

تحلیل تصاویر ریزآرایه به منظور تشخیص نوع سرطان سینه

نسترن دهقان: دانشجوی ارشد هوش مصنوعی، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران. dehghan_nastaran@shahroodut.ac.ir
***حمید حسن پور:** استاد و متخصص پردازش سیگنال، دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران (*نویسنده مسئول). h.hassanpour@shahroodut.ac.ir
محمد رضا عباس زادگان: استاد و متخصص ژنتیک انسانی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران. abbaszadeganmr@mums.ac.ir

تاریخ پذیرش: ۹۵/۹/۱۵

تاریخ دریافت: ۹۵/۶/۲۷

چکیده

زمینه و هدف: فناوری ریزآرایه به عنوان ابزاری بسیار قدرتمند جهت مطالعه و تحلیل رفتار هزاران ژن به طور همزمان می‌باشد. تصاویر حاصل از فناوری ریزآرایه نقش مهمی در کشف و درمان بیماری‌ها دارند. هدف از این پژوهش، ارائه یک روش خودکار جهت استخراج و تحلیل داده‌ها از تصاویر ریزآرایه به منظور تشخیص بیماری‌های سرطانی می‌باشد.

روش کار: سیستم پیشنهادی شامل سه فاز اصلی پردازش تصویر، داده‌کاوی و شناسایی بیماری می‌باشد. در فاز پردازش تصویر عملیاتی مانند شناسایی مکان ژن‌ها، حذف پس‌زمینه و استخراج داده‌های خام از تصویر انجام می‌گیرد. فاز دوم شامل نرمال‌سازی داده‌های استخراج شده و انتخاب ژن‌های مؤثرتر می‌باشد. در فاز سوم، با توجه به داده‌های استخراج شده عمل شناسایی و تشخیص بیماری انجام می‌گیرد.

یافته‌ها: در این پژوهش مجموعه تصاویر ریزآرایه سرطان سینه از پایگاه داده دانشگاه استنفورد مورد استفاده قرار گرفته است. دقت روش پیشنهادی جهت تعیین مکان ژن‌ها و تشخیص نوع سرطان سینه به ترتیب بالغ بر ۹۸٪ و ۹۵٫۴۵٪ می‌باشد.

نتیجه‌گیری: نتایج به دست آمده از این روش نسبت به دیگر روش‌های موجود در زمینه تحلیل تصاویر و داده‌های ریزآرایه از دقت بالاتری برخوردار است؛ همچنین نسبت به آزمایش‌های بیولوژی از دسترسی آسان‌تر و هزینه کمتری برخوردار است.

کلیدواژه‌ها: ریزآرایه، پردازش تصویر، داده‌کاوی، سرطان سینه

مقدمه

معنی شناسایی موقعیت هر خال در تصویر می‌باشد. از مهم‌ترین روش‌های شبکه‌بندی می‌توان به روش استفاده از ماشین بردار پشتیبان (Support Vector Machine-(SVM)) اشاره کرد (۱). در این روش خال‌های شناسایی شده توسط تکنیک‌های لبه‌یابی به عنوان داده‌ی ورودی به یک ماشین بردار پشتیبان اعمال می‌شوند. خط جدا کننده، فاصله‌ی مابین سطر و ستون خال‌ها را تخمین می‌زند. در صورت عدم‌شناسایی موقعیت تقریبی خال‌ها، روش بیان شده به درستی عمل نمی‌کنند. در پژوهشی دیگری (۲) برای شناسایی موقعیت هر خال، از محاسبه هیستوگرام تصویر به کمک تکنیک‌های حد‌آستانه استفاده شده است. این روش برای تصاویر نویزیاز دقت پایینی برخوردار است. همچنین در (۳) فرض می‌شود که

در سال‌های اخیر، استفاده از فناوری ریزآرایه (Microarray) امکان بررسی هزاران ژن را به طور همزمان فراهم کرده است. به طور ساده "ریزآرایه" عبارت است از فناوری بررسی ده‌ها، صدها و هزاران ژن می‌باشد. به هر ژن بر روی تراشه ریزآرایه یک خال گفته می‌شود. فناوری ریزآرایه به دو بخش کلی انجام آزمایش بالینی و تحلیل تصاویر حاصل از آزمایش تقسیم می‌شود. این دو بخش لازم و ملزوم یکدیگر هستند. تحلیل تصاویر ریزآرایه شامل سه فاز اصلی پردازش تصویر، داده‌کاوی و شناسایی بیماری می‌باشد. در فاز پردازش تصویر، کمی‌سازی تصاویر ریزآرایه در سه مرحله شبکه‌بندی (Gridding)، قطعه‌بندی و استخراج داده انجام می‌شود. شبکه‌بندی تصویر به

می‌کنند. در شرایطی که تصاویر دارای خرابی‌هایی مانند نویز و زاویه چرخش نباشند، این نرم‌افزارها از کارایی خوبی برخوردار خواهند بود. از مهم‌ترین این نرم‌افزارها می‌توان GenePix، ScanAlyze و ImaGene اشاره نمود (۵، ۹، ۱۰).

با کمی‌سازی داده‌ها در فاز پردازش تصاویر، حجم زیادی از ویژگی‌ها (ژن) استخراج می‌شود. بنابراین در فاز دوم باید قبل از عمل دسته‌بندی با کمک تکنیک‌های داده‌کاوی ژن‌های مفید و تاثیرگذار از بین هزاران ژن انتخاب شوند؛ تا بتوان عملیات دسته‌بندی را بهتر انجام داد. در فاز سوم، دسته‌بندی (Classification) داده‌های ریزآرایه جهت تشخیص نوع سرطان، با استفاده از الگوریتم‌های موجود در یادگیری ماشین انجام می‌شود که به منظور تعیین مرحله‌ی درمانی بیمار یا تجویز دارو نقش به‌سزایی دارد. Golub و همکارانش در سال ۱۹۹۹ به عنوان پیشگامان طبقه‌بندی سرطان به وسیله داده‌های بیان ژن ریزآرایه، احتمال تشخیص سرطان را با روش‌های آماری نشان دادند (۱۱). همچنین در (۱۲) با استفاده از ماشین بردار پشتیبان توانستند نمونه‌های سالم از توموری را از هم مجزا کنند. در پژوهش دیگر با استفاده از الگوریتم k نزدیک‌ترین همسایه به صورت فازی داده‌های ریزآرایه خود را دسته‌بندی کرده‌اند (۱۳). برخی محققین برای بهبود نتایج، از ترکیب دسته‌بندی کننده‌های متفاوت استفاده می‌کنند، همانند (۱۴) که با ادغام دو روش بردار ماشین پشتیبان و شبکه عصبی فازی داده‌های خود را دسته‌بندی می‌کنند.

