌بند NB و SVM و KNN به ترتیب دقت‌های ۸۰ و ۹۶/۶۷ و ۹۰ درصد را گزارش کردند. با وجود برتری روش DSCFS در تعداد ژن انتخابی، روش پیشنهادی با توجه به معیار دقت، عملکرد بهتری دارد.

سان[34] و همکاران (۳۳) در سال ۲۰۱۹ روشی به نام ECOC-MDC برای تبدیل مسئله دسته‌بندی چندکلاسه به دسته‌بندی‌های دوکلاسه بر اساس مفاهیم

[21] Tan
[22] Top scoring pair
[23] Prediction analysis of microarrays
[24] Wang
[25] Top scoring genes
[26] Leave-one-out cross validation
[27] Wang
[28] Receiver Operator Characteristic

[29] Chen
[30] Relative simplicity based direct classification
[31] One versus rest
[32] Wang
[33] Wei
[34] sun

پیچیدگی داده‌ها ارایه دادند. آن‌ها از سه روش انتخاب ویژگی ROC و T-test و Wilcoxon برای انتخاب ژن‌های با ارزش‌تر استفاده کردند. آن‌ها با استفاده از روش ECOC-MDC و دسته‌بند NB و با ۱۰۰ ژن انتخابی الگوریتم ROC، با دقت ۹۶٪ انواع سرطان پستان را تشخیص دادند. همچنین با دسته‌بند SVM و با ۱۰۰ ژن انتخابی الگوریتم T-test، به دقت ۹۷/۳٪ و با دسته‌بند SVM و با ۱۰۰ ژن انتخابی الگوریتم Wilcoxon، به دقت ۹۸/۷ دست یافتند.

در مجموع دست یافتن به دقت ۱۰۰ درصدی، نقطه قوت روش پیشنهادی محسوب می‌شود. هر چند از نظر تعداد ژن‌های انتخابی، برخی روش‌های ذکر شده عملکرد بهتری داشتند ولی با در نظر گرفتن همزمان دو پارامتر دقت دسته‌بندی و تعداد ژن انتخابی، روش پیشنهادی عملکرد قابل قبولی نسبت به سایر روش‌ها دارد.

علاوه بر این، با توجه به این‌که امتیاز هر ژن در هر کلاس به صورت مجزا تعیین می‌شود، می‌توان در صورت نیاز از این روش برای یافتن ژن‌های مشترک بین کلاس‌های بیماری استفاده کرد. به این ترتیب که ژن‌هایی که دارای امتیاز بالا در دو یا چند کلاس باشند، می‌توانند به عنوان ژن‌های نامزد برای مشخص شدن ژن‌های موثر مشترک، مورد مطالعه قرار گیرند.

نتیجه‌گیری

در این مطالعه، از ارتباط بین بیان غیرنرمال ژن و سرطان به عنوان دانش در مسئله انتخاب ژن در داده‌های ریزآرایه استفاده شد که باعث گردید روش پیشنهادی علاوه بر تشخیص ۱۰۰ درصدی انواع سرطان پستان، قادر به تفکیک ژن‌های انتخابی باشد. توانایی تفکیک ژن‌ها در هر کلاس سرطان به‌علاوه تشخیص بیان بیشتر از حد یا کمتر از حد ژن‌های انتخابی، از خصوصیات ویژه روش پیشنهادی است که می‌تواند مورد استفاده متخصصین و پژوهشگران حوزه تشخیص و درمان بیماری سرطان پستان قرار گیرد.

تقدیر و تشکر

این پژوهش نتیجه بخشی از فعالیت‌های صورت گرفته در قالب رساله دکتری با عنوان "انتخاب ویژگی مبتنی بر تئوری اطلاعات برای انتخاب ژن‌های موثر در داده‌های ریزآرایه" است که در تاریخ ۱۳۹۶/۱۱/۱۱ در شورای تحصیلات تکمیلی دانشگاه یزد با شماره ۳۹۶۹۰ به تصویب رسیده است. از کلیه همکاران گرامی آزمایشگاه هوش محاسباتی دانشگاه یزد که در انجام این پژوهش همراهی نمودند، تشکر و قدردانی می‌شود.

تعارض منافع

نویسندگان اعلام می‌دارند که هیچ تعارض منافعی در پژوهش حاضر وجود ندارد.

