

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

بسمه تعالی



دانشگاه صنعتی شاهرود

دانشکده مهندسی کامپیوتر و فناوری اطلاعات

گروه مهندسی کامپیوتر - هوش مصنوعی

دسته‌بندی مجموعه داده‌های ریز آرایه دی‌ان‌ای به منظور تشخیص سرطان

نسترن دهقان خلیل‌آباد

استاد راهنما

دکتر حمید حسن‌پور

استاد مشاور

دکتر محمدرضا عباس‌زادگان

پایان نامه ارشد جهت اخذ درجه کارشناسی ارشد

اسفند ۱۳۹۴

با پاس از سه وجود مقدس:

آنان که ناتوان شدند تا ما به توانایی برسیم...

موباشان سپید شد تا ما رو سفید شویم...

و عاشقانه سوختند تا که ما بخش وجود ما و روشمگر را همان باشند...

پدرانمان

مادرانمان

استادانمان

و همچنین خواهرانم همراهان همیشگی و پشتوانه‌های زندگیم...

تقدیر و تشکر

ضمن سپاس و ستایش به درگاه ایزد منان که به من توانایی داد که با استعانت از او بتوانم این پژوهش را انجام دهم، بر خود لازم می‌بینم از دکر می و تشویق اساتید و دوستان که در مکارش این مجموعه مرایاری نمودند، قدردانی نمایم:

جناب آقای دکتر حمید حسن پور، استاد راهنما و دکتر محمد رضا عباس زادگان استاد مشاور، که در طول مکارش این مجموعه بارهای بسیاری عالمند و بجایشان، سکندار شایسته‌ای در هدایت این پایان نامه بوده‌اند.

در ادامه از کلیه اساتید ارجمند در طول دوره تحصیل از جمله آقایان دکتر پویان، دکتر زاهدی و خانم دکتر مشایخی قدردانی و تشکر می‌نمایم.

و در خاتمه از بهرامان زندگیم، پدر، مادر و سایر عزیزان و همچنین دوستان ارجمندم آقای شهابی و خانم‌ها، طهرانی،، زارع، انصاری، چرامین، اسکو و دیگر دوستانی که مراد انجام این تحقیق یاری نمودند، قدردانی و تشکر می‌نمایم.

نسترن دهقان

اسفند ماه ۱۳۹۴

تعهد نامه

اینجانب نسترن دهقان خلیل اباد دانشجوی دوره کارشناسی ارشد رشته مهندسی کامپیوتر- هوش مصنوعی دانشکده مهندسی فناوری اطلاعات و کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان نامه دسته بندی مجموعه داده های ریزآرایه دی ان ای به منظور تشخیص سرطان تحت راهنمایی دکتر حمید حسن پور متعهد می شوم .

- تحقیقات در این پایان نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است.
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان نامه ، در مواردی که از موجود زنده (یا یافته های آن ها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

- کلیه حقوق معنوی این اثر و محصولات آن (مقالات مستخرج ، کتاب ، برنامه های رایانه ای ، نرم افزار ها و تجهیزات ساخته شده است) متعلق به دانشگاه صنعتی شاهرود می باشد . این مطلب باید به نحو مقتضی در تولیدات علمی مربوطه ذکر شود .
- استفاده از اطلاعات و نتایج موجود در پایان نامه بدون ذکر مرجع مجاز نمی باشد.

* متن این صفحه نیز باید در ابتدای نسخه های تکثیر شده پایان نامه وجود داشته باشد .

چکیده

در دهه اخیر، علم ژنتیک و به ویژه ژنتیک مولکولی دچار تحولات بسیاری شده است. یکی از این تحولات، ظهور 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، تشخیص و درمان بیماری‌ها بر عهده داشته و نحوه تحلیل این تصاویر، تأثیر مستقیمی در نتایج بدست آمده دارد. اما سیستم‌هایی که به منظور تحلیل این تصاویر وجود دارند تنها قادر به کمی‌سازی تصاویر (استخراج داده‌های خام از تصویر) ریزآرایه هستند و عموماً بصورت دستی عمل می‌کنند. هدف از این پایان‌نامه، ارائه یک سیستم خودکار جهت استخراج و تحلیل داده‌ها از تصاویر ریزآرایه به منظور تشخیص بیماری‌های سرطانی می‌باشد. سیستم پیشنهادی شامل سه فاز اصلی پردازش تصویر، داده‌کاوی و شناسایی بیماری می‌باشد. در فاز پردازش تصویر عملیاتی مانند حذف نویز، رفع چرخش، تعیین مکان ژن، حذف پس‌زمینه و استخراج داده‌های خام از تصویر انجام می‌گیرد. فاز داده‌کاوی نیز 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.

در این پژوهش از مجموعه تصاویر ریزآرایه سرطان سینه، سرطان خون و غدد لنفاوی تهیه شده توسط دانشگاه استنفورد به منظور بررسی و آزمایش سیستم پیشنهادی استفاده شده است. نتایج بدست آمده از آزمایش نشان می‌دهد که دقت این سیستم در مرحله تعیین مکان ژن‌ها بر روی تصاویر ریزآرایه بیش از ۹۸٪ است. همچنین این سیستم قادر به شناسایی نوع سرطان در پایگاه داده‌ی سرطان سینه، سرطان غدد لنفاوی و سرطان خون به ترتیب با دقت ۹۵/۴۵٪، ۱۰۰٪ و ۹۴/۱۱٪ می‌باشد.

کلمات کلیدی: ریزآرایه، پردازش تصویر، داده‌کاوی، شناسایی بیماری.

لیست مقالات مستخرج از پایان نامه

- [1]: N. Dehghan, H. Hassanpour and M. R. Abbaszadegan “Fully Automatic Classification of Breast Cancer Microarray Images”, Journal of Electrical Systems and Information Technology (JESIT), Elsevier, October 2015. (Submitted)
- [2]: N. Dehghan, H. Hassanpour and M. R. Abbaszadegan “Microarray images analysis to detect breast cancer”, Razi Journal of Medical Sciences, January 2016. (Submitted), (Persian).
- [3]: N. Dehghan and H. Hassanpour “Design of Automatic system for microarray image analysis”, Journal of Advances in Computer Research (JACR), February 22, 2016, (To be submitted).

فهرست مطالب

فصل اول - آشنایی با فناوری ریزآرایه

۱-۱	مقدمه	۲
۲-۱	معرفی ساختار دی‌ان‌ای	۲
۳-۱	فناوری ریزآرایه	۵
۴-۱	تحلیل تصاویر ریزآرایه	۸
۱-۴-۱	شبکه‌بندی	۱۰
۲-۴-۱	قطعه‌بندی	۱۱
۳-۴-۱	انتخاب ویژگی	۱۴
۴-۴-۱	طبقه‌بندی داده‌ها	۱۵
۵-۱	کاربرد ریزآرایه در سرطان	۱۶
۶-۱	اهداف پایان‌نامه	۱۸
۷-۱	ساختار پایان‌نامه	۱۹
۸-۱	نتیجه‌گیری	۱۹

فصل دوم - مروری بر تحقیقات پیشین

۱-۲	مقدمه	۲۲
۲-۲	مروری بر نرم‌افزارهای موجود	۲۲
۱-۲-۲	نرم‌افزار GENEPIX	۲۳
۲-۲-۲	نرم‌افزار IMAGE	۲۸
۳-۲-۲	نرم‌افزار SCANALYZE	۲۹
۳-۲	مروری بر الگوریتم‌های تحلیل تصاویر ریزآرایه	۳۲
۴-۲	نتیجه‌گیری	۳۸

فصل سوم - سیستم تحلیل تصاویر ریزآرایه

۱-۳	مقدمه	۴۲
۲-۳	سیستم پیشنهادی	۴۲
۱-۲-۳	تشخیص زاویه چرخش و تصحیح آن	۴۴
۲-۲-۳	بلوک‌بندی تصویر	۴۷
۳-۲-۳	شبکه‌بندی هر بلوک	۴۹
۴-۲-۳	قطعه‌بندی خال‌ها	۵۹
۵-۲-۳	محاسبه بیان ژن	۶۳
۶-۲-۳	نرمال‌سازی	۶۶
۷-۲-۳	انتخاب ویژگی	۶۸
۴-۲-۳	دسته‌بندی داده‌ها	۷۰

۷۲.....	نتیجه‌گیری ۳-۳
---------	----------------

فصل چهارم- ارزیابی عملکرد و تحلیل نتایج

۷۴.....	مقدمه ۱-۴
۷۴.....	پایگاه داده ۲-۴
۷۴.....	۱-۲-۴ پایگاه داده سرطان سینه
۷۵.....	۲-۲-۴ پایگاه داده سرطان غدد لنفاوی
۷۵.....	۳-۲-۴ پایگاه داده سرطان خون
۷۶.....	۳-۴ بررسی و تحلیل نتایج
۷۶.....	۱-۳-۴ نتایج تنظیم چرخش تصویر
۷۷.....	۲-۳-۴ نتایج بلوک‌بندی و شبکه‌بندی تصاویر ریزآرایه
۸۱.....	۳-۳-۴ نتایج قطعه‌بندی خال از تصویر
۸۲.....	۴-۳-۴ نتایج محاسبه بیان ژن از تصاویر ریزآرایه
۸۴.....	۵-۳-۴ نرمال‌سازی داده‌های ریزآرایه
۸۶.....	۶-۳-۴ نتایج استخراج ویژگی بر روی پایگاه داده ریزآرایه
۸۷.....	۷-۳-۴ نتایج دسته‌بندی داده‌های ریزآرایه برای تعیین نوع سرطان
۹۰.....	۴-۴ مقایسه عملکرد سیستم پیشنهادی با نتایج دیگر روش‌ها
۹۱.....	۱-۴-۴ تشخیص و تصحیح زاویه چرخش
۹۲.....	۲-۴-۴ بلوک‌بندی و شبکه‌بندی
۹۳.....	۳-۴-۴ قطعه‌بندی و تصحیح پس‌زمینه
۹۵.....	۴-۴-۴ محاسبه شدت روشنایی خال‌ها
۹۶.....	۵-۴-۴ طبقه‌بندی داده‌های ریزآرایه
۹۷.....	۵-۴ نتیجه‌گیری

فصل پنجم- نتیجه‌گیری و پیشنهادات آتی

۱۰۰.....	مقدمه ۱-۵
۱۰۰.....	نتیجه‌گیری ۲-۵
۱۰۲.....	پیشنهادات آتی ۳-۵
۱۰۵.....	پیوست
۱۱۱.....	منابع

فهرست نمودارها

فصل سوم

- نمودار ۳ - ۱- تعیین محدوده بلوک‌ها با استفاده از هیستوگرام تصویر. ۴۹.....
- نمودار ۳ - ۲- هیستوگرام عمودی مربوط به تصویر ریزآرایه. ۵۵.....
- نمودار ۳ - ۳- منظم‌سازی و حذف نویز هیستوگرام به کمک تابع خود همبستگی. ۵۷.....
- نمودار ۳ - ۴- برجسب گذاری نواحی خال‌ها به صورت مجزا. ۵۸.....
- نمودار ۳ - ۵- تعیین محدوده خال‌ها. ۵۸.....
- نمودار ۳ - ۶- پراکندگی داده‌های ریزآرایه. ۶۵.....
- نمودار ۳ - ۷- بیان ژن‌ها با محاسبه میزان شدت روشنایی هر خال در تصویر ریزآرایه. ۶۶.....
- نمودار ۳ - ۸- تخمین منحنی برازش بر روی داده‌های ریزآرایه. ۶۸.....
- نمودار ۳ - ۹- محاسبه داده‌های نرمال با استفاده از منحنی برازش. ۶۸.....
- نمودار ۳ - ۱۰- مقایسه داده‌های مصنوعی ریزآرایه قبل و بعد از عمل نرمال‌سازی بر روی نمودار M_A. ۶۹.....

فصل چهارم

- نمودار ۴ - ۱- مقایسه پراکندگی داده‌های خام استخراج شده از تصویر ریزآرایه با تبدیلات لگاریتمی آن‌ها. ۸۴.....
- نمودار ۴ - ۲- بیان ژن‌های محاسبه شده از تصویر ریزآرایه. ۸۴.....
- نمودار ۴ - ۳- تعیین منحنی برازش بر روی داده‌های ریزآرایه با استفاده از روش Lowess. ۸۵.....
- نمودار ۴ - ۴- مقایسه داده‌های ریزآرایه قبل و بعد از عمل نرمال‌سازی در نمودار M_A. ۸۵.....

فهرست جدول‌ها

فصل دوم

جدول ۲-۱- معرفی نرم‌افزارهای تحلیل تصاویر ریزآرایه. ۲۳

فصل چهارم

جدول ۴-۱- محاسبه میانگین دقت روش پیشنهادی برای تصحیح زاویه چرخش بر روی دو تصویر. ۷۷

جدول ۴-۲- محاسبه میانگین دقت روش پیشنهادی برای تعیین موقعیت بلوک‌ها و خال‌ها در سه تصویر ریزآرایه. ۸۰

جدول ۴-۳- محاسبه میانگین خطا برای مقایسه دو روش ریخت‌شناسی و $TV+L1$ برای حذف پس‌زمینه. ۸۲

جدول ۴-۴- ده ژن برتر استخراج شده از سرطان سینه با اعمال روش بهره اطلاعات. ۸۶

جدول ۴-۵- ده ژن برتر استخراج شده از سرطان غدد لنفاوی با اعمال روش بهره اطلاعات. ۸۶

جدول ۴-۶- ده ژن برتر استخراج شده از سرطان خون با اعمال روش بهره اطلاعات. ۸۷

جدول ۴-۷- مقایسه میانگین خطای مربوط به قطعه‌بندی تصاویر با اعمال روش‌های موجود و پیشنهادی. ۹۵

جدول ۴-۸- مقایسه عملکرد الگوریتم‌های درخت تصمیم‌گیری برای تشخیص نوع تومور. ۹۶

فهرست شکل‌ها

فصل اول

- شکل ۱-۱- نمونه‌ای از بیان ژن‌ها در دو سلول متفاوت..... ۴
- شکل ۱-۲- مدل قرارگیری تک رشته‌های دی‌ان‌ای بر روی تراشه‌ی ریزآرایه..... ۵
- شکل ۱-۳- شمای کلی مراحل مختلف آزمون داده‌های ریزآرایه..... ۶
- شکل ۱-۴- بخشی از تصویر ریزآرایه..... ۷
- شکل ۱-۵- چگونگی انتخاب مناطق محلی برای تخمین پس‌زمینه در نرم‌افزارهای مختلف..... ۱۳

فصل دوم

- شکل ۲-۱- محیط کاربری نرم‌افزار GenePix..... ۲۴
- شکل ۲-۲- فواصل مابین خال‌ها و بلوک‌های نمایش داده شده در نرم‌افزار GenePix..... ۲۵
- شکل ۲-۳- استفاده از روش قطعه‌بندی خال‌ها براساس دایره‌ی افقی در نرم‌افزار GenePix..... ۲۶
- شکل ۲-۴- تعیین مناطق محلی خال برای تخمین پس‌زمینه در نرم‌افزار GenePix..... ۲۷
- شکل ۲-۵- نمودار پراکندگی داده‌های ریزآرایه در نرم‌افزار GenePix..... ۲۸
- شکل ۲-۶- تعیین مناطق محلی خال برای تخمین پس‌زمینه در نرم‌افزار ImaGene..... ۲۹
- شکل ۲-۷- تعیین مناطق محلی خال برای تخمین پس‌زمینه در نرم‌افزار ScanAlyze..... ۳۰
- شکل ۲-۸- شبکه‌بندی تصویر ریزآرایه با استفاده از نرم‌افزار GenePix..... ۳۱

فصل سوم

- شکل ۳-۱- شمای کلی از سیستم ارائه شده برای تحلیل تصاویر ریزآرایه..... ۴۳
- شکل ۳-۲- نحوه‌ی تشخیص زاویه چرخش در یک بلوک از تصویر ریزآرایه با استفاده از روش [۵۶]..... ۴۴
- شکل ۳-۳- نحوه‌ی محاسبه تبدیل رادن از یک تصویر دو بعدی..... ۴۶
- شکل ۳-۴- تشخیص زاویه چرخش و رفع آن با استفاده از روش تبدیلات رادن..... ۴۷
- شکل ۳-۵- تعیین موقعیت بلوک‌ها در یک تصویر ریزآرایه..... ۵۱
- شکل ۳-۶- تصویر ریزآرایه‌ی مصنوعی به همراه نویز بالا..... ۵۲
- شکل ۳-۷- شمای کلی تصحیح نویز سراسری پس‌زمینه در تصویر ریزآرایه..... ۵۳
- شکل ۳-۸- شمای کلی حذف نویز مشعل در تصویر ریزآرایه..... ۵۵
- شکل ۳-۹- نمایش شبکه‌بندی تصویر ریزآرایه..... ۵۹
- شکل ۳-۱۰- حذف پس‌زمینه از تصاویر ریزآرایه مصنوعی..... ۶۳