هدف از این تحقیق، ارائه یک سیستم خودکار جهت استخراج اطلاعات از تصاویر ریزآرایه و دسته‌بندی داده‌ها به منظور تشخیص نوع سرطان سینه می‌باشد. در این سیستم در گام پردازش تصویر با اعمال تکنیک‌های آستانه‌گذاری نظیر اتسو (Otsu) موقعیت خال‌ها شناسایی می‌شود (۱۵). سپس با بکارگیری روش منظم‌سازی تغییرات کلی (Total Variation Regularization (TV)) و حداقل‌سازی نرم $L1$ (۱۶) میزان پس‌زمینه هر خال شناسایی و از تصویر اصلی حذف می‌گردد و بدین ترتیب خال‌ها قطع‌بندی

تمامی خال‌ها در فواصل برابری از یکدیگر قرار گرفته‌اند. در این روش ابتدا تصاویر ورودی به تصاویر دودویی تبدیل می‌شوند، سپس مجموع سطرها / ستون‌های آن محاسبه می‌شود. مقادیر نزدیک به صفر نماینده فضای خالی مابین دو خال می‌باشند. میانگین این فواصل خالی، به عنوان فاصله مابین تمامی خال‌ها در نظر گرفته می‌شود. این روش نیز از دقت پایینی برخوردار است. در مرحله قطع‌بندی، تصاویر به دو دسته پیش‌زمینه (Foreground) و پس‌زمینه (Background) تقسیم می‌شوند. ناحیه‌ی اول شامل نقاط متعلق به خال‌ها بوده و ناحیه‌ی دوم شامل فضای موجود در اطراف خال‌ها می‌باشد. مشهورترین تکنیک‌های قطع‌بندی شامل: قطع‌بندی براساس دایره ثابت (Fixed Circle Segmentation)، قطع‌بندی براساس شکل وقفی (Adaptive Shape Segmentation)، قطع‌بندی براساس هیستوگرام (Segmentation Histogram) و قطع‌بندی براساس دایره وقفی (Adaptive Circle Segmentation) می‌باشند. این روش‌ها در نرم‌افزارهای ScanAlyze و GenePix مورد استفاده قرار می‌گیرند (۴، ۵). در برخی از پژوهش‌ها با استفاده از تکنیک‌های خوشه‌بندی، عمل قطع‌بندی خال‌ها صورت می‌گیرد؛ به عنوان نمونه در (۶) با در نظر گرفتن دو خوشه خال‌های موجود در تصویر به روش K -means قطع‌بندی می‌شوند. روش مذکور برای تصاویری که دارای نویز بالایی هستند یا خال‌های موجود، شدت روشنایی پایینی دارند به خوبی عمل نمی‌کند. در یک پژوهش انجام شده در (۷) از سه خوشه برای قطع‌بندی استفاده می‌شود. خوشه‌ای که کمترین مقدار را دارد به عنوان پس‌زمینه و دو خوشه دیگر به عنوان پیش‌زمینه تعریف می‌شوند. در گام آخر از پردازش تصویر، میزان روشنایی متناسب با سطح خاکستری هر یکاز خال‌ها استخراج و سطح بیان آن‌ها با مقایسه دو نمونه در کانال سبز و قرمز از تصویر (نمونه سرطانی و سالم) محاسبه می‌شود (۸). چندین نرم‌افزار برای کمی‌سازی تصاویر ریزآرایه به بازار عرضه شده است که تمامی آن‌ها به صورت نیمه‌خودکار عمل

می باشد و در قالب عدد صحیح دو بایتی نمایش داده می شوند. هر نمونه از تصاویر شامل ۴۱۷۶۰ ژن می باشد. این پایگاه داده شامل ۷۱ تصویر ریزآرایه می باشد که ۲۲ نمونه آن مربوط به سرطان سینه خوش خیم می باشند (۱۸) و مابقی داده ها مربوط به نمونه های سرطانی است که تومور به صورت موضعی و حاد پیشرفت داشته است (۱۹). در سیستم طراحی شده داده های سرطان سینه به دو دسته آموزش و آزمایش تقسیم می شوند. از این مجموعه داده ۴۹ (تقریباً ۷۰٪) نمونه به عنوان مجموعه آموزش و ۲۲ (تقریباً ۳۰٪) نمونه به عنوان مجموعه آزمایش در نظر گرفته می شود.