References

1. Smith RA, Andrews KS, Brooks D, Fedewa SA, Manassaram-baptiste D, Saslow D, et al. Cancer Screening in the United States, 2018: A Review of Current American Cancer Society Guidelines and Current Issues in Cancer Screening. CA Cancer J Clin. 2018;68(4):297-316.
2. Enayat R, Salehiniya H. An investigation of changing patterns in breast cancer incidence trends among Iranian women. J Sabzevar Univ Med Sci. 2015;22(1):27-35.
3. Wang YA, Johnson SK, Brown BL, McCarragher LM, Al-Sakkaf K, Royds JA, et al. Enhanced anti-cancer effect of a phosphatidylinositol-3 kinase inhibitor and doxorubicin on human breast epithelial cell lines with different p53 and oestrogen receptor status. Int J Cancer. 2008;123(7):1536–44.
4. Sheikhpour R, Ghassemi N, Yaghmaei P, Ardekani JM, Shiryazd M. Immunohistochemical assessment of P53 protein and its correlation with clinicopathological characteristics in breast cancer patients. Indian J Sci Technol. 2014;7(4):472-9.
5. Vallejos CS, Gómez HL, Cruz WR, Pinto JA, Dyer RR, Velarde R, et al. Breast Cancer Classification According to Immunohistochemistry Markers: Subtypes and Association With Clinicopathologic Variables in a Peruvian Hospital Database. Clin Breast Cancer. 2010;10(4):294-300.
6. Cheang MCU, Chia SK, Voduc D, Gao D, Leung S, Snider J, et al. Ki67 index, HER2 status, and

prognosis of patients with luminal B breast cancer. JNCI J Natl Cancer Inst. 2009; 101(10): 736-50.

7. Perou CM, Sørlie T, Eisen MB, Van De Rijn M, Jeffrey SS, Rees CA, et al. Molecular portraits of human breast tumours. Nature. 2000; 406(6797): 747-52.

8. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci. 2001; 98(19):10869-74.

9. Dai X, Li T, Bai Z, Yang Y, Liu X, Zhan J, et al. Breast cancer intrinsic subtype classification, clinical use and future trends. Am J Cancer Res. 2015; 5(10):2929-43.

10. Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges. ACM SIGKDD Explor Newsl. 2003; 5(2):1-5.

11. Zhang L, Lin X. Some considerations of classification for high dimension low-sample size data. Stat Methods Med Res. 2013; 22(5):537-50.

12. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003; 3(3):1157-82.

13. Glazier AM. Finding Genes That Underlie Complex Traits. Science. 2002; 298(5602): 2345-9.

14. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007; 23(19):2507-17.

15. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science. 1999; 286(5439):531-7.

16. Tsafrir D, Bacolod M, Selvanayagam Z, Tsafrir I, Shia J, Zeng Z, et al. Relationship of gene expression and chromosomal abnormalities in colorectal cancer. Cancer Res. 2006; 66(4): 2129-37.

17. Prelich G. Gene overexpression: uses, mechanisms, and interpretation. Genetics. 2012; 190(3): 841-54.

18. Gray JW, Collins C. Genome changes and gene expression in human solid tumors. Carcinogenesis. 2000; 21(3):443-52.

19. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science (80). 2015; 349(6255):1483-9.

20. Zainulabadeen A, Yao P, Zare H. Underexpression of specific interferon genes is associated with poor prognosis of melanoma. PLoS One. 2017; 12(1):e0170025.

21. Pinter CC. A Book of SET THEORY. Dover Publications; 2014.

22. Zhao Z, Wang L, Liu H, Ye J. On Similarity Preserving Feature Selection. IEEE Trans Knowl Data Eng. 2013; 25(3):619-32.

23. Wang R, Tang K. Feature Selection for Maximizing the Area Under the ROC Curve. In: 2009 IEEE International Conference on Data Mining Workshops. 2009; 400-5.

24. Chen X, Wasikowski M. FAST: A Roc-based Feature Selection Metric for Small Samples and Imbalanced Data Classification Problems. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2008; 124-32.

25. Sun L, Wang J, Wei J. AVC : Selecting discriminative features on basis of AUC by maximizing variable complementarity. BMC bioinformatics. 2017; 18(3):50.

26. Tan AC, Naiman DQ, Xu L, Winslow RL, Geman D. Simple decision rules for classifying human cancers from gene expression profiles. Bioinformatics. 2005; 21(20):3896-904.

27. Geman D, d'Avignon C, Naiman DQ, Winslow RL. Classifying gene expression profiles from pairwise mRNA comparisons. Stat Appl Genet Mol Biol. 2004; 3(1):1-19.

28. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc Natl Acad Sci. 2002; 99(10):6567-72.

29. Wang H, Zhang H, Dai Z, Chen M, Yuan Z. TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. BMC Med Genomics. 2013; 6(1):S3.

30. Zhou C, Wang S, Blanzieri E, Liang Y. An entropy-based improved k-top scoring pairs (TSP) method for classifying human cancers. African J Biotechnol. 2012;11(45):10438-45.

31. Chen Y, Wang L, Li L, Zhang H, Yuan Z. Informative gene selection and the direct classification of tumors based on relative simplicity. BMC Bioinformatics. 2016;17(1):44.

32. Wang S, Wei J. Feature selection based on measurement of ability to classify subproblems. Neurocomputing. 2017; 224:155-65.

33. Sun M, Liu K, Wu Q, Hong Q, Wang B, Zhang H. A novel ECOC algorithm for multiclass microarray data classification based on data complexity analysis. Pattern Recognit. 2019; 90: 346-62.