فصل چهارم

- شکل ۴-۱- تشخیص و تصحیح زاویه چرخش تصویر ریزآرایه..... ۷۸
- شکل ۴-۲- تعیین موقعیت بلوک‌ها در یک تصویر ریزآرایه..... ۷۹
- شکل ۴-۳- شبکه‌بندی تصویر ریزآرایه..... ۸۰
- شکل ۴-۴- مقایسه روش پیشنهادی و روش ریخت‌شناسی برای حذف پس‌زمینه از تصاویر ریزآرایه..... ۸۲

- شکل ۴-۵- مقادیر پیکسل‌های یک خال در تصویر ریزآرایه و حذف پس‌زمینه با روش پیشنهادی. ۸۳
- شکل ۴-۶- درخت تصمیم‌گیری برای تشخیص نوع سرطان سینه. ۸۸
- شکل ۴-۷- درخت تصمیم‌گیری برای تشخیص نوع سرطان غدد لنفاوی. ۸۹
- شکل ۴-۸- درخت تصمیم‌گیری برای تشخیص نوع سرطان خون. ۹۰
- شکل ۴-۹- مقایسه عملیات شبکه‌بندی با استفاده از نرم‌افزار GenePix با روش پیشنهادی. ۹۲
- شکل ۴-۱۰- مقایسه‌ی روش پیشنهادی برای تخمین میزان پس‌زمینه از روش ریخت‌شناسی. ۹۴

فصل پنجم

- شکل ۵-۱- روابط مابین ژن‌ها در یک تصویر ریزآرایه. ۱۰۳
- شکل ۵-۲- نمونه‌ای از خرابی‌های رایج در تصویر ریزآرایه. ۱۰۳

فصل اول

آشنایی با فناوری ریزآرایه

۱-۱ مقدمه

شناخت و کشف علت بیماری کمک شایانی به پیشگیری از بیماری، تشخیص به موقع و اعمال روش‌های درمانی مناسب می‌کند. در گذشته، تشخیص امراض مهم و خطرناک مانند سرطان عموماً بر پایه علائم غیرمولکولی مثل نوع بافت تومور، آزمایشات مختلف، ویژگی‌های آسیب‌شناسی، علائم و اثرات شهودی بیماری، مراحل و خصوصیات بالینی بوده است [۱].

یکی از مهم‌ترین و دقیق‌ترین راه‌ها برای کشف بیماری‌ها و حتی پیش‌بینی آن‌ها استفاده از دی‌ان‌ای^۱ افراد و اطلاعات ژنتیکی آن‌ها می‌باشد. این اطلاعات می‌تواند بسیار کارا و مفید باشد و همچنین در دسته‌بندی انواع گوناگون سرطان‌ها به ما کمک کند. در سال‌های اخیر، فناوری ریزآرایه^۲ امکان بررسی بیان هزاران ژن را به طور همزمان فراهم کرده است، و به دنبال آن حجم عظیمی از داده‌های مربوط به ژن‌های انسانی در اختیار محققین قرار گرفت. یکی از مشکلات پیش‌رو، انبوه داده‌ها و اطلاعات به دست آمده از داده‌های مولکولی و ژنتیکی است که لزوم استفاده از روش‌های آماری و تکنولوژی اطلاعات و همچنین یادگیری ماشین را متذکر می‌شود. از آنجایی که داده‌های بالینی شامل نمونه‌های اندک و هر نمونه دارای ویژگی‌ها یا ژن‌های فراوانی می‌باشد، بررسی تمام این ژن‌ها و به دست آوردن نتایجی که پزشکان را در تشخیص بیماری‌ها یاری کند، کاری مشکل است. استفاده از روش‌های کامپیوتری و هوش مصنوعی برای تحلیل تصاویر ریزآرایه، از سالیان پیش مرسوم بوده و کمک شایانی به علم پزشکی کرده است [۲].

۲-۱ معرفی ساختار دی‌ان‌ای

دی‌ان‌ای، مولکول بسیار طولی است که از دو رشته مجزا تشکیل می‌شود. این دو رشته مانند یک زیپ به هم متصل شده‌اند. اطلاعات زیاد موجود در دی‌ان‌ای، در واحدهایی به نام "ژن" ذخیره می‌شود. شکل هر ژن قسمتی از مولکول دی‌ان‌ای است. ژن‌ها برای ساختن پروتئین یا آر‌ان‌ای^۳ مورد استفاده قرار

^۱ DNA (Deoxyribo Nucleic Acid)

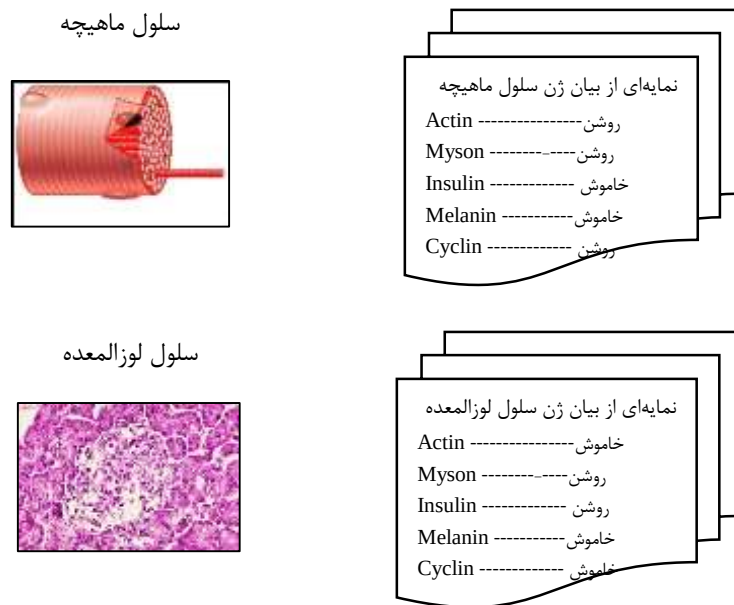
^۲ Microarray

^۳ RNA (RiboNucleic Acid)

می‌گیرند. یک ملکول دی‌ان‌ای مانند واگن‌های یک قطار به دنبال هم قرار گرفته‌اند و دور محور فرضی تابیده‌اند [۳]. برای کسب اطلاعات بیشتر در مورد ساختار دی‌ان‌ای به بخش پیوست پایان نامه مراجعه کنید.

سلول‌ها از همه ژن‌های خود به صورت همزمان استفاده نمی‌کنند. زمانی که یک ژن مورد استفاده قرار می‌گیرد، آن ژن فعال شده (روشن) در غیر این صورت ژن غیرفعال (خاموش) می‌باشد. بنابراین تنظیم بیان ژن به عنوان میزان فعال یا غیر فعال بودن هر ژن معرفی می‌شود. ماده ژنتیکی همه سلول‌ها یکی است، اما شکل و کار سلول‌های بدن مانند سلول‌های ماهیچه‌ای و عصبی، با هم متفاوت است، به این دلیل که در هر نوع سلول فقط بعضی از ژن‌ها بیان می‌شود و مابقی فعال نیستند (شکل ۱-۱). در این شکل در دو بافت متفاوت بدن با توجه به کارآیی آن‌ها برخی ژن‌ها خاموش و برخی روشن هستند [۴]. تاکنون برای مطالعه خواص، فعالیت و نقش ژن‌ها و پروتئین‌ها، اکثراً یک ژن یا حداکثر چند ژن بطور خاص و جداگانه مورد بررسی و مطالعه قرار می‌گرفت. بررسی همزمان هزاران ژن و پروتئین کار بسیار دشواری بود. به همین دلیل پروژه‌ای به نام ژنوم^۱ تعریف شد. پروژه ژنوم انسان یک تلاش بین‌المللی بود که ۱۳ سال به طول انجامید و به طور رسمی در اکتبر ۱۹۹۰ آغاز به کار کرد. پیش‌بینی می‌شد که انجام این پروژه ۱۵ سال به طول انجامد، اما پیشرفت سریع فناوری، باعث تسریع در فرآیند تکمیل آن شد و پروژه در سال ۲۰۰۳ به اتمام رسید. اهداف پروژه تعیین توالی کامل ۳ میلیارد زیر واحد دی‌ان‌ای، تعیین همه ژن‌های انسان و قابل دسترسی بودن اطلاعات ژنی برای مطالعات زیستی بیشتر بود. در ملکول دی‌ان‌ای تشخیص مرزهای ژن کار مشکلی است به همین دلیل ژن‌ها به صورت مبهم بیان می‌شوند. علاوه بر این هر ساله یک یا چند ژن جدید شناسایی و معرفی می‌شود [۵-۷].

¹ Genome



شکل ۱-۱- نمونه‌ای از بیان ژن‌ها در دو سلول متفاوت.

پس از خاتمه پروژه ژنوم انسانی و تعیین ردیف کامل ژنوم انسان که شامل ۳ میلیارد زیر واحد دی‌ان‌ای بود، به صورت تقریبی این واحدهای دی‌ان‌ای ۳۰ هزار ژن را رمز نمودند. با بررسی این اطلاعات مشخص شد که مطالعه یک ژن یا یک پروتئین و کشف یک فرایند بطور جداگانه کمک بسیاری به حل مشکل نمی‌کند. زیرا فرایندهای زیستی بسیار پیچیده است و گاهی ده‌ها و صدها ژن در بروز یک فعالیت نقش دارند.

در سال‌های اخیر با تکمیل پروژه ژنوم انسان و امکان بررسی تعداد قابل توجهی از ژن‌ها، فناوری ریزآرایه^۱ توجه زیادی را به خود جلب کرده است. در همین مسیر فناوری‌های ژنومیک^۲ (بررسی چندین ژن به صورت همزمان) و پروتئومیک^۳ (بررسی چندین پروتئین به صورت همزمان) کمک بسیاری جهت بررسی فعالیت و تغییرات مجموعه ژن‌ها و پروتئین‌ها نمود که هنوز نیز به عنوان روش‌های مفید استفاده می‌شود [۸].

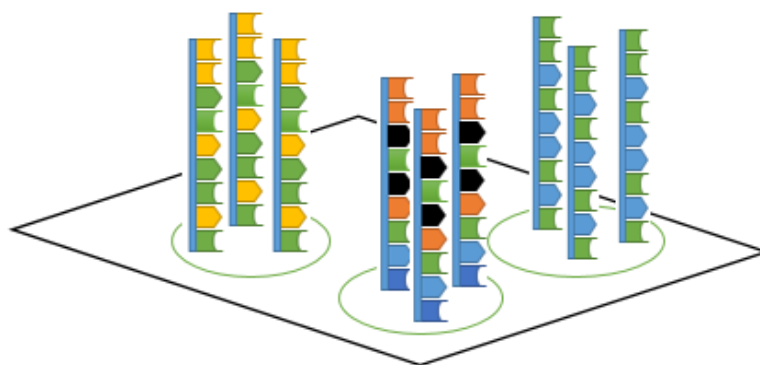
^۱ Microarray

^۲ Genomic

^۳ Proteomic

۳-۱ فناوری ریزآرایه

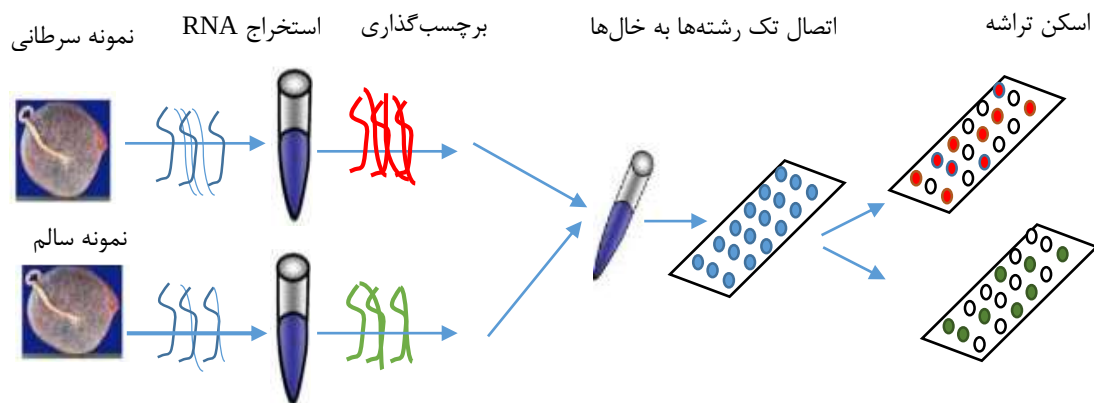
هزاران ژن و محصولات آن‌ها (مانند آران‌ای و پروتئین) همگی در یک مسیر پیچیده و منظم در بدن یک موجود زنده به فعالیت مشغولند. با این حال روش‌های آزمایشگاهی در سال‌های گذشته در زیست‌شناسی مولکولی بر اساس یک ژن در یک آزمایش کار می‌کنند که در این صورت بدست آوردن بررسی جامع از میزان فعالیت ژن‌ها سخت و غیرممکن به نظر می‌رسد. در چند سال اخیر فناوری ریزآرایه توجه دانشمندان بیولوژی مولکولی و دیگر رشته‌های علمی را به خود جلب کرده است. این فناوری به محققان کمک می‌کند تا دید بهتر و گسترده‌تری از تعامل میان هزاران ژن به طور همزمان داشته باشند. مفهوم اساسی فناوری ریزآرایه، سیستم سنتی تشخیصی-آنالیزی زیستی است. به طور ساده ریزآرایه عبارت است از فناوری بررسی ده‌ها، صدها و هزاران ژن در یک سطح کوچک (حدود چند سانتی مربع) جهت مقایسه یا بررسی تغییراتی نظیر کاهش، افزایش یا عدم تغییر در ساختار و فعالیت ژن‌ها یا پروتئین‌های نمونه با نمونه‌های شاهد می‌باشد. بر روی تراشه ریزآرایه، هزاران تک رشته‌ی دی‌ان‌ای به عنوان نماینده‌ی از ژن‌های متفاوت قرار گرفته‌اند [۹]. این قرارگیری توسط ربات‌هایی با نام arrayer صورت می‌گیرد. به هر ژن بر روی تراشه ریزآرایه یک خال گفته می‌شود (شکل ۱-۲).



شکل ۱-۲ مدل قرارگیری تک رشته‌های دی‌ان‌ای بر روی تراشه‌ی ریزآرایه.

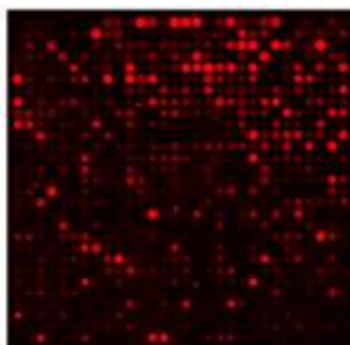
به طور کلی تکنولوژی ریزآرایه را می‌توان به دو بخش کلی تقسیم نمود: بخش اول شامل عملیات آزمایشگاهی این تکنیک می‌باشد که عبارت است از: ساخت تراشه ریزآرایه، نمونه‌گیری از نمونه‌های

مختلف (برای مثال نمونه سرطانی و سالم)، استخراج آران ای و ساخت مکمل دیان ای از روی آن و همزمان نشان دار کردن نمونه ها با رنگ های سبز و قرمز (به ترتیب برای نمونه سالم و بیمار)، ریختن هر دو نمونه بر روی اسلاید، مراحل هیبرید شدن دیان ای (اتصال تک رشته های دیان ای به تراشه ریز آرایه)، شستشو و خشک کردن تراشه و در نهایت اسکن کردن تراشه ریز آرایه می باشد. به صورت مختصر مراحل آزمون در شکل (۱-۳) مشاهده می شود. بخش دوم تحلیل تصاویر حاصل شده از آزمایش می باشد. این دو بخش لازم و ملزوم یکدیگرند و باید سعی شود که به درستی انجام گیرد. در غیر این صورت نتیجه خوبی از آزمایش حاصل نخواهد شد [۹].

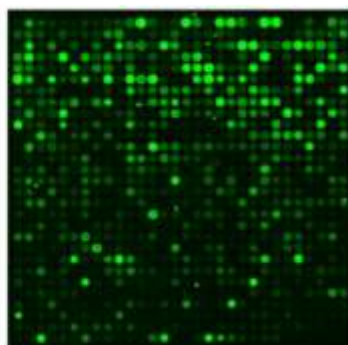


شکل ۱-۳- شمای کلی مراحل مختلف آزمون داده های ریز آرایه.