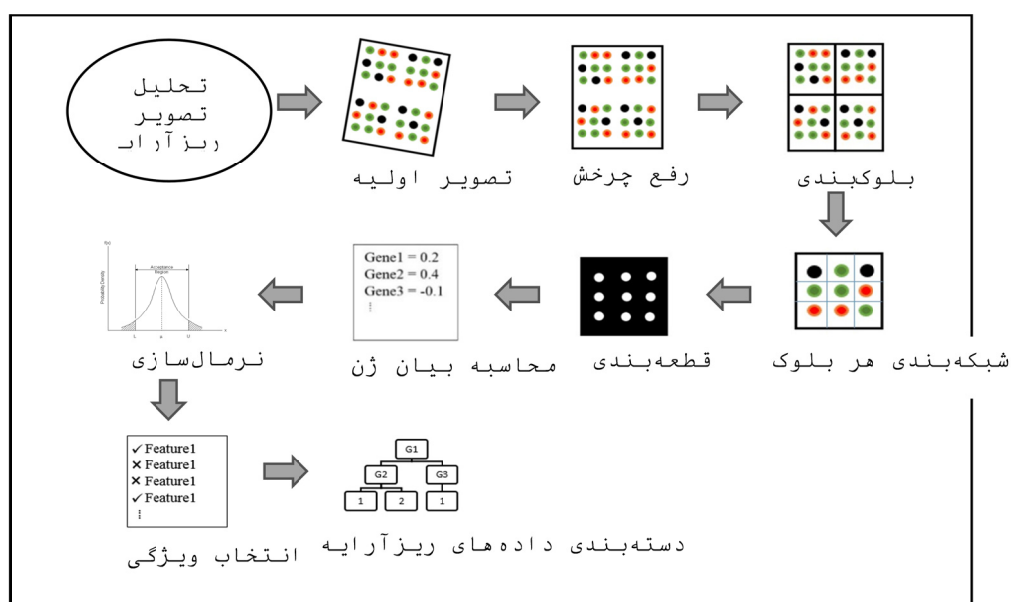
در این قسمت، سیستمی برای کمی سازی تصاویر ریزآرایه، تحلیل و دسته بندی داده های خام حاصل از این تصاویر ریزآرایه با استفاده از تکنیک های پردازش تصویر، داده کاوی و یادگیری ماشین ارائه می شود. ورودی این سیستم، تصویر ریزآرایه و خروجی آن دسته بندی نوع تومور سرطان سینه در آزمون بیمار می باشد (شکل ۱). برخی از تصاویر ریزآرایه دارای چرخش می باشند؛ بنابراین در مرحله اول میزان چرخش تصویر با استفاده از روش تبدیلات رادن (Transform Radon) تشخیص و اصلاح لازم بر

می شوند. در نهایت با محاسبه ی مقادیر شدت روشنایی هر خال و مقایسه ژن ها در هر نمونه (نمونه سالم و سرطانی) میزان بیان آن ها محاسبه می گردد. برای انجام دسته بندی نوع سرطان، ابتدا از بین ژن های موجود ۱۰۰ ژن مؤثرتر با کمک روش بهره اطلاعاتی مشخص می شوند. در نهایت این سیستم با تشکیل یک درخت تصمیم J48 برای دسته بندی و شناسایی نوع سرطان سینه در دو زیرکلاس می تواند اطلاعات به دست آمده را به صورت خودکار تحلیل نماید.

در این مقاله ابتدا در بخش دوم روشی برای تحلیل تصاویر ریزآرایه معرفی می گردد. سپس یافته های حاصل از پژوهش در بخش سوم بررسی و تحلیل می شوند. بخش پنجم نیز به بحث و نتیجه گیری کلی در مورد روش پیشنهادی پرداخته می شود.

روش کار

مطالعات صورت گرفته در این پژوهش مربوط به تصاویر ریزآرایه سرطان سینه تهیه شده از دانشگاه استنفورد می باشد (۱۷). هر تصویر رنگی ریزآرایه، به دو تصویر خاکستری مربوط به کانال قرمز و سبز تقسیم می شود. کانال قرمز نماینده ژن های بیمار و کانال سبز نماینده ژن های فرد سالم

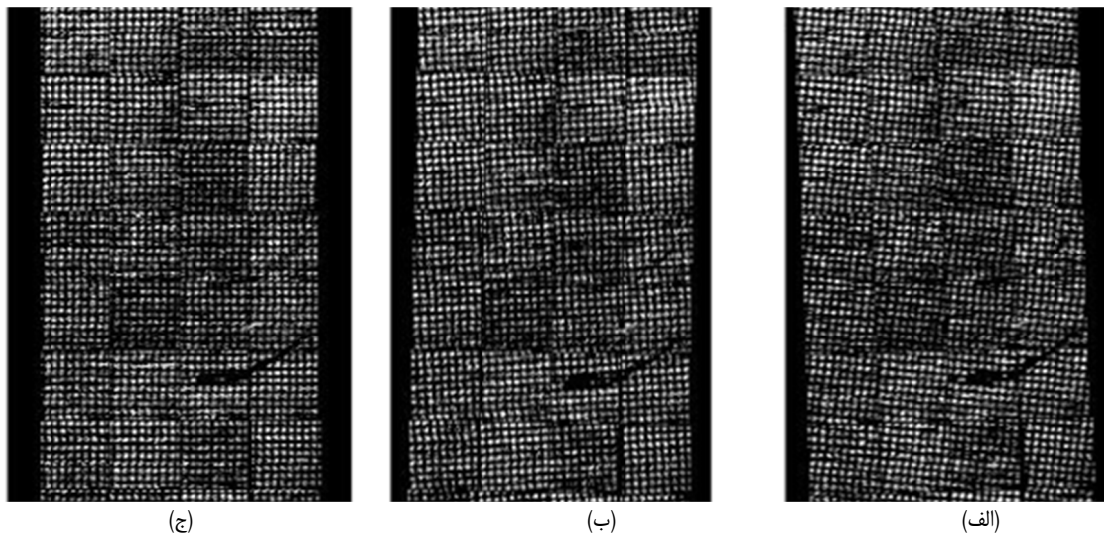


شکل ۱- شمای کلی از سیستم ارائه شده برای تحلیل تصاویر ریزآرایه

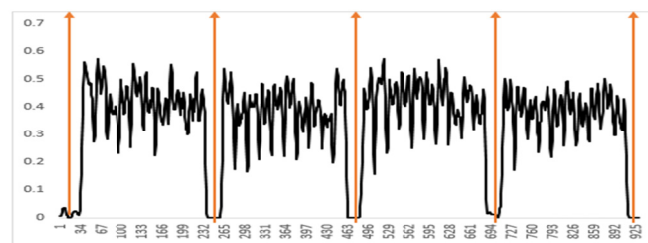
شکل (۳) رسم می‌شود. در این پژوهش با استفاده از یافتن کمترین مقدار محلی، حدآستانه هر بلوک مشخص می‌شود.

با توجه به اینکه هر بلوک شامل چندین خال می‌باشد، تعیین موقعیت هر خال در یک زیر بلوک، هدف اصلی مرحله سوم می‌باشد. در این مرحله ابتدا میانگین شدت روشنایی پیکسل‌های افقی / عمودی تصویر محاسبه و هیستوگرام آن رسم می‌شود (شکل ۴-الف). سپس برای یافتن الگوهای تکراری نظیر سیگنال تناوبی در شکل (۴-الف) که تحت نویز تصویر قرار گرفته است، از رابطه خود همبستگی (Autocovariance) استفاده می‌گردد. پس از اعمال رابطه خود همبستگی، مجدداً تابع یک بعدی نظیر شکل (۴-ب) حاصل می‌شود. سپس نقاطی که دارای بیشینه محلی هستند مشخص گردیده و میانگین آن‌ها تخمین زده می‌شود. میانگین این مقادیر، نشان‌دهنده فاصله تقریبی بین بیشترین و کمترین مقادیر متوالی محلی در هیستوگرام شکل (۴-الف)

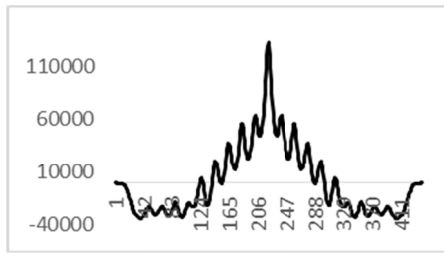
روی آنان صورت گیرد (۲۰-۲۲). با توجه به اینکه زاویه چرخش تصاویر ریزآرایه بسیار اندک است، بنابراین برای محاسبه میزان چرخش تبدیل رادن برای هر یک از تصاویرمابین زاویه -1° تا $+1^\circ$ محاسبه می‌شود. با پیدا کردن حداکثر مقدار رادن، مقدار زاویه چرخش تشخیص داده می‌شود. برای درک بهتر اثر تبدیلات رادن بر روی تصاویر ریزآرایه مثالی در شکل (۲) نمایش داده شده است. در این نمونه ابتدا بخشی از تصویری با زاویه چرخش صفر درجه انتخاب، سپس به صورت دستی در دو زاویه $1/5^\circ$ و $-1/5^\circ$ در جهت عقربه ساعت و عکس آن چرخانده (شکل ۲-الف-ب)، سپس به کمک الگوریتم تبدیلات رادن میزان انحراف زاویه در تصویر تشخیص و سپس تصحیح می‌گردد (شکل ۲-ج). سپس، در مرحله دوم تعداد بلوک‌ها محاسبه و موقعیت آن‌ها در تصویر شناسایی می‌شود. بدین منظور مجموع شدت روشنایی هر سطر / ستون از پیکسل‌ها در تصویر محاسبه و هیستوگرامی سطری / ستونی همانند



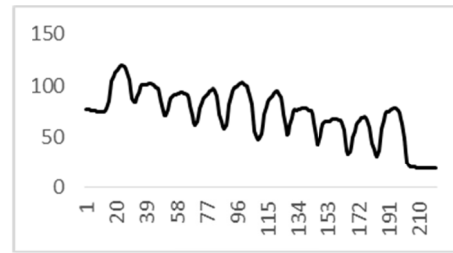
شکل ۲- نمایش تشخیص و تصحیح زاویه چرخش تصویر ریزآرایه. (الف) چرخش تصویر با زاویه $+1/5^\circ$. (ب) چرخش تصویر با زاویه $-1/5^\circ$. (ج) تصحیح زاویه چرخش به کمک روش رادن.



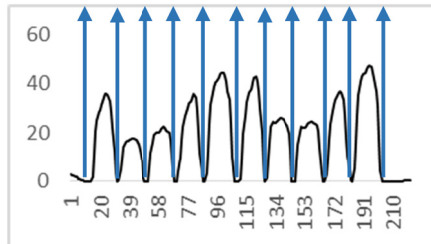
شکل ۳- تعیین محدوده‌ی بلوک‌ها با استفاده از هیستوگرام سطری تصویر.



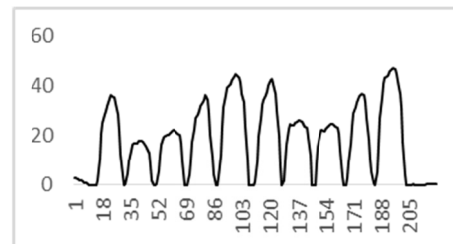
ب)



الف)



د)



ج)

شکل ۴- نحوه تعیین موقعیت خال‌ها در یک بلوک از تصویر ریزآرایه. الف) رسم هیستوگرام عمودی مربوط به یک بلوک در تصویر ریزآرایه. ب) رسم تابع خودهمبستگی حاصل از هیستوگرام عمودی مربوط به تصویر ریزآرایه. ج) حذف نویز به کمک تابع خود همبستگی و با اعمال عملگر ریخت‌شناسی بر روی هیستوگرام عمودی اولیه تصویر و ترسیم منظم هیستوگرام. د) تخمین فواصل مابین خال‌ها و رسم خطوط به کمک روش اتسو.

کلی (Total Variation Regularization-TV) و حداقل سازی نرم L1 می‌باشد (۱۶). این روش در سال ۲۰۰۴ برای تجزیه تصویر به ویژگی‌های متفاوت معرفی شد (۲۵). در این روش تصویر یک قسمت‌شامل رنگ پس‌زمینه و مرزهای مهم است و مابقی تصویر به عنوان بافت شناخته می‌شوند. این روش، میزان شدت سطح خاکستری مربوط به خال را برگردانده و پس‌زمینه تصویر را حذف می‌کند.