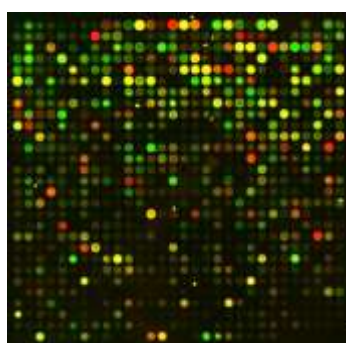
نتیجه آزمایش ریز آرایه به صورت تصویر است. بخشی از تصویر اسکن شده در شکل (۱-۴) مشاهده می شود. این تصویر به عنوان داده در اختیار ما قرار می گیرد. هر یک از نواحی دایره ای در تصویر ریز آرایه، یک خال نامیده می شود و در بردارنده اطلاعاتی در مورد ژن های تحت آزمایش می باشد. همانگونه که در این شکل مشاهده می شود، تصاویر ریز آرایه دارای نظم ماتریسی خاصی هستند. در شکل (۱-۴-الف) لایه سبز رنگ تصویر نماینده ژن های سالم می باشد. شدت روشنایی هر یک از خال ها نمایان گر این است که به چه میزان دیان ای در نمونه ی محلول سالم با ژن های موجود در خال ها اتصال برقرار کرده اند.



(ب)



(الف)



(ج)

شکل ۱-۴- بخشی از تصویر ریزآرایه. الف) لایه سبز رنگ در تصویر ریزآرایه، ب) لایه قرمز رنگ در تصویر ریزآرایه، ج) ادغام دو لایه تصویر ریزآرایه.

نقاط تاریک ژن‌هایی هستند که هیچ اتصالی با سلول‌های سالم برقرار نکرده‌اند. به مقدار اتصال نمونه‌ها بر روی تراشه‌ی ریزآرایه میزان هیبرید شدن یا هیبریداسیون^۱ ژن‌ها می‌نامند. برای خال‌های قرمز رنگ در شکل (۱-۴-ب) نیز می‌توانیم به صورت مشابه استدلال کنیم. با این تفاوت که خال‌های قرمز رنگ، نماینده‌ی دی‌ان‌ای در محلول نمونه‌ی سرطانی هستند. اگر رنگ سبز یا قرمز دارای شدت روشنائی بیشتری باشند، اتصال ژن‌های سالم یا بیمار قوی‌تر بوده است. دستگاه می‌تواند خال‌های موجود در هر دو تصویر را با هم ادغام کند تا شکل (۱-۴-ج) حاصل شود. خال‌های زرد به معنای بیان برابر دو ژن در دو نمونه (توموری و سالم) هستند؛ به عبارت دیگر اتصال ژن‌های سالم و بیمار در آن دو خال با هم برابر

¹ Hybridization

بوده و جهشی رخ نداده است. در مکانی که خال‌های سیاه موجود است، نمونه‌های بیمار و سالم هیچ اتصالی با ژن‌های آن قسمت برقرار نکرده‌اند [۱۰].

در بخش تحلیل تصاویر و داده‌های ریزآرایه با توجه به متنوع بودن فرمت تصاویر، وجود انواع مختلف تراشه‌های ریزآرایه، الگوریتم‌ها و تکنیک‌های زیادی ارائه شده است. با توجه به این تنوع و قطعی نبودن روش‌های ارائه شده جهت آنالیز درست داده‌ها هنوز هم محققین زیادی روی قسمت‌های مختلف آنالیز این داده‌ها تحقیق می‌کنند و مقالات زیادی در این زمینه ارائه می‌دهند.

۴-۱ تحلیل تصاویر ریزآرایه

با استفاده از فناوری ریزآرایه زمان لازم جهت دستیابی به اطلاعات ژن‌ها کاهش چشمگیری یافته است، ولی یکی از مشکلات اصلی در این زمینه زمان لازم جهت استخراج و تحلیل داده‌های خام بدست آمده از این آزمایش‌ها می‌باشد. تعداد خال‌ها در هر تصویر بسیار زیاد بوده و در برخی کاربردها ممکن است تا چندین هزار خال در یک تصویر وجود داشته باشد. بنابراین پردازش خودکار این تصاویر از اهمیت ویژه‌ای برخوردار است [۱۱-۱۳].

پس از تولید تصاویر ریزآرایه، جهت بدست آوردن اطلاعات حاصل از آزمایش، مراحل پردازشی مختلفی بر روی آن انجام می‌شود. بدین منظور از الگوریتم‌ها و تکنیک‌های پردازش تصویر به منظور کمی‌سازی تصاویر ریزآرایه یا به عبارت دیگر تعیین میزان بیان هر ژن استفاده می‌شود. سپس به دلیل وجود تعداد ژن‌های زیاد، حجم بالایی از اطلاعات در اختیار ما قرار می‌گیرد. به کمک تکنیک‌های داده کاوی و یادگیری ماشین تمامی داده‌های ریزآرایه تحلیل و دسته‌بندی می‌شوند. بنابراین مهم‌ترین مراحل برای تحلیل تصاویر ریزآرایه عبارت است از :

۱. پردازش تصویر

پردازش تصویر یک گام مهم در آزمون‌های ریزآرایه می‌باشد. این گام موجب افزایش دقت و صحت تحلیل داده‌های ریزآرایه می‌شود. در این مرحله کمی‌سازی سطح بیان ژن از تصاویر ریزآرایه در پنج مرحله صورت می‌گیرد، این مراحل عبارت‌اند از:

۱-۱- تشخیص زاویه چرخش: چرخش تصویر در هر دو جهت عقربه ساعت و بلعکس بر روی

محور x تخمین زده می‌شود و میزان چرخش تصحیح می‌گردد.

۲-۱- بلوک‌بندی: کل تصویر ریزآرایه توسط ماتریسی از چندین بلوک با اندازه‌های مشابه

تشکیل می‌شود. هدف از این بخش تعیین محل دقیق هر یک از بلوک‌ها می‌باشد.

۳-۱- شبکه‌بندی^۱: این مرحله شامل تقسیم‌بندی خال‌ها به سلول‌های مجزا و تعیین مکان هر

یک از خال‌ها در تصویر است.

۴-۱- قطعه‌بندی^۲: در این قسمت هدف جداسازی خال‌ها از پس‌زمینه است.

۵-۱- استخراج روشنایی^۳: در این مرحله هدف تعیین میزان روشنایی هر یک از خال‌ها است

که با استفاده از آن سطح بیان هر خال محاسبه می‌شود.

۲. داده‌کاوی

نکته مهمی که در داده‌های ریزآرایه وجود دارد، زیاد بودن تعداد ژن‌ها (ویژگی‌ها) و اندک بودن تعداد نمونه‌ها می‌باشد. این عامل باعث بروز مشکلاتی در تحلیل داده‌ها می‌شود. مشکلاتی از قبیل اطلاعات غیرمفید در برخی از ژن‌ها نه تنها در تحلیل اطلاعات مفید نیستند، بلکه باعث طبقه‌بندی نادرست اطلاعات می‌شوند. بنابراین تکنیک‌های داده‌کاوی جهت حل این مشکلات و ارائه مدل ساده‌تر و دقیق‌تر استفاده می‌شود. در نتیجه با انتخاب ویژگی^۴ مناسب می‌توانیم ژن‌هایی که دارای اطلاعات مفیدی هستند را استخراج و تنها آن‌ها را مورد بررسی قرار دهیم.

^۱Gridding

^۲Segmentation

^۳Intensity Extraction

^۴ Feature selection

۳. یادگیری ماشین

دسته‌بندی^۱ صحیح داده‌های ریزآرایه در جهت تشخیص نوع سرطان، با استفاده از الگوریتم‌های موجود در یادگیری ماشین برای تعیین مرحله‌ی درمانی بیمار یا تجویز دارو نقش به‌سزایی دارد. دسته‌بندی دارای تکنیک‌های متنوعی است که در تحقیقات مختلف از آن‌ها استفاده می‌شود. مهم‌ترین و اساسی‌ترین گام‌ها شامل شبکه‌بندی، قطعه‌بندی، انتخاب ویژگی و دسته‌بندی تومورهای سرطانی با استفاده از تصاویر ریزآرایه و بیان ژن‌ها می‌باشد. در ادامه هر گام به صورت مختصر شرح داده می‌شود.

۱-۴-۱ شبکه‌بندی

شبکه‌بندی تصویر به معنی جداسازی خال در تصویر ریز آرایه است، به عبارت دیگر به شناسایی موقعیت هر خال در تصویر ریز آرایه اطلاق می‌شود. روش‌های متفاوتی برای انجام این کار وجود دارد، از مهم‌ترین روش‌ها می‌توان به شبکه‌بندی با استفاده از روش ماشین بردار پشتیبان^۲ اشاره کرد. در این روش ابتدا خال‌ها بر روی تصویر شناسایی، سپس تمامی آن‌ها به عنوان داده‌ی ورودی در ماشین بردار پشتیبان دو کلاس مورد استفاده قرار می‌گیرند. خط جدا کننده، فاصله‌ی مابین سطر و ستون را تخمین می‌زند [۱۴].

در برخی از پژوهش‌ها با کمک تکنیک‌های آستانه‌گیری نظیر اتسو^۳ موقعیت هر خال در تصویر شناسایی می‌شود. در این روش با محاسبه‌ی مجموع پیکسل‌های سطری / ستونی، تابع یک بعدی حاصل می‌شود که به کمک آستانه‌گذاری اتسو، فواصل مابین تمامی خال‌ها تخمین زده می‌شود [۱۵]. این روش برای تصاویر نویزی از دقت پایینی برخوردار است.

برخی از محققین نیز فرض را بر این بنا داشته‌اند که تمامی خال‌ها در فواصل برابری از یکدیگر قرار گرفته‌اند. با توجه به این فرضیه روشی خودکار برای تعیین موقعیت هر خال طراحی کرده‌اند. در این

¹ Classification

² Support Vector Machine (SVM)

³ Otsu

روش ابتدا بر روی تصویر پیش‌پردازشی همانند اعمال عملگر ریخت‌شناسی^۱ صورت می‌گیرد. سپس با محاسبه مجموع سطرها / ستون‌ها فاصله بین خال‌ها محاسبه می‌شود. بدین منظور بعد از محاسبه مجموع پیکسل‌ها، تعداد پیکسل‌های نزدیک به صفر و غیرصفر، تعیین می‌شود. پیکسل‌های صفر نماینده فضای خالی بین دو خال می‌باشد. مقدار محاسبه شده از میانگین‌گیری این فواصل به عنوان فاصله نهایی مابین تمامی خال‌ها در نظر گرفته می‌شود [۱۶].

۲-۴-۱ قطعه‌بندی

در فرآیند پردازش تصویر، قطعه‌بندی عبارت است از تقسیم‌بندی تصویر به نواحی مختلف و مستقل به طوری که هر ناحیه دارای خصوصیات خاصی می‌باشد. قطعه‌بندی در تصاویر ریزآرایه عبارت است از دسته‌بندی تصویر به دو ناحیه‌ی پیش‌زمینه^۲ و پس‌زمینه^۳. ناحیه‌ی اول شامل نقاط متعلق به خال‌ها بوده و ناحیه‌ی دوم شامل فضای موجود در اطراف خال‌ها می‌باشد. از مشهورترین روش‌های قطعه‌بندی تصاویر ریزآرایه که از سال ۲۰۰۲ تاکنون مورد استقبال محققین قرار گرفته است، می‌توان به چهار روش زیر اشاره کرد [۱۷]:

- *قطعه‌بندی براساس دایره ثابت*^۴: از یک ماسک^۵ دایره‌ای با قطر ثابت برای جداسازی خال‌ها استفاده می‌شود. این روش فقط قادر است خال‌هایی که قطرشان برابر این ماسک است را قطعه‌بندی کند. پیاده‌سازی این روش بسیار ساده است، اما نتایج قابل قبولی ندارد [۱۸، ۱۹].
- *قطعه‌بندی براساس شکل وقتی*^۶: در این روش، قطعه‌بندی خال‌ها کاملاً مستقل از شکل هندسی آن‌ها صورت می‌گیرد. روش رایج برای این قطعه‌بند براساس روش رشد نقاط اولیه^۷ می‌باشد. این روش خودکار بوده و تنها نقطه ضعف آن مشخص کردن تعداد و محل نقاط شروع الگوریتم

^۱ Morphology

^۲ Foreground

^۳ Background

^۴ Fixed Circle Segmentation

^۵ Mask

^۶ Adaptive Shape Segmentation

^۷ Seeded region growing

می‌باشد. انتخاب این نقاط آغازین، معمولاً توسط کاربر به صورت دستی انجام می‌شود [۱۹].

[۲۰].

- *قطعه‌بندی براساس هیستوگرام*^۱: در این روش، براساس رسم هیستوگرام تصویر، خال‌های مربوطه از تصویر جدا می‌گردد. در این روش یک حدآستانه در نظر گرفته می‌شود، مقداری که بیشتر از حدآستانه باشد به عنوان مقادیر خال‌ها و مابقی به عنوان پس‌زمینه در نظر گرفته می‌شود. [۲۱].

- *قطعه‌بندی براساس دایره وقفی*^۲: در این روش نیز مانند روش اول از ماسک‌های دایره‌ای جهت قطعه‌بندی استفاده می‌شود، ولی شعاع این ماسک‌ها برای هر خال، به صورت مستقل توسط نرم‌افزار تخمین زده می‌شود. بخش‌بندی براساس دایره وقفی زمانی از کارایی لازم برخوردار است که شکل خال‌ها در تصویر به صورت دایره‌ای باشد [۱۸].

بسیاری از نرم‌افزارها از جمله GenePix, ScanAlyze و ImaGene برای استخراج خال‌ها از روش‌های مذکور الهام گرفته‌اند [۲۲-۲۴].

در برخی از پژوهش‌ها با الهام از روش‌های خوشه‌بندی همانند k-means و ماشین بردار پشتیبان عمل قطعه‌بندی خال‌ها صورت می‌گیرد. در فصل دوم به صورت مفصل این روش‌ها توضیح داده می‌شود.

از آنجایی که مقدار پس‌زمینه در هر یک از اسلایدها شامل برخی از رشته‌های دی‌ان‌ای جفت نشده یا مربوط به مواد شیمیایی است که در زمان آزمایش بر روی تراشه باقی مانده‌اند. بنابراین با استفاده از روش‌های موجود پس از قطعه‌بندی، مرحله‌ای دیگر به منظور محاسبه صحیح میزان سطح خاکستری زن‌ها به نام تصحیح پس‌زمینه نیاز است. روش‌های متفاوتی برای تخمین و حذف پس‌زمینه در مقالات مختلف در نظر گرفته شده است. از جمله: ۱- پس‌زمینه محلی ۲- عملگر ریخت‌شناسی ۳- پس‌زمینه ثابت ۴- در نظر نگرفتن پس‌زمینه در ادامه به صورت خیلی مختصر هر یک از روش‌ها توضیح می‌دهیم:

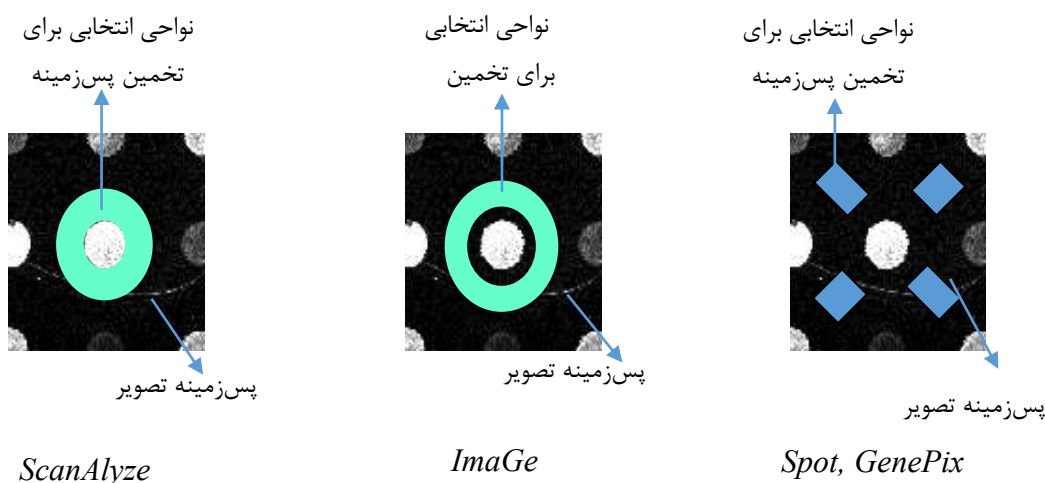
^۱ Histogram Segmentaion

^۲ Adaptive Circle Segmentation

۱. پس زمینه محلی:

یک قسمت کوچک نزدیک به خال انتخاب و میانگین پیکسل‌ها محاسبه می‌شود. مقدار محاسبه شده، به عنوان مقدار پس زمینه در نظر گرفته می‌شود. (نکته: برای هر یک از خال‌ها به صورت مجزا این مقدار محاسبه می‌شود).