در مرحله پنجم پس از قطعه‌بندی خال‌ها، میزان شدت روشنایی متناسب با سطح خاکستری هر کدام به صورت مجزا محاسبه می‌شود. از آنجایی که دلیل اصلی بیماری سرطان ایجاد جهش ژنتیکی می‌باشد. برای تشخیص جهش کافی است شدت روشنایی دو نمونه سالم و توموری با هم مقایسه شوند. بدین منظور از معادله (۱) استفاده می‌شود.

$$\begin{aligned} \text{Gene - expression} &= \log_2 \left(\frac{\text{int}(cy5)}{\text{int}(cy3)} \right) \\ &= \log_2(\text{int}(cy5)) \\ &\quad - \log_2(\text{int}(cy3)) \end{aligned} \quad (1)$$

می‌باشد. مقدار میانگین محاسبه شده به منظور یافتن الگوی تکراری در هیستوگرام اولیه، توسط عملگرهای ریخت‌شناسی استفاده می‌شود. برای حذف نویز هیستوگرام اولیه از عملگر ریخت‌شناسی استفاده می‌گردد. با اعمال عملگر ریخت‌شناسی هیستوگرام منظمی همانند شکل (۴-ج) حاصل می‌شود. در نهایت برای شبکه‌بندی و شناسایی موقعیت خال‌ها بر روی هیستوگرام بدست آمده، از روش حد‌آستانه اتسو (۲۳) استفاده شده و با تخمین فواصل خالی مابین خال‌ها و رسم خطوط عمودی مابین آن‌ها، موقعیت هر یک از خال‌ها شناسایی می‌گردد (شکل ۴-د).

پس از شناسایی موقعیت هر خال، قطعه‌بندی بر روی تصویر صورت می‌گیرد. وجود پس‌زمینه ناهموار در تصویر ریزآرایه مهم‌ترین مشکل در این مرحله تلقی می‌شود. در طی مطالعات صورت گرفته (۲۴) مقدار پس‌زمینه مربوط به مواد شیمیایی است که در زمان آزمایش بالینی بر روی تراشه باقی مانده‌اند. یکی از بهترین روش‌های موجود جهت تشخیص و حذف پس‌زمینه در تصاویر ریزآرایه، روش ترکیبی منظم‌سازی تغییرات

گره‌ها انتخاب شود. بعد از اتمام کار قواعد مورد نظر از بررسی ریشه تا برگ‌ها بدست می‌آیند. شکل کلی این قواعد به صورت زیر می‌باشد.

*if gene(i) is A_i and...and gene(k) is A_k
then class C_j*

در قاعده بالا A_i, \dots, A_k مقادیر بیان ژن‌های $gene(i) \dots gene(k)$ است و C_j کلاس خروجی یا به عبارت دیگر نوع تومور را نشان می‌دهد.

یافته‌ها

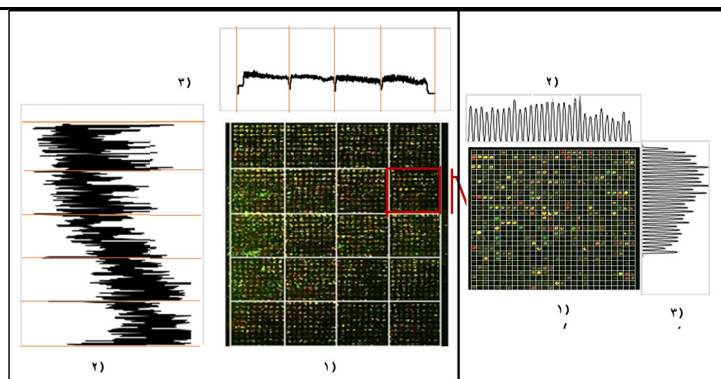
در این پژوهش یک روش تمام خودکار جهت استخراج و تحلیل اطلاعات از تصاویر ریزآرایه، برای تشخیص نوع سرطان سینه ارائه گردید. در این بخش عملکرد بخش‌های مختلف روش پیشنهادی بر روی تصاویر و داده‌های ریزآرایه مورد تجزیه و تحلیل قرار می‌گیرد.

نتایج بلوک‌بندی و شبکه‌بندی: برای شناسایی محل بلوک‌ها و خال‌ها در تصاویر ریزآرایه از روش آستانه‌گذاری هیستوگرام استفاده شد (طبق شکل ۵). به منظور ارزیابی روش پیشنهادی در تعیین موقعیت بلوک‌ها و شناسایی مکان خال‌ها، الگوریتم

$int(cy3)$ و $int(cy5)$ به ترتیب نماینده شدت روشنایی‌ها در نمونه‌های توموری و سالم می‌باشد. با کمی‌سازی داده‌ها در فاز پردازش تصاویر، حجم زیادی از داده‌های خام (ژن‌ها) از تصویر استخراج می‌شود که نیاز به پیش‌پردازش و مدیریت دارد. بدین منظور در مرحله ششم و هفتم، ویژگی‌های موجود در این مجموعه پس از نرمال‌سازی به کمک روش (Locally Weighted Linear Regression) (۲۶)، توسط روش بهره اطلاعاتی مورد بررسی قرار می‌گیرند. این روش، یک روش آماری جهت تعیین ارزش هر ویژگی در دسته‌بندی نمونه‌ها، براساس نتیجه مطلوب می‌باشد (۲۷). در آخرین مرحله الگوریتم درخت تصمیم‌گیری J48 به منظور طبقه‌بندی داده‌های ریزآرایه استفاده می‌شود. در این الگوریتم برای انتخاب ریشه از بین تمامی ویژگی‌ها از یک آزمون آماری استفاده می‌شود تا مشخص گردد هر ویژگی تا چه حد قادر است به تنهایی نمونه‌های آزمایشی را دسته‌بندی کند. به این منظور از آنتروپی و بهره اطلاعات استفاده می‌شود. به دنبال آن، عملیات فوق برای مثال‌های قرار گرفته در هر شاخه تکرار می‌شود تا بهترین ویژگی برای تمامی

جدول ۱- محاسبه میانگین دقت روش پیشنهادی برای تعیین موقعیت بلوک‌ها و خال‌ها در سه تصویر ریزآرایه.