نرم‌افزارهایی از قبیل *GenePix*, *ScanAlyze* و *ImaGene* با استفاده از روش پس زمینه محلی، این مقدار را تخمین می‌زنند. با توجه به شکل (۱-۵) هر روش منطقه متفاوتی را برای تخمین پس زمینه در نظر می‌گیرد. ممکن است گاهی بدلیل وجود نویز مقدار میانگین شدت روشنایی پیکسل‌ها در منطقه محلی انتخابی، بیشتر از مقدار خال‌ها باشد. بنابراین هنگام تصحیح پس زمینه محلی ممکن است پیکسل‌های مربوط به خال مقادیر منفی به خود بگیرند. این تخمین اشتباه منجر به ایجاد مشکلاتی در ادامه کار می‌شود [۲۵].



شکل ۱-۵- چگونگی انتخاب مناطق محلی برای تخمین پس زمینه در نرم‌افزارهای مختلف.

۲. پس زمینه با مقدار ثابت

این روش به صورت سراسری برای تمامی خال‌ها استفاده می‌شود. بدین صورت، ابتدا میزان شدت روشنایی خال محاسبه و سپس سه درصد آن به عنوان پس زمینه در نظر گرفته می‌شود. نرم افزار

ScanAlyze علاوه بر روش قبلی از پس‌زمینه با مقدار ثابت برای تخمین این مقدار استفاده می‌کند.

این روش به ندرت در برنامه‌های کاربردی بکار برده می‌شود [۲۶].

۳. روش ریخت‌شناسی

در سال ۱۹۹۹ این روش توسط Soille بکار گرفته شد [۲۷]. در این روش شدت روشنایی پس‌زمینه با استفاده از یک فیلتر غیرخطی تخمین زده می‌شود. بنابراین به صورت غیرخطی اساساً تمام تصویر هموار می‌شود. علاوه بر خال‌ها، لکه‌ها و لبه‌های تیز نیز حذف یا هموار می‌گردند و تصویر حاصل شده به عنوان تصویر پس‌زمینه خال‌ها برآورد می‌شود. با کم کردن این تصویر از تصویر اصلی، پس‌زمینه از خال‌ها حذف می‌گردد و مقادیر باقی‌مانده نماینده میزان هیبرید شدن ژن‌های مورد نظر در خال‌ها می‌باشد. در این روش بعد از محاسبه پس‌زمینه و کم کردن آن از تصویر اصلی، به ندرت اتفاق می‌افتد که مقادیر خال منفی شود. استفاده از عملگرهای سایش و سپس گسترش در ریخت‌شناسی برای این روش پیشنهاد شده است. اصولاً تعیین اندازه و ساختار عنصری برای عملگرهای ریخت‌شناسی در تصاویر متفاوت ریزآرایه، به صورت آزمایش و خطا بدست می‌آید.

همان‌طور که ذکر شد در اکثر روش‌های موجود ابتدا خال‌ها با در نظر گرفتن پس‌زمینه قطعه‌بندی شده و سپس با اجرای الگوریتم‌هایی نظیر ریخت‌شناسی، مقدار پس‌زمینه آن‌ها تشخیص و تصحیح می‌گردد. در حالیکه با استفاده از یک روش قوی تخمین پس‌زمینه، نیازی به اجرای هر دو مرحله نیست. با اجرای الگوریتم تخمین پس‌زمینه بر روی هر یک از خال‌ها به صورت مجزا و حذف مقدار پس‌زمینه از تصویر، قطعه‌بندی و تصحیح پس‌زمینه همزمان صورت می‌گیرد.

۱-۴-۳ انتخاب ویژگی

به دلیل حجم انبوهی از داده‌های به دست آمده از ریزآرایه، بررسی و کشف اطلاعات مفید بدون استفاده از روش‌های آماری و تکنیک‌های یادگیری ماشین ممکن نیست. در نتیجه باید قبل از عمل دسته‌بندی نوعی ساده‌سازی و کاهش اطلاعات یا به عبارتی انتخاب ژن‌های مفید و تاثیرگذار از بین هزاران ژن

صورت گیرد؛ تا بتوان تفسیر دقیق‌تری از نقش ژن‌ها در بیماری‌ها ارائه داد و طبقه‌بندی بهتری از بیماران با این داده‌ها بدست آورد. روش‌های مختلفی برای انتخاب ژن‌ها وجود دارد که عبارتند از: Sung-bae cho و همکارانش در سال ۲۰۰۳ برای انتخاب ژن از چندین روش انتخاب ویژگی از جمله ضریب همبستگی پیرسون، ضریب کوسینوسی، استفاده از اطلاعات متقابل استفاده کرده‌اند. هر نمونه شامل ۷۱۲۹ ژن بوده که از بین آن‌ها ۲۵ ژن برای دسته‌بندی داده‌ها انتخاب شده‌است [۲۸]. Ming-Chi Tasi و همکارش در سال ۲۰۰۷ برای انتخاب ویژگی تنها با به کارگیری روش بهره‌ی اطلاعاتی^۱، ژن‌های مفید را استخراج کرده است [۲۹]. همچنین Guesh Kumer و همکارانش در سال ۲۰۱۲ با استفاده از روش اطلاعات متقابل از بین ۲۰۰۰ ژن، تنها ۱۰ ژن مفید برای تعیین نوع بیماری را انتخاب کردند [۳۰]. در دو مرجع [۳۱، ۳۲] در سال ۲۰۱۴ ژن‌های مفید از بین هزاران ژن با ترکیب روش‌های تفکیک کلاسی^۲، آزمون t^3 و استفاده از بهره اطلاعاتی، استخراج شده است.

در بین پژوهشگران ایرانی نیز سید محمد تقی جوادی و همکارانش در اسفند سال ۱۳۸۷ و همچنین فاطمه امین‌زاده و همکارانش در سال ۱۳۸۹ برای کاهش ابعاد داده‌های ریزآرایه‌ی خود به ترتیب روش‌های بهره اطلاعاتی و $FCBF^4$ بر روی داده‌های خود اعمال کرده‌اند [۳۳، ۳۴]. در تمامی روش‌های بالا هدف از انتخاب ویژگی، انتخاب یک زیر مجموعه‌ی کوچکی از ویژگی‌ها است، به طوری که توزیع آن‌ها حد امکان به توزیع کلاس اصلی نزدیک باشد.

۵-۴-۱ طبقه‌بندی داده‌ها

دسته‌بندی صحیح داده‌های ریزآرایه مربوط به بیماری‌های متفاوت، با استفاده از تحلیل میزان بیان ژن‌های استخراج شده برای تعیین مرحله درمانی بیمار، یا تجویز دارو نقش به‌سزایی دارد. برای دسته‌بندی تکنیک‌های متنوعی وجود دارد که در تحقیقات مختلف از آن‌ها استفاده می‌شود. از تکنیک‌های

¹ Information Gain

² Class Separability

³ T_test

⁴ Fast Correlation-Based Filter

رایج در طبقه‌بندی داده‌های ریزآرایه می‌توان به مواردی از جمله، شبکه عصبی^۱، ماشین بردار پشتیبانی، طبقه‌بندی بیزین^۲، سیستم‌های خبره^۳، الگوریتم‌های ژنتیک^۴ و درخت تصمیم‌گیری^۵ اشاره کرد. 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 [۳۵]. برخی محققین نیز با این تفکر که اغلب کلاس‌بندها وابسته به کاربرد خاصی هستند، به منظور بهبود نتایج، از ترکیب دسته-بندکننده‌های متفاوت استفاده می‌کنند، همانند Wang و همکارانش که در سال ۲۰۰۷ با ادغام دو روش بردار ماشین پشتیبان و شبکه عصبی فازی داده‌های خود را دسته‌بندی می‌کنند [۳۶]. همچنین Halder و همکارانش در سال ۲۰۱۵ با استفاده از الگوریتم k نزدیک‌ترین همسایه به صورت فازی داده‌های ریزآرایه خود را دسته‌بندی کرده‌اند [۳۷].

همان‌طور که مطالعه کردید، روش‌های فراوانی برای دسته‌بندی داده‌های ریزآرایه پیشنهاد شده است. هر کدام از این روش‌ها با توجه به پایگاه داده‌ی مورد استفاده خودشان، دقت بالایی برای تفکیک آن‌ها ارائه می‌دهند. الگوریتم‌های دسته‌بندی باید با توجه به نیاز فرد خبره و کارایی آن‌ها انتخاب می‌شود. تقریباً در تمامی روش‌های ذکر شده، برای رسیدن به نتیجه نهایی الگوریتم‌ها همانند جعبه سیاه عمل می‌کنند و مسیر طی شده برای دسته‌بندی داده‌های ریزآرایه در اختیار فرد خبره قرار نمی‌گیرد.

۵-۱ کاربرد ریزآرایه در تشخیص سرطان

سرطان چهارمین بیماری شایع و یکی از علل بسیار مهم مرگ و میر در سراسر جهان است که هر ساله منجر به بیش از ۶ میلیون مورد مرگ می‌شود [۳۸]. علی‌رغم پیشرفت‌های تکنولوژی طی یک دهه‌ی گذشته در زمینه سرطان، موقعیت‌های محدودی برای مقابله با این بیماری ویران‌گر حاصل شده است. بنابراین تشخیص دقیق و به موقع آن، از اهمیت به‌سزایی در مطالعات و تحقیقات برخوردار است. لذا به

¹ Neural Network

² Bayesian Classification

³ Expert System

⁴ Genetic Algorithms

⁵ Decision Tree

کارگیری مناسب تکنولوژی جدید همانند فناوری ریزآرایه به منظور پیش‌بینی و دسته‌بندی سرطان از اهمیت روزافزونی برخوردار شده است.

فناوری ریزآرایه در سال ۱۹۹۹ معرفی و با عناوین آرایه‌های دی‌ان‌ای، تراشه‌های ژنی، تراشه‌های دی-ان‌ای و تراشه‌های زیستی نام‌گذاری شده است. فناوری ریزآرایه یکی از آخرین پیشرفت‌ها در زمینه زیست‌شناسی ملکولی است که اجازه نظارت بر بیان هزاران ژن را به صورت همزمان تنها در یک آزمایش هیبریداسیون می‌دهد. علاوه بر این پتانسیل علمی این فناوری در مطالعه‌ی بنیادی بیان ژن، یعنی تنظیم و تعاملات ژن‌ها، کاربردهای مهمی در پژوهش‌های دارویی و کلینیکی دارد. برای مثال، با مقایسه بیان ژن در سلول‌های سالم و ناسالم، ریزآرایه می‌تواند در شناسایی ژن‌های ناسالم برای داروهای درمانی یا ارزیابی تاثیر آن‌ها استفاده شود [۳۸].

در مورد بیماری‌های خطرناک مانند انواع سرطان‌ها، ریزآرایه به عنوان ابزار نیرومندی برای مطالعه‌ی دی‌ان‌ای و همچنین بیان ژن در سرطان محسوب می‌شود. چرا که یکی از مهم‌ترین عوامل سرطان ایجاد جهش‌های ژنتیکی در بیمار است. برای تشخیص جهش کافی است تفاوت دو نمونه سالم و توموری را محاسبه کنیم [۳۹]. بنابراین راهکارهای مورد استفاده در مطالعات الگوی سرطان با استفاده از بیان ژن در ریزآرایه عبارتند از:

- *مطالعه تومور در مقابل نمونه مرجع (شاهد)*، که در آن الگوی بیان ژن این دو، با هم مقایسه می‌شود تا اختلافات و تشابهات آن‌ها معین شود.
- *طبقه‌بندی سرطان*، که در آن الگوی بیان ژن در نمونه‌های متعدد یک نوع سرطان مقایسه می‌شود تا زیر گروه‌های مجزا، آشکار و سرطان به صورت مولکولی طبقه‌بندی شود.
- *ارزیابی دوره سرطان*، که در آن الگوی بیان ژن در نمونه‌های تومور در مراحل مختلف مقایسه می‌شوند تا اختلافات بین مراحل اولیه و پیشرفته بیماری روشن شود.

۶-۱ اهداف پایان نامه

یکی از پیشرفت‌های مهم در زمینه‌ی ژنتیک، استفاده از فناوری ریزآرایه است. امروزه ریزآرایه‌ها به ابزار بسیار قدرتمندی جهت مطالعه‌ی همزمان رفتار هزاران ژن تبدیل شده است. صنایع عظیم داروسازی جهان با آگاهی از توانمندی‌های گسترده این فناوری در آینده آزمایشگاه‌های پزشکی و سایر علوم، سرمایه‌گذاری‌های کلانی در این زمینه تخصیص داده و در حال حاضر انواع سیستم‌های تهیه ریزآرایه‌ها و نرم‌افزارهای مورد نیاز آن را با قیمت‌های نسبتاً بالایی به بازار عرضه می‌کنند. همچنین در برخی از کارخانه‌های داروسازی، سالانه ده‌ها هزار تصویر ریزآرایه با ابعاد مختلف تولید می‌شود. پردازش این تصاویر، از مراحل زمان‌بر در آزمایش بوده است. بنابراین پردازش خودکار این تصاویر و تحلیل داده‌های ریزآرایه از اهمیت بالایی برخوردار است.

متأسفانه در ایران آزمایشگاهی برای تحلیل تصاویر ریزآرایه وجود ندارد. بنابراین تمامی تصاویر تهیه شده به وسیله آزمون ریزآرایه، به کشورهای مختلفی از جمله آمریکا و فرانسه فرستاده می‌شود. افراد متخصص در زمینه‌ی بیوانفورماتیک نیز برای دریافت پاسخ، هزینه گزاف و زمان زیادی صرف می‌کنند. همچنین سیستم‌های موجود در این زمینه اکثراً به صورت نیمه‌خودکار عمل می‌کنند و در تصاویر پر نویز اطلاعات را به درستی به فرد خبره ابلاغ نمی‌نمایند.

هدف از این تحقیق، ارائه یک سیستم خودکار جهت استخراج و تحلیل اطلاعات از تصاویر ریزآرایه، از جمله تشخیص نوع بیماری سرطان می‌باشد. این سیستم از سه گام اصلی تشکیل شده است. در گام اول با کمک پردازش تصویر، تصاویر ریزآرایه کمی‌سازی شده و در دو گام بعدی به ترتیب داده‌های مهم از بین هزاران ژن انتخاب و نوع بیماری تشخیص داده می‌شود. با عرضه‌ی این سیستم، کمک بسیار شایانی در صنعت فناوری ریزآرایه در ایران می‌شود. چرا که این روش از بازده بسیار بالایی برخوردار بوده و در مدت زمان کوتاهی قادر به تحلیل میزان قابل توجهی از اطلاعات می‌باشد. همچنین این سیستم منجر به کاهش هزینه و زمان در تحلیل تصاویر ریزآرایه در آزمایشگاه‌های ایران می‌شود و می‌تواند در یافتن راه حل‌های درمانی بسیاری از بیماری‌های مهم همانند سرطان پیش‌تاز باشد.

۷-۱ ساختار پایان نامه

در فصل دوم مروری بر تحقیقات پیشین در زمینه‌ی نرم‌افزارهای تحلیل ریز آرایه ، صورت می‌گیرد. در فصل سوم سیستم جامعی معرفی خواهد شد که در ابتدا با استفاده از روش‌های پردازش تصویر داده‌های خام تصاویر ریزآرایه استخراج شود، سپس با کمک تکنیک‌های داده‌کاوی و یادگیری ماشین داده‌های محاسبه شده به منظور تشخیص نوع سرطان دسته‌بندی می‌نماید. در فصل چهارم سیستم طراحی شده بر روی سه پایگاه داده که شامل سه نوع از تصاویر ریزآرایه سرطانی هستند، آزمایش می‌شود و نتیجه‌ی عملکرد بخش‌های مختلف سیستم بر روی تصاویر و داده‌های ریزآرایه مورد تجزیه و تحلیل قرار می‌گیرد و در نهایت سیستم پیشنهادی با چندین سیستم موجود مقایسه می‌شود. در آخرین فصل نیز به جمع‌بندی کلی پایان‌نامه پرداخته و پیشنهاداتی برای بهبود عملکرد سیستم جهت کارهای آینده مطرح می‌گردد.

۸-۱ نتیجه‌گیری

در این فصل ابتدا به صورت مختصر ساختار دی‌ان‌ای و فناوری ریزآرایه معرفی گردید. سپس قسمت‌های مهم و اساسی تحلیل تصاویر حاصل از این فناوری در چهار گام متفاوت شبکه‌بندی، قطعه‌بندی، انتخاب ویژگی و طبقه‌بندی معرفی و مورد بررسی قرار گرفت. در پایان، کاربردهای گوناگون این فناوری در زمینه‌ی درمان بیماری و پیشگیری سرطان شرح داده شد.

فصل دوم

مروری بر تحقیقات پیشین

۱-۲ مقدمه

ریزآرایه‌ها ابزار بسیار مناسبی جهت مطالعه و آنالیز رفتار هزاران ژن به طور همزمان بوده و نقش مهمی را در کشف و تشخیص بیماری‌ها و همچنین بررسی روش‌های درمان بر عهده دارند. یکی از مهم‌تری بخش‌های آزمایش ریزآرایه، پردازش تصویر و آنالیز تصاویر حاصل از آن است. بخش تحلیل تصاویر ریزآرایه به صورت نرم‌افزار در آزمایشگاه‌های بیولوژیکی مورد استفاده قرار می‌گیرد. سرمایه‌گذاری‌های زیادی که در دهه‌ی اخیر در جهت توسعه و گسترش آن به عمل آمده، هر ساله میزان قابل توجهی بر حجم دانش ژنتیک ملکولی بشر افزوده و اکثر نسخه‌های به روز نرم‌افزارهای تحلیل این تصاویر در اختیار عموم قرار نمی‌گیرد. بنابراین متخصصان بیولوژیکی برای تحلیل هر تصویر ریزآرایه‌ی خود، مجبور به پرداخت هزینه‌های هنگفت می‌شوند. نمونه‌ای از این هزینه‌ها در مراجع [۴۰-۴۲] آورده شده است. در بخش بعدی، عملکرد نرم‌افزارهای پیشین و الگوریتم‌ها متفاوت، بررسی و توصیف می‌شوند.

۲-۲ مروری بر نرم‌افزارهای موجود

به منظور بررسی عملکرد سیستم‌های موجود در زمینه‌ی کمی‌سازی تصاویر ریزآرایه، نسخه‌نهایی سه نرم‌افزار مشهور GenePix، ScanAlyze و ImaGene در قالب نمایشی^۱ تهیه و مورد تحلیل قرار گرفته است. در جدول (۱-۲) اطلاعات جامع و مختصری برای هر نرم‌افزار ذکر شده است. این نرم‌افزارها فقط برای مقاصد پژوهشی استفاده می‌شوند و تقریباً عملکرد مشابهی دارند.

همانطور که قبلاً ذکر شد، مهم‌ترین مراحل برای تحلیل تصاویر ریزآرایه شامل: تعیین زاویه چرخش و تصحیح آن، شناسایی موقعیت بلوک‌ها، شناسایی موقعیت تمامی خال‌ها در هر بلوک، قطعه‌بندی، تصحیح پس‌زمینه، محاسبه‌ی بیان ژن‌ها، کاهش ویژگی داده‌ها و در نهایت دسته‌بندی داده‌های ریزآرایه می‌باشد. در نرم‌افزارهای ذکر شده، تنها قابلیت کمی‌سازی تصاویر ریزآرایه‌ی مربوط به بخش پردازش

¹ Demo

تصویر وجود دارد و تحلیل داده‌های ریزآرایه در نرم‌افزار امکان‌پذیر نیست. به عبارت دیگر امکان مقایسه دو داده‌ی ریزآرایه به صورت همزمان میسر نمی‌باشد.

جدول ۲-۱- معرفی نرم‌افزارهای تحلیل تصاویر ریزآرایه.

نرم‌افزار	محصول	تاریخ شروع	نسخه نهایی	سال عرضه	آدرس دسترسی به نرم‌افزار
GenePix	Axon Instruments, Molecular Devices	1999	7.2.29	2013	http://mdc.custhelp.com
ImaGene	BioDiscovery	1997	9.0	2012	http://biodiscovery.com
ScanAlyze	Michael Eisen Lab	1999	2.50	2002	http://eisenlab.org

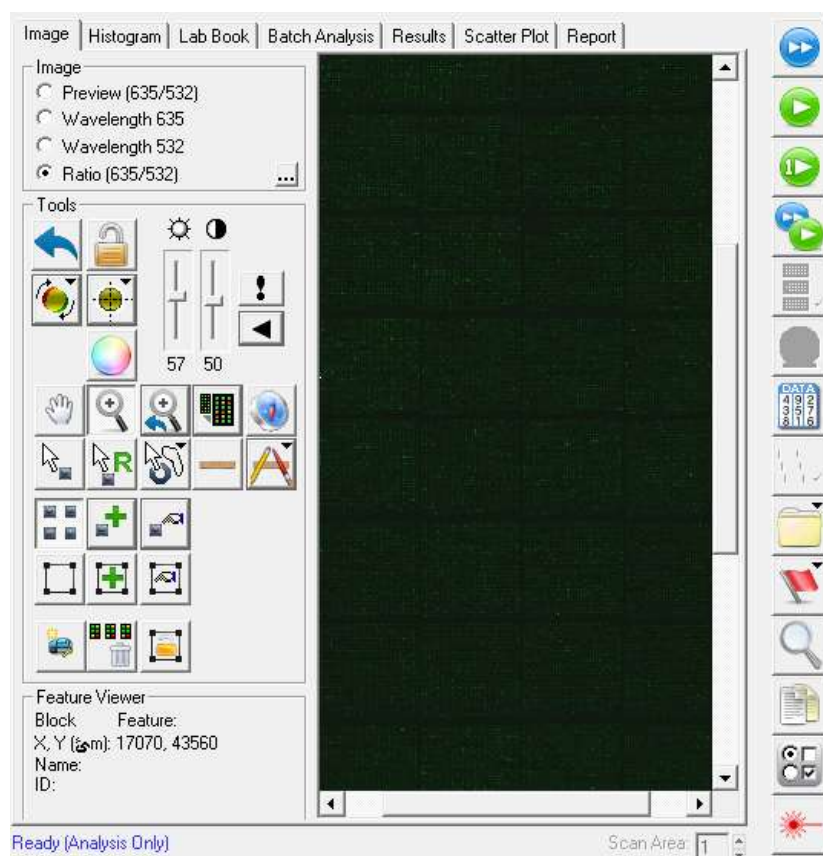
در ادامه تمامی مراحل کمی‌سازی تصاویر ریزآرایه در دو نرم‌افزار GenePix، ImaGene و ScanAlyze شرح داده می‌شود.

۲-۲-۱ نرم‌افزار GenePix

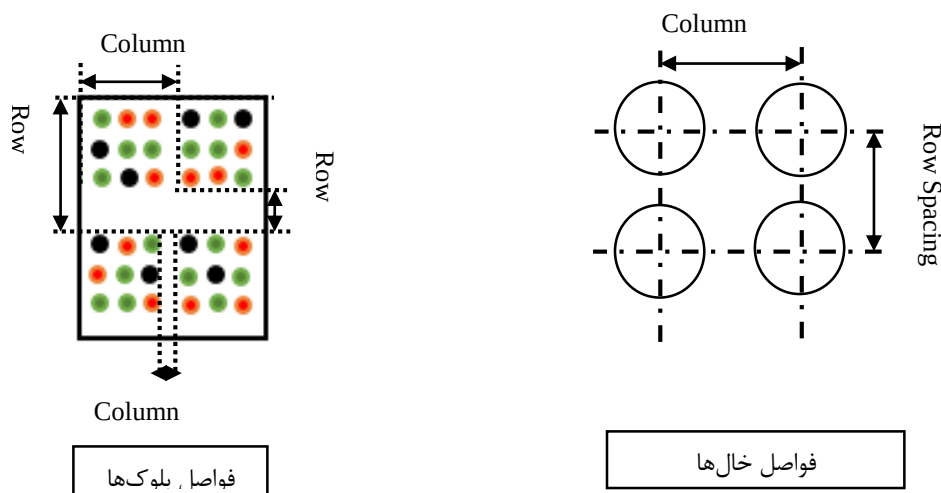
نرم‌افزار GenePix محصول مشترک دو شرکت Axon Instruments و Molecular Devices توجه بسیاری از افراد را به خود جلب کرده است. این شرکت فعالیت خود را در سال ۱۹۹۹ شروع و هر سال نسخه‌های جدیدی به بازار عرضه می‌دهد. این نرم‌افزار برای تحلیل تصاویر ریزآرایه و فقط به منظور مقاصد پژوهشی در نظر گرفته شده است [۴۳]. فرمت متفاوتی از متن و تصویر توسط GenePix پشتیبانی می‌شود. مهم‌ترین این فرمت‌ها عبارت‌اند از: فرمت *atf gal gpr gpl jpeg tif*. تصاویر ورودی نرم‌افزار باید در فرمت tif یا jpeg تهیه شوند و مابقی فرمت‌های تصویر در این نرم‌افزار شناسایی نمی‌شوند. فضای کاربری نرم‌افزار GenePix در شکل (۲-۱) نمایش داده شده است.

گاهی برخی از تراشه‌های ریزآرایه به صورت زاویه‌دار داخل دستگاه اسکن قرار گرفته می‌شوند و منجر به ایجاد تصاویر به صورت زاویه دار می‌گردند. بنابراین اولین مرحله در هر یک از این نرم‌افزارها؛ تشخیص و تصحیح میزان زاویه چرخش (مرحله اصلاح چرخش تصویر) می‌باشد. میزان چرخش تصویر در هر دو جهت عقربه ساعت و بالعکس امکان پذیر می‌باشد. این نرم‌افزار قابلیت تشخیص میزان زاویه چرخش و تصحیح زاویه به صورت خودکار را ندارد. بدین منظور کاربر به صورت آزمایش و خطا میزان چرخش

تصویر را تعیین کرده و به سیستم اعمال می‌کند. سیستم با توجه به زاویه‌ی دریافتی، جهت تصویر ریزآرایه را تغییر می‌دهد. اگر تصویر ریزآرایه متشکل از چندین بلوک باشد، نرم‌افزار قابلیت چرخاندن تصویر را به صورت کامل ندارد و برای رفع چرخش تصویر، برای هر بلوک میزان چرخش مجزایی دریافت، و یک به یک بلوک‌ها به صورت جداگانه تصحیح می‌گردند. بعد از تصحیح تقریبی زاویه‌ی چرخش، برای شناسایی موقعیت بلوک‌ها و خال‌ها، کاربر باید از تعداد دقیق آن‌ها در هر سطر و ستون آگاهی داشته باشد. علاوه بر این، با توجه به این حقیقت که نرم‌افزار فاصله مابین خال‌ها و بلوک‌ها را برای کاربر اندازه‌گیری نمی‌کند، این فواصل باید به صورت دستی از طریق کاربر وارد سیستم شوند. این مقادیر نیز می‌توانند به صورت آزمایش و خطا تعیین شوند (شکل ۲-۲). در صورت بروز هرگونه خطا در این مراحل، مابقی روند کار نیز با خطا تعیین خواهد شد.



شکل ۲-۱ - محیط کاربری نرم‌افزار GenePix.



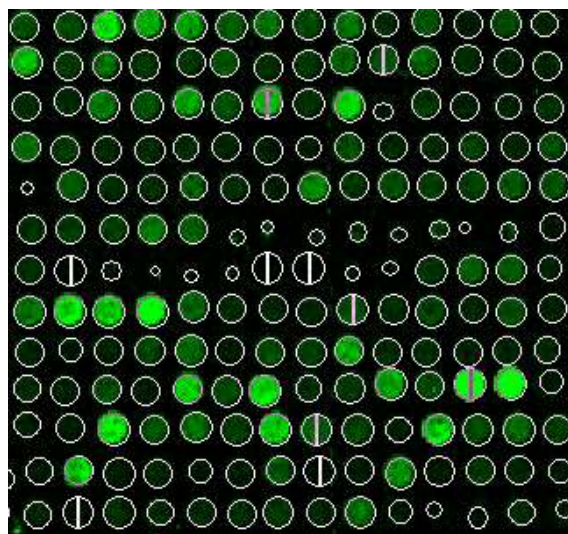
شکل ۲-۲- فواصل مابین خال‌ها و بلوک‌ها نمایش داده شده در نرم‌افزار GenePix.

در این نرم‌افزار برای قطعه‌بندی خال‌ها در تصویر ریزآرایه همان‌طور که در فصل اول شرح داده شد از روش قطعه‌بندی براساس دایره‌ی افقی، دایره‌ی ثابت و شکل افقی الهام گرفته شده است. با انتخاب هر روش، عملیات قطعه‌بندی بر روی تصویر اعمال می‌شود. نمونه‌ای از قطعه‌بندی براساس دایره‌ی افقی در شکل (۲-۳) مشاهده می‌شود. در این قطعه‌بندی کاربر موظف است دامنه تغییرات شعاع دایره‌ها را وارد سیستم نمایند. سپس شعاع هر خال با توجه به دامنه‌ی تعریف شده توسط نرم‌افزار تخمین زده می‌شود. یکی از مهم‌ترین قسمت‌های این مرحله انتخاب و تنظیم دامنه تغییرات شعاع می‌باشد و تمامی مراحل بعد را تحت تاثیر خود قرار می‌دهد. در این روش اگر خال‌هایی وجود داشته باشند که یکی از شرایط زیر را شامل شوند، ممکن است تشخیص داده نشوند:

الف) خال‌هایی که کمتر از ۵ پیکسل داشته باشند.

ب) خال‌هایی که قطر آن‌ها بزرگتر از مقدار تعیین شده باشد.

ج) خال‌هایی که با هم هم‌پوشانی داشته باشند.

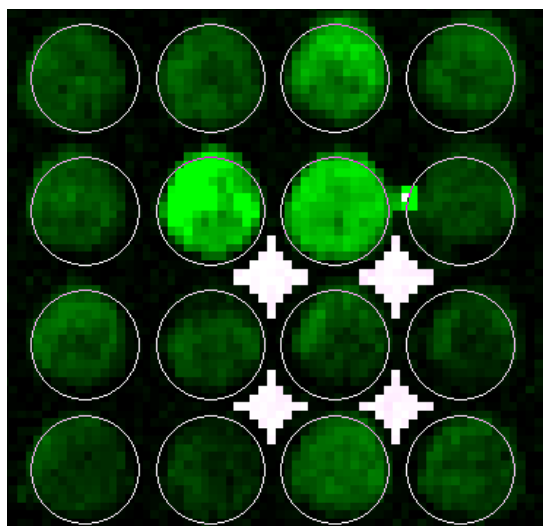


شکل ۲-۳- استفاده از روش قطعه‌بندی خال‌ها براساس دایره‌ی افقی در نرم‌افزار GenePix.

در گام بعدی یعنی تصحیح پس‌زمینه، نرم‌افزار از سه روش تصحیح پس‌زمینه محلی^۱، تصحیح پس‌زمینه سراسری^۲ و ریخت‌شناسی بهره می‌برد. در روش محلی ابتدا یک قسمت کوچک نزدیک به خال انتخاب، سپس میانگین پیکسل‌ها محاسبه می‌شود. مقدار محاسبه شده به عنوان مقدار پس‌زمینه در نظر گرفته می‌شود. در شکل (۲-۴) منطقه‌ای که نرم‌افزار برای تخمین در نظر می‌گیرد، نمایش داده می‌شود. در روش محلی مقدار پس‌زمینه برای هر خال به صورت مجزا تخمین زده می‌شود. در روش سراسری، ابتدا میزان شدت روشنایی هر خال محاسبه، سپس به صورت تقریبی ۳٪ آن مقدار به عنوان شدت روشنایی پس‌زمینه در نظر گرفته می‌شود. در روش ریخت‌شناسی، با استفاده از یک فیلتر غیر خطی، مقدار پس‌زمینه تخمین زده می‌شود. پس از تصحیح پس‌زمینه، میزان روشنایی هر خال محاسبه می‌گردد. با اتمام همه مراحل ذکر شده، اطلاعات استنتاج شده از تصویر برای کاربر نمایش داده می‌شود. این اطلاعات شامل موقعیت و شناسه هر بلوک و خال، میزان شدت روشنایی، محاسبه میانگین و میانه، لگاریتم شدت روشنایی هر خال می‌باشد.