نوع تصویر	تصویر با کیفیت خوب	تصویر با کیفیت متوسط	تصویر با کیفیت پایین
دقت بلوک‌بندی	٪۱۰۰	٪۱۰۰	٪۱۰۰
دقت شبکه‌بندی - تعیین موقعیت خال‌ها	٪۱۰۰	٪۹۸٫۸۳	٪۹۹٫۷۶



شکل ۵- تعیین موقعیت بلوک‌ها و خال‌ها در بخشی از یک تصویر ریزآرایه، بخش (الف) مربوط به تعیین موقعیت بلوک‌ها و بخش (ب) مربوط به شناسایی موقعیت خال‌ها می‌باشد. (الف-۱) شناسایی موقعیت بلوک‌ها در تصویر ریزآرایه، (الف-۲) نمایش هیستوگرام افقی و رسم خطوط مربوط به تعیین موقعیت بلوک‌ها. (الف-۳) رسم هیستوگرام عمودی و رسم خطوط مربوط به تعیین موقعیت بلوک. (ب-۱) شناسایی موقعیت خال‌ها در یک بلوک. (ب-۲) رسم هیستوگرام افقی برای تعیین موقعیت خال‌ها. (ب-۳) رسم هیستوگرام عمودی برای تعیین موقعیت خال‌ها.

جدول ۲- مقایسه میانگین خطای مربوط به قطعه‌بندی با اعمال روش‌های موجود و پیشنهادی.

روش تخمین پس‌زمینه	روش محلی	روش سراسری	روش ریخت‌شناسی	روش پیشنهادی
میانگین خطا	۳٫۴	۸٫۰۱	۴٫۰۳	۱٫۰۳

نتایج استخراج ویژگی و شناسایی سرطان سینه: با اعمال بهره اطلاعات بر روی داده‌های استخراج شده از تصویر ریزآرایه، ۱۰۰ ژن مفید از پایگاه داده‌ی سرطان سینه به منظور دسته‌بندی داده‌ها استخراج شده است. در جدول (۳) از بین ۱۰۰ ژن، ده ژن برتر به همراه مقدار بهره اطلاعاتی‌شان نمایش داده می‌شود.

با شناسایی ژن‌های موثر، به کمک الگوریتم درخت تصمیم‌گیری نوع سرطان سینه شناسایی می‌شود. بدین منظور ۱۰۰ ژن برتر به عنوان داده‌های ورودی به درخت تصمیم‌گیری اعمال شده و خروجی آن بیان‌گر نوع سرطان سینه می‌باشد. پس از اعمال درخت J48، قواعد با خواندن از ریشه تا برگ نوشته می‌شود. مجموعه قوانین استخراج شده برای تشخیص نوع سرطان سینه از درخت J48 به صورت زیر در آمده است.

- R1: $if(gene_{11965} \leq -0.0903) then class 1$
 R2: $if(gene_{11965} > -0.0903) and (gene_{5428} > 0.0467) and (gene_{11965} > 0.2532) then class 2$
 R3: $if(gene_{11965} > -0.0903) and (gene_{5428} \leq 0.0467) then class 2$
 R4: $if(gene_{11965} > -0.0903) and (gene_{5428} > 0.0467) and (gene_{11965} \leq 0.2532) then class 1$

نتایج نشان می‌دهد که تنها با استفاده از چند ژن، نوع سرطان سینه شناسایی می‌شود. برای ارزیابی عملیات مورد نظر، از پارامتر دقت معادله (۴) استفاده می‌شود. در این روش داده‌های ریزآرایه مربوط به سرطان سینه با دقت ۹۵٫۴۵٪ دسته‌بندی می‌شوند.

$$Accuracy = \frac{\text{Number of Correct Classified Instance}}{\text{Number of All Instance}} \quad (4)$$

بر روی سه تصویر ریزآرایه مربوط به سرطان سینه با کیفیت‌های متفاوت آزمایش شده است. نتایج ارزیابی در جدول (۱) بیانگر دقت روش پیشنهادی بر روی سه تصویر می‌باشد. تعیین دقت شبکه‌بندی هر تصویر توسط معادله (۲) محاسبه می‌گردد.

$$Accuracy = \left(\frac{N(\text{correctspot})}{N(\text{total spot})} \right) * 100 \quad (2)$$

در معادله بالا تعداد خال‌هایی که به صورت صحیح شبکه‌بندی شده‌اند $N(\text{correctspot})$ ، با تعداد کل خال‌ها، $N(\text{total spot})$ ، مقایسه می‌شوند.

نتایج قطعه‌بندی: برای تخمین پس‌زمینه، معروف‌ترین روش‌ها استفاده از عملگرهای ریخت‌شناسی، پس‌زمینه محلی و پس‌زمینه سراسری می‌باشد. این سه روش اغلب در سه نرم‌افزار ScanAlyze، GenePix و ImaGene به کار گرفته می‌شود؛ اما در این پژوهش روش ارائه شده نتایج خروجی آن همواره از روش‌های مذکور بهتر و دقیق‌تر است. برای درک بهتر هر چهار روش بر روی ماتریس یک خال اعمال (مطابق شکل ۶) و میانگین خطای متوسط (معادله ۴) محاسبه می‌گردد. برای عملگرهای ریخت‌شناسی از پنجره 7×7 استفاده می‌شود و ضریب لاگراتز برای تخمین پس‌زمینه با مقدار تقریبی ۰٫۳ انتخاب شده است. میانگین خطای متوسط شدت روشنایی هر تصویر مطابق معادله (۳) محاسبه می‌شود.

$$Accuracy = \sum_{i=1}^m \frac{\sum_{j=1}^n |u_{i,j}^{true} - u_{i,j}^{est}|}{m \times n} \quad (3)$$

$u_{i,j}^{est}$ و $u_{i,j}^{true}$ به ترتیب مربوط به شدت روشنایی واقعی و تخمینی پس‌زمینه حاصل از اعمال الگوریتم پیشنهادی در ماتریس $m \times n$ می‌باشند. مقدار متوسط خطا در جدول (۲) نمایش داده می‌شود. با توجه به جدول (۲) روش پیشنهادی همواره از دیگر روش‌ها بهتر عمل می‌کنند.