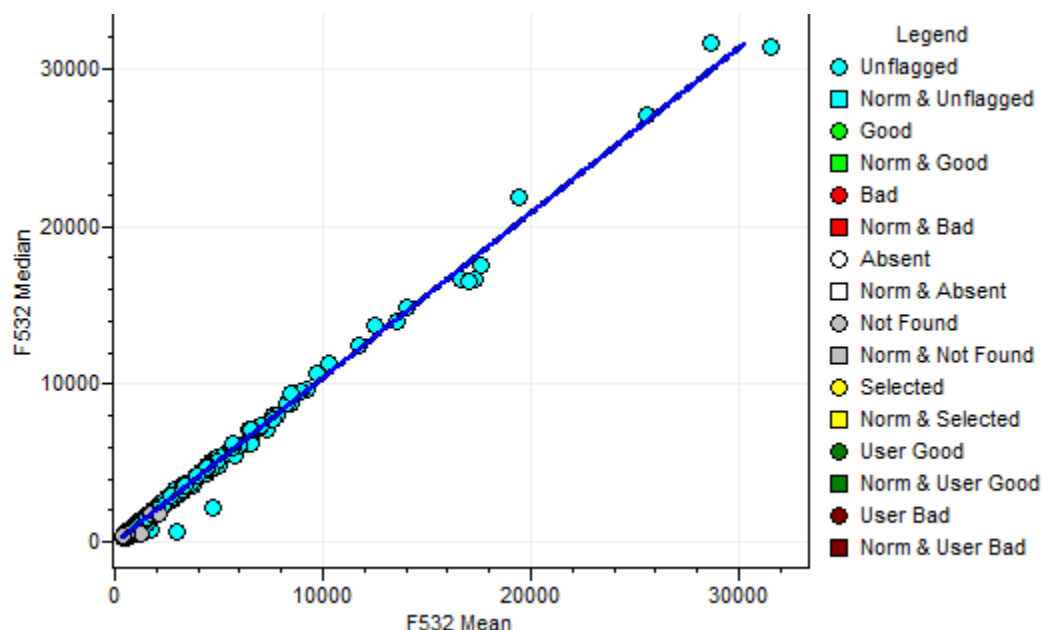
¹ Local Background Correction

² Constant Background Correction



شکل ۲-۴- تعیین مناطق محلی خال برای تخمین پس‌زمینه در نرم‌افزار GenePix.

در اکثر اوقات به دلیل خرابی تصویر یا اشکالاتی که در حین آزمایش به وجود می‌آید، داده‌ها به صورت نامنظم و پراکنده تخمین زده می‌شوند. نرمال‌سازی داده‌ها، در بسیاری از موارد باعث نظم بهتر بین آن‌ها و باعث کاهش پراکندگی می‌شود. بدین منظور، روش نرمال‌سازی سراسری براساس میانگین داده‌ها استفاده می‌شود. در نهایت تمامی اطلاعات استنتاجی به همراه داده‌های نرمال در فرم متنی ذخیره می‌شود. یکی از مهم‌ترین قابلیت نرم‌افزار GenePix رسم نمودار پراکندگی داده‌های محاسبه شده از تصویر و هیستوگرام تمامی آن‌ها می‌باشد. در این قسمت مقدار متغیر در نمودار افقی و عمودی به صورت دلخواه تعریف و تمامی اطلاعات استنتاجی نسبت به هم همانند شکل (۲-۵) نمایش داده می‌شود. در این نمودار هر دایره نماینده یک خال در تصویر می‌باشد. در این نمودار مقدار میانگین و میانه مقدار بیان هر ژن توسط نرم‌افزار رسم شده است. داده‌های استخراج شده مربوط به یک بلوک از تصویر ریزآرایه سرطان سینه می‌باشد، که توسط نرم‌افزار GenePix از تصویر استخراج شده است.

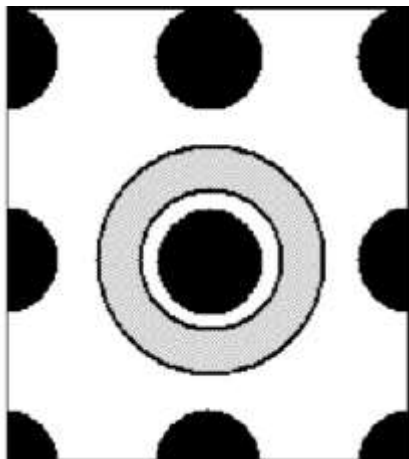


شکل ۲-۵- نمودار پراکندگی داده‌های ریزآرایه در نرم‌افزار GenePix.

۲-۲-۲ نرم‌افزار ImaGene

نرم‌افزار ImaGene محصول شرکت BioDiscovery از سال ۱۹۹۷ در کالیفرنیا به عنوان یک شرکت پیشگام در زمینه‌ی انفورماتیک ریزآرایه تاسیس و معرفی شد. این نرم‌افزار به طور مشابه همانند نرم‌افزار قبلی عمل می‌کند. در این بخش، به صورت خلاصه عملکرد نرم‌افزار مورد بررسی قرار می‌گیرد. برای تعیین موقعیت بلوک‌ها و خال‌ها همانند نرم‌افزار GenePix برخی از پارامترها همانند فواصل مابین خال‌ها و بلوک‌ها به صورت دستی توسط کاربر وارد سیستم می‌شود. در حالت استاندارد، تصویر ابتدا قطعه‌بندی شده، سپس پس‌زمینه‌ی آن تصحیح می‌گردد. برخلاف نرم‌افزار قبلی الگوریتم‌های کمتری برای تعیین این دو مرحله بکار گرفته شده است. برای قطعه‌بندی خال‌ها دو روش دستی و خودکار تعریف می‌شود. در روش دستی، کاربر موظف به تعیین پارامتر محدوددهی پس‌زمینه در اطراف خال، دامنه تغییرات پس‌زمینه و پیش‌زمینه می‌باشد. این مقادیر اکثراً از طریق آزمایش و تجربه بدست می‌آیند. روش خودکار به صورت قطعه‌بندی براساس هیستوگرام عمل می‌نماید. این روش برای تصاویر بدون نویز بکار گرفته می‌شود. برای تصحیح پس‌زمینه نیز تنها دو روش محلی و سراسری در این نرم‌افزار

تعریف شده است [۲۴]. در شکل (۶-۲) منطقه‌ای که نرم‌افزار برای تخمین پس‌زمینه در نظر می‌گیرد، نمایش داده می‌شود.



شکل ۶-۲- تعیین مناطق محلی خال برای تخمین پس‌زمینه در نرم‌افزار ImaGene.

با محاسبه‌ی شدت روشنایی خال‌ها و کاستن مقدار پس‌زمینه، مقدار واقعی هر خال تخمین زده می‌شود. برای نرمال‌سازی داده‌های ریزآرایه از چند روش از جمله^۱ Lowess و نرمال‌سازی بر اساس میانگین داده‌ها استفاده می‌شود. در نرمال‌سازی به روش Lowess داده‌ها براساس منحنی برازش رگرسیون غیرخطی مبتنی بر داده نرمال می‌شوند. سایر قسمت‌های این سیستم مشابه به GenePix عمل می‌کند. در نهایت تمامی اطلاعات استخراج شده از تصویر در فرم متن با فرمت txt ذخیره می‌شود.

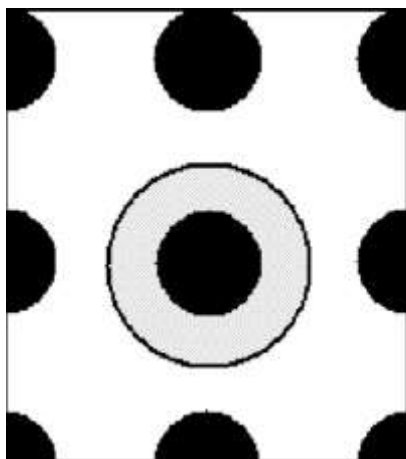
۳-۲-۲ نرم‌افزار ScanAlyze

نرم‌افزار ScanAlyze محصول آزمایشگاه مایکل ایزن^۲ در موسسه پزشکی هوارد در دانشگاه کالیفرنیا می‌باشد. اولین نسخه‌ی آن در سال ۱۹۹۹ به بازار عرضه گردید. فعالیت اصلی آن‌ها بر روی داده‌های ژنتیک و تحلیل آن‌ها متمرکز می‌باشد. این نرم‌افزار دقیقاً مشابه دو نرم‌افزار قبلی عمل می‌کند، با این تفاوت که برای قطعه‌بندی خال‌ها از روش دایره ثابت و دایره وفقی استفاده می‌شود. در روش دایره وفقی، نرم‌افزار قابلیت انتخاب دستی شعاع دایره برای تعیین ماسک هر خال را به صورت مجزا دارد.

^۱ Locally Weighted Linear Regression

^۲ Michael Eisen

البته این عمل به علت تعداد بسیار زیاد خال‌ها در هر تصویر، می‌تواند بسیار وقت‌گیر باشد [۴۴]. در گام تصحیح پس‌زمینه به روش محلی، منطقه‌ای که نرم‌افزار برای تخمین پس‌زمینه در نظر می‌گیرد، در شکل (۷-۲) نمایش داده شده است.

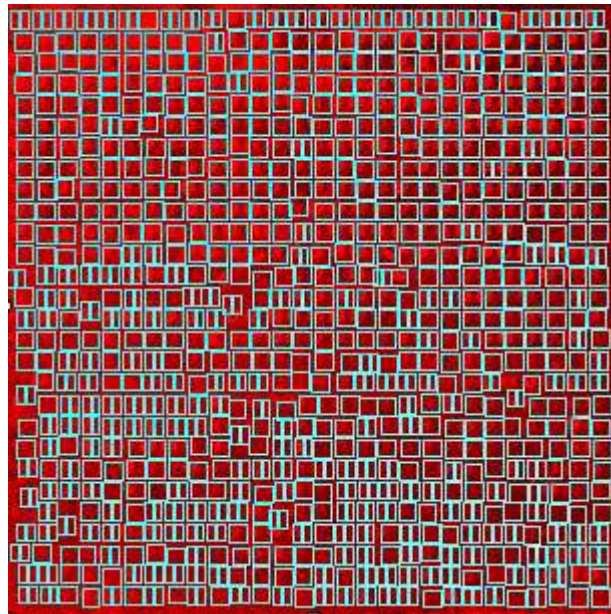


شکل ۷-۲- تعیین مناطق محلی خال برای تخمین پس‌زمینه در نرم‌افزار ScanAlyze.

تقریباً سه نرم‌افزار ImaGene, GenePix, و ScanAlyze مشابه یکدیگر عمل می‌کنند. از مهم‌ترین معایب مشترک آن‌ها عبارت است از:

- آدرس‌دهی دستی بلوک‌ها و خال‌ها در هر تصویر؛
 - قطعه‌بندی و تصحیح پس‌زمینه در دو مرحله جداگانه؛
 - تعیین دستی پارامترهای مربوط به تصحیح پس‌زمینه؛
 - پشتیبانی قالب‌های محدودی از تصویر ورودی؛
 - عدم مقایسه و تحلیل دو داده‌ی ریزآرایه به صورت خودکار؛
 - اجرای عملیات بلوک‌بندی، شبکه‌بندی و قطعه‌بندی به صورت مجزا در دو کانال رنگی؛
- نرم‌افزارها با استفاده از الگوریتم‌های متفاوت در این تلاشند تا فرد خبره با توجه به نیازهای خود، تصاویر ریزآرایه را کمی‌سازی نموده و داده‌های مناسبی را از آن‌ها استخراج نماید. تمامی نرم‌افزارهای ذکر شده به صورت نیمه‌خودکار عمل می‌کنند. در شرایطی که تصاویر به صورت ایده‌آل تهیه گردند و عاری از نویز باشند، نرم‌افزارها از کارایی خوبی برخوردار خواهند بود. اما در شرایطی که تصاویر دارای خرابی

زیادی باشد، ممکن است برخی از خال‌ها به درستی شناسایی نشوند. این خطا منجر به اعلام اطلاعات غلط به فرد متخصص بیولوژیکی می‌شود. در شکل (۲-۸) نمونه‌ای از این خطا مشاهده می‌شود. در این شکل به دلیل وجود نویز، عملیات شبکه‌بندی با خطای بالایی انجام شده است. این خطا منجر به نتایج اشتباه در ادامه کار می‌شود.



شکل ۲-۸- شبکه‌بندی تصویر ریزآرایه با استفاده از نرم‌افزار GenePix.

این نرم‌افزارها اغلب برای تصاویر ریزآرایه یک کانال تعریف می‌شوند. زیرا در تصاویر ریزآرایه دو کانال یکی از معایب بزرگ این نرم‌افزارها همان طور که ذکر شد، اجرای عملیات بلوک‌بندی، شبکه‌بندی و قطعه‌بندی به صورت مجزا در دو کانال رنگی می‌باشد، در حالی که موقعیت تمام بلوک‌ها و خال‌ها در هر دو کانال با هم برابر می‌باشد. بنابراین بهتر است برای افزایش دقت این مراحل دو کانال رنگی را با هم ادغام، سپس الگوریتم مربوطه اجرا گردد. با توجه به نواقص نرم‌افزارهای موجود، در این پژوهش سعی شده یک سیستم تمام خودکار برای تحلیل تصاویر ریزآرایه به همراه دقت بالا ارائه گردد. در فواصل بعدی نرم‌افزار ارائه شده به طور کامل شرح داده خواهد شد.

۳-۲ مروری بر الگوریتم‌های تحلیل تصاویر ریز آرایه

همان‌طور که در فصل اول بیان شد، تصاویر ریز آرایه با استفاده از الگوریتم‌های پردازش تصویر تبدیل به داده‌های خام و سپس به کمک روش‌های داده کاوی و یادگیری ماشین دسته‌بندی می‌شوند. با استفاده از این تکنیک‌ها سه گام اساسی بر روی تصاویر ریز آرایه، شامل: شبکه‌بندی، قطعه‌بندی و دسته‌بندی داده‌های ریز آرایه صورت می‌گیرد. پیشینه و الگوریتم‌های مربوط به این گام‌ها در ادامه به صورت مختصری شرح داده می‌شود.

۱-۳-۲ شبکه‌بندی

منظور از شبکه‌بندی، تعیین موقعیت هر خال در تصویر یا به عبارت دیگر تقسیم‌بندی خال‌ها به سلول‌های مجزا می‌باشد. صحت شبکه‌بندی، نتایج مابقی گام‌ها را به شدت تحت تاثیر قرار می‌دهد، به همین دلیل شبکه‌بندی در بسیاری از مقالات به عنوان مهم‌ترین گام یاد می‌شود [۴۵].

Dimitris Bariamis و همکارانش در سال ۲۰۰۹ با کمک الگوریتم ماشین بردار پشتیبان موقعیت خال‌ها را شناسایی کردند. در این روش ابتدا موقعیت خال‌ها بر روی تصاویر با استفاده از عملگرهای تحلیل تصاویر مانند لبه‌یابی شناسایی می‌شود. سپس خال‌ها به عنوان داده‌های ورودی در ماشین بردار پشتیبان دو کلاس مورد استفاده قرار می‌گیرد. برای استفاده از ماشین بردار پشتیبان ابتدا فواصل مابین ستون‌ها و سطرهای متوالی با استفاده از انحراف معیار به صورت تقریبی تخمین زده می‌شود. برای اجرای ماشین بردار پشتیبان بر روی تصویر، یک سطر همانند k در نظر گرفته می‌شود، خال‌هایی که قبل از سطر k قرار دارند در کلاس ۱ و مابقی در کلاس ۱- جای می‌گیرند. در نهایت، خط جداساز بین دو کلاس به وسیله ماشین بردار پشتیبان ترسیم می‌شود. به همین ترتیب ماشین بردار پشتیبان به تمامی سطرها و ستون‌های تخمینی وارد و خطوط جداسازی هر یک رسم می‌شود. در ۹۶/۴٪ موارد، موقعیت نقاط به درستی تشخیص داده می‌شود. نقطه ضعف مهم این روش شناسایی چندین خال در همه‌ی سطرها / ستون‌ها است. اگر در یک سطر / ستون به دلیل بیان کم زن‌ها، هیچ خالی شناسایی نشود خط جداساز نیز اشتباه تخمین زده می‌شود [۱۴].

Fatma El-Zahraa Labib به کمک همکارانش در سال ۲۰۱۲ یک روش تمام خودکار برای شبکه‌بندی خال‌ها معرفی کردند. در این روش با اعمال تکنیک‌های آستانه‌گیری در هیستوگرام بدست آمده از تصویر، فواصل مابین خال‌ها تخمین زده می‌شود. بدین منظور ابتدا تصویر رنگی به تصویر خاکستری تبدیل می‌شود، سپس با محاسبه‌ی مجموع پیکسل‌های سطرها / ستون‌ها هیستوگرام مربوطه به آن ترسیم می‌شود. با اعمال ضریب خود همبستگی بر روی هیستوگرام بدست آمده، فواصل مابین خال‌ها به وضوح از هم تشخیص داده می‌شوند. با تعیین نقاط کمینه‌ی محلی بر روی هیستوگرام، موقعیت خال‌ها شناسایی می‌شود. این روش در تصاویر همراه با نویز بالا به خوبی عمل نمی‌کند. در تصاویر ایده‌آل دقت روش ذکر شده به صورت تقریبی بر روی هر تصویر ۹۸/۶٪ می‌باشد [۱۵].

Harikiran و همکارانش در سال ۲۰۱۴ فرض را بر این بنا داشته‌اند که تمامی خال‌ها در فواصل برابری از یکدیگر قرار گرفته‌اند. با این فرض روشی خودکار برای شبکه‌بندی خال‌ها ارائه دادند. الگوریتم آن‌ها به طور کلی و خلاصه شامل سه گام مهم بود. در گام اول تصویر ریزآرایه لبه‌یابی می‌شود. در گام دوم عملگرهای ریخت‌شناسی بر روی تصویر بدست آمده اعمال می‌گردد. در نهایت با محاسبه مجموع سطرها / ستون‌ها فاصله بین خال‌ها محاسبه می‌شود. بدین منظور بعد از محاسبه مجموع پیکسل‌ها، تعداد پیکسل‌هایی که برابر صفر و غیرصفراند تعیین می‌شود. پیکسل‌های صفر نماینده فضای خالی بین دو خال می‌باشد. با محاسبه فاصله بدست آمده هر دو خال به صورت تقریبی، میانگینی از تمامی فواصل بدست می‌آید. این میانگین به عنوان فاصله نهایی تمام خال‌ها در نظر گرفته می‌شود. یکی از مهم‌ترین نقطه ضعف‌های این پژوهش، فرض برابری فاصله خال‌ها در تصویر ریزآرایه می‌باشد. در اکثر تصاویر ریزآرایه به دلیل خرابی‌های ایجاد شده همانند وجود نویز و چرخش تصویر، در زمان آزمایش یا اسکن، تصاویر ایده‌آل نیستند. به همین دلیل ممکن است تصاویر با فواصل کمی متفاوت در کنار یکدیگر قرار گیرند [۱۶].

۲-۳-۲ قطعه‌بندی

قطعه‌بندی در تصاویر ریزآرایه به معنی دسته‌بندی تصاویر به دو دسته‌ی پیش‌زمینه (خال) و پس‌زمینه می‌باشد. در فصل اول مشهورترین روش‌های قطعه‌بندی به نام‌های قطعه‌بندی براساس دایره ثابت، قطعه‌بندی براساس شکل وقفی، قطعه‌بندی براساس هیستوگرام و قطعه‌بندی براساس دایره وقفی به صورت کامل شرح داده شد. همان‌طور که مشاهده شد، در بسیاری از نرم‌افزارها از جمله ScanAlyze، GenePix و ImaGene برای استخراج خال‌ها از این چهار روش الهام گرفته‌اند.

در برخی از پژوهش‌ها با الهام از روش‌های خوشه‌بندی، عمل قطعه‌بندی خال‌ها صورت می‌گیرد. ایده‌ی الگوریتم‌های خوشه‌بندی، از جمله الگوریتم k-means بر این مبنا است که اشیایی که متعلق به یک گروه هستند شباهت‌های بیشتری نیز به هم دارند. در نوع ساده‌ای از این روش ابتدا به تعداد خوشه‌های مورد نیاز نقاطی به صورت تصادفی انتخاب می‌شود. سپس داده‌ها با توجه به میزان نزدیکی (شباهت) به یکی از این خوشه‌ها نسبت داده می‌شوند و بدین ترتیب خوشه‌های جدیدی حاصل می‌شود. با تکرار همین روال می‌توان در هر تکرار با میانگین‌گیری از داده‌ها مراکز جدیدی برای آنها محاسبه کرد و مجدداً داده‌ها را به خوشه‌های جدید نسبت داد. این روند تا زمانی ادامه پیدا می‌کند که دیگر تغییری در دسته‌بندی داده‌ها حاصل نشود. برای تصاویر ریزآرایه، پس‌زمینه و پیش‌زمینه به عنوان خوشه‌ی مجزا از هم در نظر گرفته می‌شود. این روش با در نظر گرفتن دوخوشه، در سال ۲۰۰۲ توسط Bozinov و Rahmenfuhrer برای قطعه‌بندی خال‌ها در تصاویر ریزآرایه استفاده شد [۴۶]. روش مذکور برای تصاویری که دارای نویز بالایی هستند و خال‌ها مقدار بیان ضعیفی دارند به خوبی عمل نمی‌کند. نویزها گاهی در این روش به عنوان خال شناسایی می‌شوند. بعدها در سال‌های ۲۰۰۹، ۲۰۱۰ و ۲۰۱۵، روش‌هایی برای بهبود خوشه‌بندی K-means به منظور قطعه‌بندی بهتر هر خال از تصویر ریزآرایه ارائه گردید [۴۷-۴۹]. در یک پژوهش انجام شده توسط Wang و همکارانش در سال ۲۰۱۵ به جای دو خوشه، از سه خوشه برای قطعه‌بندی استفاده شده است. با توجه به این حقیقت که مقادیر پیکسل‌های خاکستری پس‌زمینه از پیش‌زمینه کمتر است، خوشه‌ای که کمترین مقدار را دارد به عنوان پس‌زمینه و دو خوشه

دیگر به عنوان پیش‌زمینه تعریف می‌شوند [۴۹]. با در نظر گرفتن زمان اجرای الگوریتم، بهبود یافته الگوریتم K-means زمان بیشتری را برای اجرای الگوریتم خود نسبت به حالت استاندارد مصرف می‌کند. اما نسبت به روش قبلی در برابر نویز و بیان پایین ژن‌ها استحکام بیشتری دارد. در برخی از پژوهش‌ها با بکارگیری خوشه‌بندی ماشین بردار پشتیبان دو یا سه کلاس، خال‌ها قطعه‌بندی می‌شوند. برای مثال در سال ۲۰۱۳ روش دسته‌بندی سه کلاسه توسط Shao و همکارانش ارائه گردید که هر کلاس شامل، پس‌زمینه، پیش‌زمینه و نویز در تصویر ریزآرایه می‌باشد. مرز هر دسته با خط جداکننده مشخص می‌شود. مزیت این روش نسبت به روش‌های قبلی در نظر گرفتن نویز می‌باشد. اما ضعف این روش شناسایی نمونه‌ها به صورت صحیح می‌باشد. در صورتیکه نمونه‌های نویز، پس‌زمینه و پیش‌زمینه به درستی تخمین زده نشود مرز جداکننده مابین آن‌ها نیز به درستی بدست نمی‌آید [۵۰].

۳-۳-۲ دسته‌بندی داده‌های ریزآرایه

دسته‌بندی دارای تکنیک‌های متنوعی است که در تحقیقات مختلف از آن‌ها استفاده می‌شود. از تکنیک‌های رایج در طبقه‌بندی داده‌های ریزآرایه می‌توان به مواردی از جمله، شبکه عصبی، ماشین بردار پشتیبانی، طبقه‌بندی بیزین، سیستم‌های خبره، الگوریتم‌های ژنتیک و درخت تصمیم‌گیری اشاره کرد. با توجه به نگرش متفاوت پژوهشگران، روش‌های متفاوتی برای دسته‌بندی داده‌های ریزآرایه ارائه شده است. برخی از محققان تنها با بکارگیری یک تکنیک یادگیری ماشین داده‌های ریزآرایه‌ای را تفکیک می‌کنند. اما برخی از محققان با این تفکر که اغلب کلاس‌بندها وابسته به کاربرد خاصی هستند، به منظور بهبود نتایج، از ترکیب دسته‌بندی‌کننده‌های متفاوت استفاده می‌کنند. در بعضی موارد، فرد خبره نیازمند به دسترسی مراحل کار و توصیف قوانین قابل فهم می‌باشد، که در این حالت محقق خود را ملزم دانسته سیستم‌خبره‌ای طراحی کند تا فرد خبره بتواند از طریق تبادل اطلاعات، داده‌های ریزآرایه‌ای خود را دسته‌بندی کند. بنابراین روش‌های متفاوت دسته‌بندی داده‌های ریزآرایه به صورت کلی به سه روش زیر می‌توان تقسیم‌بندی کرد.

الف) دسته‌بندی‌کننده‌های منفرد: از نمونه‌های بکارگیری تکنیک یادگیری ماشین به صورت منفرد می‌توان موارد زیر اشاره کرد:

- 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. آن‌ها توانستند با استفاده از روش شبکه عصبی خودسازمان‌ده^۱ داده‌های سرطان خون را تحلیل و نمونه سالم و توموری را از هم تفکیک کنند [۳۵].

- Keerthika و همکارش در سال ۲۰۱۳ با استفاده از بهینه‌سازی حداقل ترتیبی^۲ توانستند داده‌های ریزآرایه را دسته‌بندی کنند [۵۱].

- Halder و همکارانش در سال ۲۰۱۵ با استفاده از الگوریتم k جستجوی نزدیک‌ترین همسایه^۳ به صورت فازی داده‌های ریزآرایه خود را دسته‌بندی کرده‌اند [۳۷].

ب) دسته‌بندی‌کننده‌های ترکیبی: در سال‌های اخیر، مطالعات بر روی دسته‌بندی ترکیبی رشد چشمگیری داشته است. یکی از مهم‌ترین معایب استفاده از تکنیک‌های ترکیبی افزایش پیچیدگی و بالا بردن زمان اجرای سیستم می‌باشد. در این حالت نیز می‌توان به موارد زیر اشاره کرد:

- Zhenyu Wang و همکارنش در سال ۲۰۰۷ با ادغام دو روش بردار ماشین پشتیبان و شبکه عصبی فازی داده‌های خود را دسته‌بندی کردند [۳۶]. در این مقاله از سه پایگاه داده سرطان خون، سرطان روده بزرگ و سرطان غده لنفاوی استفاده کرده و توانستند تمامی نمونه‌ها را به ترتیب با دقت ۱۰۰٪، ۹۵/۸۵٪، ۹۵/۶۵٪ دسته‌بندی کنند.

^۱ SOM(self-organizing map)

^۲ SMO (Sequential Minimal Optimization)

^۳ K Nearest Neighbor

- Sajid nagi و همکارش در سال ۲۰۱۳ به منظور بهبود نتایج، از ترکیب روش‌های دسته‌بندی کیسه کردن^۱ و روش تقویتی^۲ و تعمیم پشته‌ای^۳ کمک گرفته و نه پایگاه داده‌ی ریزآرایه خود را تقریباً با دقت بالایی دسته‌بندی کردند [۵۲].

- Jaison و همکارانش در سال ۲۰۱۵ با ترکیب روش‌های دسته‌بندی بیزین ساده، ماشین بردار پشتیبان و k جستجوی نزدیک‌ترین همسایه، سیستمی برای طبقه‌بندی چهار نوع پایگاه داده‌ی ریزآرایه ارائه داده است [۵۳].

- در ایران نیز در سال ۱۳۹۲ رقیه ابوالقاسمی به کمک همکارانش ژن‌هایی که حاوی اطلاعات مفید هستند را با استفاده از الگوریتم ژنتیک و شبکه عصبی خود سازمان‌ده طبقه‌بندی می‌کنند. روش پیشنهادی این مقاله موفق شد که طبقه‌بندی داده‌های سرطان سینه را به همراه دقت ۹۴٪ انجام دهد [۵۴].

- در سال ۱۳۹۳ محمد مروت با ترکیب سه روش آدابوست^۴، بگینگ^۵ و دگینگ^۶ سیستمی برای دسته‌بندی داده‌های ریزآرایه ارائه می‌دهد. نتایج حاصل از روش‌های پیشنهادی بر روی یازده مجموعه داده ریزآرایه، بیانگر کارایی روش‌های پیشنهادی در راستای تشخیص دقیق بیماری سرطان است [۵۵].

ج) دسته‌بندی کننده به صورت سیستم خبره: سیستم‌های خبره دارای الگوی منطقی هستند که یک متخصص براساس آنان تصمیم‌گیری می‌کند. نمونه‌ای از این سیستم‌های خبره در زمینه ریزآرایه عبارتند از:

- Yoon و همکارانش در سال ۲۰۱۰ برای اولین بار سیستم خبره‌ای طراحی کردند که با استفاده از چندین قانون ساده، داده‌های ریزآرایه را طبقه‌بندی می‌کنند. در این روش ابتدا تمامی قوانین

¹ Bagging Method

² Boosting Methods

³ Stacked generalization

⁴ AdaBoost

⁵ Bagging

⁶ Daging

با استفاده از داده‌های آموزشی تهیه و امتیازبندی می‌شود. سپس قوانین مشابه ادغام و قوانین اضافه حذف می‌گردد. در این پژوهش نمونه‌های توموری از نمونه‌های سالم در سه پایگاه داده‌ی متفاوت با دقت بالغ بر ۸۵٪ تفکیک می‌شود [۵۶].

- Ganesh Kumar و همکارانش در سال ۲۰۱۲ با بکارگیری الگوریتم ژنتیک و بهینه‌سازی ذرات به صورت همزمان برای تعیین توابع عضویت هر ژن و تولید قوانین، موفق به طراحی سیستم خبره‌ای به صورت فازی شده‌اند. پایگاه داده‌ی این روش شامل سرطان روده، سرطان خون و سرطان بافت لنفاوی به ترتیب با هفت، هشت و شش قانون به همراه دقت ۹۶٫۸٪، ۹۴٫۴٪ و ۹۱٫۱٪ دسته‌بندی می‌شوند [۳۰].

همان‌طور که شرح داده شد، روش‌های فراوانی برای دسته‌بندی داده‌های ریزآرایه پیشنهاد شده است. هر کدام از این روش‌ها با توجه به پایگاه داده‌ی مورد استفاده خودشان، دقت بالایی برای تفکیک آن‌ها ارائه می‌دهند. الگوریتم‌های دسته‌بندی با توجه به نیاز فرد خبره و کارایی آن‌ها انتخاب می‌شوند. تقریباً در تمامی روش‌های ذکر شده، برای رسیدن به نتیجه نهایی الگوریتم‌ها همانند جعبه سیاه عمل می‌کنند و مسیر طی شده برای دسته‌بندی داده‌های ریزآرایه در اختیار فرد خبره قرار نمی‌گیرد. با توجه به نکات ذکر شده، قوانین تولید شده و بکار گرفته شده در الگوریتم‌هایی مانند درخت تصمیم‌گیری و جدول تصمیم‌گیری، قابل استخراج و فهم می‌باشند، بنابراین این الگوریتم‌ها گزینه مناسب‌تری برای تحلیل و دسته‌بندی داده‌های ریزآرایه می‌باشند.

۴-۲ نتیجه‌گیری

در این فصل، نرم‌افزارهای موجود و تحقیقات انجام گرفته در زمینه‌ی تحلیل تصاویر و داده‌های ریزآرایه شرح داده شد. پژوهشگران در تمامی نرم‌افزارهای عرضه شده، سعی بر این داشته‌اند تا به بهترین صورت تصاویر ریزآرایه به داده‌ها کمی و خام تبدیل شوند. در این فصل الگوریتم‌های متفاوتی در زمینه شبکه‌بندی، قطعه‌بندی تصاویر و دسته‌بندی داده‌ی ریزآرایه معرفی گردید. در اکثر تحقیقات

محققان در تلاشند دقت خوبی برای تحلیل تصاویر ریزآرایه ارائه دهند. در فصل بعدی، سیستمی برای تحلیل و دسته‌بندی تصاویر و داده‌های ریزآرایه خواهد شد.