2	7	6	6	7	6	9	5	7	6	5	7	7
6	9	2	2	9	7	2	6	6	7	10	3	6
6	7	2	3	93	160	157	132	165	39	2	4	3
3	10	5	105	142	173	116	136	145	145	29	4	5
9	2	8	168	150	153	155	139	142	156	145	10	3
4	2	124	171	140	130	500	157	126	163	178	2	6
3	6	123	146	135	168	158	133	153	146	173	7	5
4	9	172	152	162	151	141	163	161	149	96	6	1
7	3	7	168	145	148	134	166	133	212	128	5	4
3	2	1	10	124	220	89	132	96	1	6	1	6
4	5	8	4	3	8	9	9	5	5	6	8	7
6	8	7	7	3	9	4	9	2	5	5	10	9

0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	89	150	151	122	155	31	0	0	0
0	0	0	96	136	169	109	132	138	136	21	0	0
0	0	0	160	144	150	145	136	140	148	143	0	0
0	0	120	167	135	123	141	156	118	161	174	0	0
0	0	120	145	130	161	151	123	147	145	165	0	0
0	0	170	151	154	146	138	159	153	142	92	0	0
0	0	0	160	137	141	133	162	131	208	122	0	0
0	0	0	0	120	212	85	127	94	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

شکل ۶- نمایش ماتریس یک خال.

الف) ماتریس واقعی خال بدون احتساب پس‌زمینه، ب) ماتریس خال به همراه پس‌زمینه ناهمگون.

جدول ۳- ده ژن برتر استخراج شده از سرطان سینه با اعمال روش بهره اطلاعات.

rank	Gene-number	IG value	rank	Gene-number	IG value
1	Gene-11965	0.707	6	Gene-7400	0.619
2	Gene-506	0.698	7	Gene-29826	0.573
3	Gene-38050	0.645	8	Gene-3144	0.565
4	Gene-9906	0.655	9	Gene-5428	0.565
5	Gene-34582	0.655	10	Gene-40030	0.559

بحث و نتیجه‌گیری

در این تحقیق، سیستمی برای تحلیل داده‌های ریزآرایه به منظور شناسایی نوع سرطان ارائه شد. این روش شامل سه فاز اصلی بوده که به ترتیب عبارتند از: پردازش تصویر به منظور کمی‌سازی تصویر، داده‌کاوی به منظور نرمال‌سازی و انتخاب ویژگی‌های مفید و در نهایت شناسایی نوع سرطان سینه. روش پیشنهادی بر روی پایگاه داده سرطان سینه اعمال شده است. در فاز اول، سیستم پیشنهادی برای تعیین محل دقیق بلوک‌ها، خال‌ها و قطعه‌بندی هر کدام از خال‌ها دارای عملکرد موفق‌آمیزی بوده است. این در حالی است که بسیاری از تصاویر ریزآرایه موجود، شامل خرابی‌هایی از جمله: وجود نویز فراوان در تصویر، چرخش تصویر، پایین بود میزان روشنای خال‌ها می‌باشد. بدین منظور ابتدا چرخش تصویر با دقت

بالایی تشخیص و تصحیح می‌گردد، سپس محل دقیق بلوک‌ها در تصویر به صورت صحیح شناسایی شده و با توجه به خرابی‌های ذکر شده بیشتر از ۹۸٪ خال‌ها در تمامی تصاویر ریزآرایه به درستی شبکه‌بندی می‌شوند. در حالی که در فاز پردازش تصویر برای کمی‌سازی تصاویر ریزآرایه سه نرم‌افزار GenePix، Imagen و ScanAlyze به بازار عرضه شده است، تمامی این نرم‌افزارها به صورت نیمه‌اتوماتیک عمل کرده و کاربر موظف است که برای شبکه‌بندی، موقعیت بلوک‌ها و فاصله مابین خال‌ها را به صورت دستی تعیین کند. روش پیشنهادی برای شبکه‌بندی تصاویر ریزآرایه به صورت تمام خودکار عمل کرده و نیاز به دخالت کاربر نمی‌باشد. همچنین در این نرم‌افزارها برای تخمین میزان پس‌زمینه از روش‌های رایج پس‌زمینه محلی، پس‌زمینه سراسری و ریخت-

جدول ۴- مقایسه عملکرد الگوریتم‌های درخت تصمیم‌گیری برای تشخیص نوع تومور.

دقت دسته‌بندها برای تشخیص نوع سرطان	نوع درخت تصمیم‌گیری
سرطان سینه	J48 Tree
% ۹۵,۴۵	Decision Stump
% ۶۶	LMT
% ۹۰,۴۷	Random Tree
% ۸۰,۹۵	Rep Tree
% ۸۵,۷۱	

2. Labib F.E.Z, Fouad I, Mabrouk M, Sharawy A. An efficient fully automated method for gridding microarray images. *AJBE*, 2012;2:115-9.

3. Harikiran J, RamaKrishna D, Avinash B, Lakshmi P, KiranKumar R, A new method of gridding for spot detection in microarray images. *CEIS*, 2014;5:25-33.

4. Eisen M, Microarray Image Analysis Software, ScanAlyze, Eisen Lab, <http://www.eisenlab.org>, Accessed Dec 1, 2015.

5. Instruments A, GenePix Pro 6.0 Microarray Acquisition and Analysis Software for GenePix Microarray Scanners User's Guide & Tutorial, Axon Instruments/ Molecular Devices Corp, Sunnyvale, CA, 2013.

6. Bozinov D, Rahnenführer J. Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. *Bioinformatics* 2002;18:747-56.

7. Wang T, Li T, Shao G, Wu S. An improved K-means clustering method for cDNA microarray image segmentation. *GMR*, 2015;14:7771-81.

8. Tarca A.L, Romero R, Draghici S. Analysis of microarray experiments of gene expression profiling. *AJBM* 2006;195:373-88.

9. Medigie C, ImaGene® 9.0 – Leading-Edge Microarray Analysis Software, BioDiscovery, <http://www.biodiscovery.com>, Accessed Dec 1, 2015.

10. Wallack D, Data Analysis with ScanAlyze, Department of Biology, Davidson College, Muhlenberg College, PA, <http://www.bio.davidson.edu>, Accessed Dec 1, 2015.

11. Golub T.R, Slonim D.K, Tamayo P, Huard C, Gaasenbeek M, Mesirov J.P, et al, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999;286:531-7.

12. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, et al. Multiclass cancer diagnosis using tumor gene expression signatures. *PNAS* 2001;98:15149-54.

13. Halder A, Dey S, Kumar A. Active learning using fuzzy k-NN for cancer classification from microarray gene expression data. *ACCS* 2015;374:103-13.

شناسی استفاده می‌شود. همان‌طور که مشاهده شد، این روش‌ها در مقابل روش پیشنهادی تخمین پس‌زمینه، ضعیف‌تر عمل می‌کنند و از دقت بالایی برخوردار نیستند. در نهایت، با استخراج داده‌های خام از تصویر ریزآرایه برای شناسایی نوع سرطان سینه از بین ژن‌های موجود، ۱۰۰ ژن مؤثرتر با استفاده از روش بهره اطلاعات انتخاب و به کمک درخت تصمیم‌گیری J48، ژن‌های برتر مورد تحلیل قرار می‌گیرند. میزان صحت طبقه‌بندی داده‌های سرطان سینه با تعداد چهار قانون با دقت % ۹۵/۴۵ تخمین زده شده است. به منظور مقایسه درخت تصمیم‌گیری ارائه شده با دیگر درخت تصمیم‌گیری، دقت روش پیشنهادی به همراه چهار روش دیگر برای شناسایی سرطان سینه در جدول (۴) گزارش شده است. کارایی روش پیشنهادی، برای مجموعه داده‌های موجود بالاتر می‌باشد. این سیستم به همراه قوانین استخراج شده، می‌تواند ابزار کمکی برای تحلیل و تشخیص بیماری سرطان سینه باشد و در درمان این بیماری پزشکان را یاری دهد. نتایج به دست آمده از این روش نشان می‌دهد که می‌توان از آن به عنوان یک سیستم مناسب برای شناسایی سرطان سینه استفاده نمود. همچنین به دلیل استفاده آسانتر و کم‌هزینه‌تر نسبت به آزمایش‌های بیولوژی، از این روش استفاده وسیع‌تری می‌توان به عمل آورد.

منابع

1. Bariamis D, Maroulis D, Iakovidis D.K, Automatic DNA microarray gridding based on Support Vector Machines, 8th IEEE International Conference on Bioinformatics and BioEngineering, IEEE, 2008, p: 1-5.

14. Wang L, Chu F, Xie W, Accurate cancer classification using expressions of very few genes. *IEEE/ACM*, 2007;4:40-53.
15. Sreedevi A, Jangamashetti D, Automatically Locating Spots in DNAMicroarray Image Using Genetic Algorithm without Gridding, *International Association of Computer science and Information Technology – Spring Conference*, 2009,p: 178-181.
16. Yin W, Chen T, Zhou S.X, Chakraborty A. Background correction for cDNA microarray images using the TV+ L1 model. *Bioinformatics* 2005;21:2410-6.
17. Stanford Microarray Database, <http://smd.stanford.edu/>, Accessed Dec 1, 2015.
18. Kao J, Salari K, Bocanegra M, Choi Y.L, Girard L, Gandhi J, et al. Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *Plos One* 2009;4:e6146.
19. Bergamaschi A, Kim Y.H, Wang P, Sørlie T, Hernandez T, Lonning PE,et al, Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene_expression subtypes of breast cancer. *GHC*, 2006;45:1033-40.
20. Fouad IA. a new robust and automatic analytical approach for block indexing of cDNA Microarray image. *IJCEA*, 2014;5:1:16
21. Rueda L. Sub-grid detection in DNA microarray images, *AIVT*,2007;4872:248-59.
22. Rueda L, Rezaeian I. Afully automatic gridding method for cDNA microarray images. *BMC Bioinformatics* 2011;12:113.
23. Otsu N. A threshold selection method from gray-level histograms. *Automatica* 1975;11:23-7.
24. Uchida S, Nishida Y, Satou K, Muta S, Tashiro K, Kuhara S. Detection and normalization of biases present in spotted cDNA microarray data: a composite method addressing dye, intensity-dependent, spatially-dependent and print-order biases. *DNA Res* 2005;12:1-7.
25. Chan TF, Esedoglu S. Aspectsof total variation regularized L 1 function approximation. *JAM*, 2005;65:1817-37.
26. Yang B, Bao X. Identification of genes associated with laryngeal squamous cell carcinoma samples based on bioinformatic analysis. *MMR*, 2015;12:3386-92.
27. Dai J, Xu Q. Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *ASP*, 2013;13:211-21.

Microarray images analysis to detect breast cancer

Nastaran Dehghan, Msc of Computer Engineering, Shahrood University of Technology, Shahrood, Iran. dehghan_nastaran@shahroodut.ac.ir

***Hamid Hassanpour**, PhD, Professor of Computer Engineering, Shahrood University of Technology, Shahrood, Iran (*Corresponding author). h.hassanpour@shahroodut.ac.ir

Mohammad Reza. Abbaszadegan, MD, Professor of Medical Genetics, Mashhad University of Medical Sciences, Mashhad, Iran. abbaszadeganmr@mums.ac.ir

Abstract

Background: Microarray technology is a powerful tool to study and analyze the behavior of thousands of genes simultaneously. Images of microarray have an important role in the detection and treatment of diseases. The aim of this study is to provide an automatic method for the extraction and analysis of microarray images to detect cancerous diseases.

Methods: The proposed system consists of three main phases of image processing, data mining, and detection of disease. The image processing phase is accompanied with some operations such as identifying the location of genes, deleting the background, and extracting the raw data from the images. The second phase includes data normalization and selection of more effective genes. The disease is identified and recognized in the third phase using the extracted data.

Results: In this study it has been used from breast cancer microarray images from Stanford University database. The accuracy of the proposed method to locate genes and diagnosis of breast cancer is up to 98 and 95.45%, respectively.

Conclusion: The obtained results indicate that the proposed method is more accurate than other existing methods in microarray analysis. In addition, the proposed method is easily implemented and less costly compared to the clinical tests.

Keywords: Microarray, Image processing, Data mining, Breast cancer