فصل سوم

سیتم تحلیل تصاویر زیر آرایه

۱-۳ مقدمه

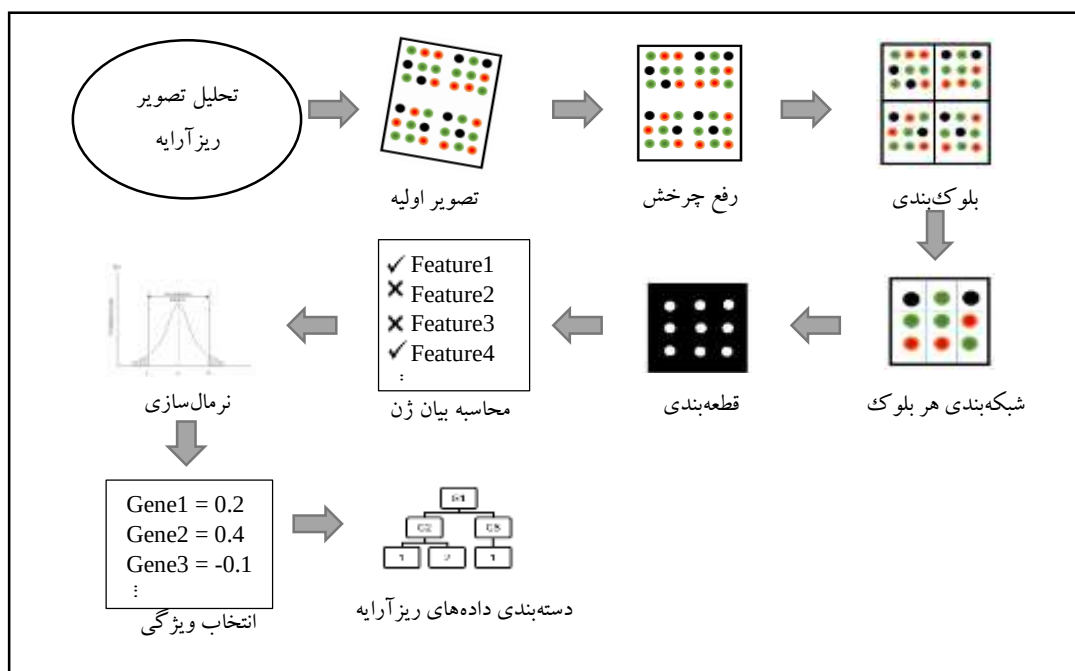
با تمایل قابل توجهی که امروزه برای استفاده از تکنولوژی ریزآرایه بوجود آمده است در هر آزمایش تحقیقاتی حجم زیادی از این تصاویر تولید می‌شود. یکی از مهم‌ترین مراحل آزمایش ریزآرایه، پردازش تصویر و تحلیل داده‌های ریزآرایه حاصل از آن است. در اغلب سیستم‌هایی که به منظور تحلیل تصاویر ریزآرایه وجود دارند (برای مثال ImaGene, ScanAlyze و Genepix) تنها قادر به کمی‌سازی تصاویر ریزآرایه می‌باشند و عموماً به صورت نیمه‌خودکار عمل می‌کنند. همچنین در شرایطی که تصاویر دارای خرابی همانند نویز باشند، خال‌ها به درستی شناسایی نمی‌شوند. علاوه بر این در این نرم‌افزارها تحلیل داده‌های ریزآرایه به منظور کاهش ابعاد و دسته‌بندی آن‌ها امکان پذیر نیست. بدین منظور سیستمی جهت پردازش و تحلیل خودکار برای این تصاویر پیشنهاد شده است. در ادامه سیستم به صورت مفصل شرح داده خواهد شد.

۲-۳ سیستم پیشنهادی

در این قسمت طی چندین مرحله، سیستمی برای استخراج داده‌ها از تصاویر ریزآرایه، تحلیل و دسته‌بندی داده‌های خام مربوط به سه نوع سرطان با استفاده از تکنیک‌های پردازش تصویر، داده‌کاوی و یادگیری ماشین ارائه داده می‌شود. با توجه به مطالعات صورت گرفته در اکثر سیستم‌ها، الگوریتم تصحیح زاویه چرخش تصویر به صورت مجزا به هر بلوک اعمال شده و بعد از تعیین مکان خال‌ها مرحله قطعه‌بندی، اکثراً در دو مرحله مجزا (قطعه‌بندی و تصحیح پس‌زمینه) صورت می‌گیرد، در حالی که در سیستم پیشنهادی سعی بر این شده که تنها با اجرای الگوریتم تخمین پس‌زمینه مقدار بیان هر ژن محاسبه گردد. همچنین نرم‌افزارهای موجود اغلب به صورت نیمه‌خودکار عمل کرده و تنها تا گام کمی‌سازی تصویر پیش می‌روند و به تحلیل داده‌های ریزآرایه به منظور دسته‌بندی آن‌ها پرداخته نمی‌شود.

در سیستم پیشنهادی سه گام پردازش تصویر، داده‌کاوی و یادگیری ماشین به صورت کامل و خودکار اجرا می‌شود. مراحل طی شده در شکل (۱-۳) نمایش داده می‌شود. از بین این مراحل چهار مرحله‌ی

اولیه مربوط به پردازش تصویر و مابقی مربوط به تحلیل داده‌ها می‌باشد. پایگاه داده مورد استفاده در این پژوهش مربوط به سه نوع سرطان و به صورت مجزا هر کدام شامل دو زیر مجموعه متفاوت می‌باشند. ورودی سیستم پیشنهادی تصویر ریزآرایه و خروجی آن نمایش نوع تومور در آزمون بیمار می‌باشد. اطلاعات مهمی که از تصویر ریزآرایه در سیستم برای فرد خبره ذخیره می‌گردد، شامل درجه‌ی هیپرید شدن هر کدام از خال‌ها در دو کانال رنگی تصویر و میزان بیان هر ژن می‌باشد.



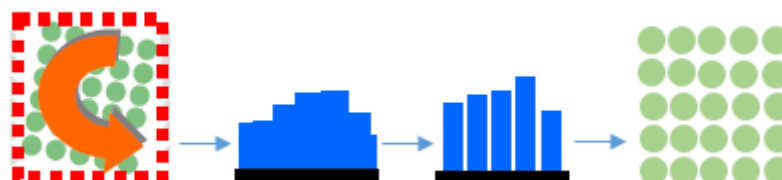
شکل ۳-۱- شمای کلی از سیستم ارائه شده برای تحلیل تصاویر ریزآرایه.

در گام اول تصاویری که به صورت زاویه دار اسکن گرفته شده، شناسایی و میزان چرخش آن‌ها محاسبه و تصحیح می‌گردد. در گام دوم و سوم به ترتیب به صورت دقیق تعداد بلوک‌ها مشخص و در هر بلوک مختصات خال‌ها به صورت جداگانه تخمین زده می‌شود. تعیین محل خال‌ها به عنوان شبکه‌بندی تصویر نامیده می‌شود. با توجه به اینکه، خطا در گام‌های اولیه می‌تواند منجر به نتایج اشتباه شود، شبکه‌بندی دقیق خال‌ها، مهم‌ترین گام برای ادامه کار می‌باشد. با توجه به مطالعات صورت گرفته، مقدار پس‌زمینه در هر یک از تصاویر، شامل برخی از رشته‌های دی‌ان‌ای جفت نشده یا مربوط به مواد شیمیایی است که در زمان آزمایش بر روی تراشه باقی مانده‌اند [۲۶]. بنابراین در گام چهارم پس‌زمینه خال‌ها تشخیص

و از تصویر اصلی کم می‌شود. بدین ترتیب خال‌ها قطعه‌بندی و مقدار پس‌زمینه حذف می‌گردد. سپس، با میانگین‌گیری پیکسل‌های هر خال در دو کانال تصویر میزان بیان ژن‌ها محاسبه می‌گردد. بعد از استخراج داده‌ها و نرمال‌سازی آن‌ها، در گام هشتم، به علت وجود تعداد ویژگی بالا (هر تصویر به صورت تقریبی شامل ۴۸۰۰۰ ژن می‌باشد) از روش‌های داده‌کاوی استفاده و ژن‌های مفید برای دسته‌بندی را انتخاب می‌کنیم. در آخرین گام برای مشخص کردن نوع تومور و دسته‌بندی آن‌ها از روش‌های یادگیری ماشین استفاده می‌کنیم. در ادامه بصورت دقیق تمامی گام‌ها توضیح داده می‌شود.

تشخیص زاویه چرخش و تصحیح آن

گاهی در هنگام قراردادن تراشه داخل دستگاه اسکن ریزآرایه، مراقبت کافی به عمل نیامده و انحراف از حالت اصلی به وجود می‌آید. در چنین مواردی، فرایند تشخیص کجی تصویر معمولاً به عنوان یک عمل پیش‌پردازش مطرح می‌شود. روش‌های متعددی در این خصوص ارائه شده است [۵۷، ۵۸]. این الگوریتم‌ها در تصاویری که شامل نویز فراوانی است به درستی عمل نمی‌کند. بعلاوه الگوریتم آن‌ها در تصاویر کامل ریزآرایه اجرا پذیر نیستند و باید هر بلوک‌ها به صورت جداگانه در دسترس باشد (شکل ۲-۳). از آنجایی که در این پایان‌نامه ابتدا چرخش تصویر ریزآرایه شناسایی و تصحیح می‌گردد، گام بعدی برای تعیین موقعیت بلوک‌ها اختصاص می‌گیرد. در این قسمت با استفاده از روش تبدیلات رادن^۱ میزان زاویه چرخش کل تصویر حاصله از تراشه را تشخیص و اصلاح لازم بر روی آنان صورت گیرد [۵۹-۶۱].



¹ Transform Radon

شکل ۳-۲- نحوه‌ی تشخیص زاویه چرخش در یک بلوک از تصویر ریزآرایه با استفاده از روش [۵۶].

۱- تبدیلات رادن

چرخش تصویر در هر دو جهت عقربه ساعت و بلعکس بر روی محور x امکان پذیر است. در مطالعه‌ی حاضر روش ساده‌تری برای تعیین میزان چرخش به کار گرفته می‌شود. اساس الگوریتم پیشنهادی بر مبنای تبدیل رادن است.

تبدیل رادن تعریف‌های متعددی دارد. از نظر ریاضی تبدیل انتگرالی است که مقدار آن برابر با انتگرال تابع بر روی یک خط است [۶۲]. معادله یک خط L در فضای دو بعدی با توجه به شکل (۳-۳) را می‌توان به صورت معادله (۱-۳) نوشت:

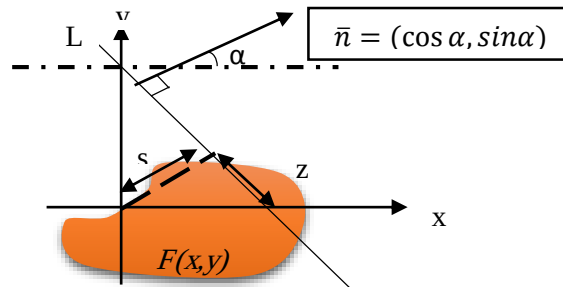
$$L(x(t), y(t)) = t(\sin \alpha, -\cos \alpha) + s(\cos \alpha, \sin \alpha) \quad (1-3)$$

که s فاصله خط از مبدا مختصات را نشان می‌دهد و $n(\alpha) = (\cos \alpha, \sin \alpha)$ بردار یکه عمود بر خط است. به همین ترتیب اگر نقطه تقاطع خط مذکور و خط عمود بر آن و گذرنده از مبدا را B بنامیم آن‌گاه t فاصله هر نقطه واقع بر خط را تا نقطه B نشان می‌دهد. همچنین $(\sin \alpha, -\cos \alpha)$ بردار راستای خط است. تبدیل رادن براساس شکل (۳-۳) و معادله (۱-۳) به صورت معادله (۲-۳) تعریف می‌شود.

$$\begin{aligned} R_{\alpha}(\alpha, s) &= \int_{-\infty}^{+\infty} f(L(x(t), y(t))) dt \\ &= \int_{-\infty}^{+\infty} f(t(\sin \alpha, -\cos \alpha) + s(\cos \alpha, \sin \alpha)) dt \end{aligned} \quad (2-3)$$

$$-\infty < s < \infty, 0 \leq \alpha < \pi$$

که R_α نماد تابع تبدیل رادن و $f(x(t), y(t))$ ، نماد تصویر به صورت پیوسته نسبت به محورهای افقی x و عمودی y است.



شکل ۳-۳ نحوه‌ی محاسبه تبدیل رادن از یک تصویر دو بعدی.

در واقع تبدیل رادن، جمع مقادیر نقاط یک تصویر در امتداد خطوطی موازی، عمود بر محوری است که با محور افقی زاویه α می‌سازد. تبدیل رادن از یک تصویر در واقع توانایی تبدیل تصاویر دو بعدی به خطوطی در دامنه پارامترهای خطی ممکن، به طوریکه هر خط در تصویر، یک ماکزیمم بدست خواهد آمد که در پارامترهای خطوط نظیر به نظیر موقعیت‌بندی می‌شود. در پژوهش فعلی برای محاسبه‌ی تبدیل رادن به ازای زاویه α ، از نرم افزار متلب استفاده شده است که تبدیل رادن برای تصویر 1852×5416 گسسته به صورت یک بردار 5727×1 در نظر گرفته می‌شود. فرم کلی تبدیل رادن به صورت معادله (۳-۳) است.

$$\rho = x \cos \alpha + y \sin \alpha \quad (3-3)$$

ρ کوچکترین فاصله از مبدا مختصات است. بنابراین تبدیل رادن برای یک مجموعه پارامترهای ρ و α انتگرال خط از تصویر است که خط نظیر به نظیر با مقدار ρ و α تعیین می‌شود. با توجه به اینکه زاویه چرخش تصاویر ریزآرایه بسیار اندک است، بنابراین برای محاسبه آن، تبدیل رادن برای تصویر از زاویه -1 تا $+1$ با گام‌های 0.1 محاسبه می‌شود. هر زاویه چرخش α یک تابع یک بعدی متفاوت است و به صورت یک بانک اطلاعاتی ذخیره می‌شود. با پیدا کردن حداکثر مقدار رادن، مقدار زاویه چرخش تشخیص و با اعمال عکس آن به تصویر، چرخش تصویر تصحیح می‌گردد. برای شبیه‌سازی الگوریتم از

یک تصویری که حرف M را نمایش می‌دهد به عنوان یک تصویر مصنوعی استفاده می‌شود. در شکل (۳-۴) تصویر نمونه مورد آزمایش به همراه چرخش ۱۰ درجه نشان داده می‌شود. در قسمت شکل (۳-۴) (ب) زاویه چرخش تصویر با استفاده از تبدیلات رادن شناسایی و تصحیح می‌شود.



(ب)



(الف)

شکل ۳-۴- تشخیص زاویه چرخش و رفع آن با استفاده از روش تبدیلات رادن. (الف) تصویر مصنوعی چرخیده شده با زاویه ۱۰ درجه، (ب) تشخیص زاویه چرخش و تصحیح آن.

بلوک‌بندی تصویر

کل تصویر ریزآرایه توسط یک ماتریسی از چندین بلوک با اندازه‌های مشابه تشکیل می‌شود. هر بلوک نیز متشکل از تعداد معینی خال در چندین سطر و ستون می‌باشد. هدف از این قسمت یافتن تعداد صحیح بلوک‌ها و شناسایی موقعیت آن‌ها در تصویر است. روش‌های زیادی برای بلوک‌بندی تصویر در سال‌های گذشته پیشنهاد شده‌اند که در تصاویر با نویز زیاد و بدون دخالت انسان به درستی عمل نمی‌کنند [۵۸، ۶۰].

در این بخش از روش آستانه‌گذاری^۱ هیستوگرام برای بلوک‌بندی استفاده می‌شود. آستانه‌گذاری تصویر یکی از مهم‌ترین تکنیک‌هایی است که به طور گسترده استفاده و در پردازش تصویر کاربردهای زیادی دارد. این تکنیک در عملیاتی از جمله قطعه‌بندی، دسته‌بندی و شناسایی اجسام در تصویر بکار گرفته می‌شود [۶۱]. برای بلوک‌بندی یا به عبارت دیگر یافتن عرض و ارتفاع هر بلوک، مجموع شدت روشنایی

¹ Thresholding

هر سطر / ستون از پیکسل‌ها در تصویر محاسبه و یک تابع گسسته یک بعدی حاصل می‌شود. بنابراین، با استفاده از تابع گسسته یک بعدی، هیستوگرامی بدست می‌آید که محور عمودی مجموع پیکسل‌ها سطر / ستون و محور افقی موقعیت آن‌ها را نمایش می‌دهد. روش تعیین حد آستانه برای مشخص کردن موقعیت بلوک‌ها متفاوت است. در بعضی مقالات از روش اتسو برای بلوک‌بندی استفاده می‌کنند [۶۱]. اما در این پژوهش از یک روش ساده‌تر یعنی یافتن کمترین مقدار محلی، حد آستانه هر بلوک مشخص می‌شود. حد آستانه برابر مجموعه‌ای از T است که به صورت معادله (۳-۴) تعریف می‌شود.

$$T = \{t_0, t_1, \dots, t_k, t_{k+1}\}, \quad (3-4)$$

$$0 = t_0 < t_1 < \dots < t_k < t_{k+1} = n, \quad t_k \in \{0\} \cup H$$

در عبارت بالا محور افقی هیستوگرام با نماد H تعریف شده و اندازه آن مابین $0 \leq H \leq n$ می‌باشد. هر کدام از مقادیر t بیانگر خطوط جداکننده بین بلوک‌ها هستند و مقدار n نماینده بزرگترین آستانه‌ی جداکننده دو بلوک می‌باشد. بدین ترتیب، یافتن موقعیت بلوک‌ها به این شرح می‌باشد:

الف) کیفیت تصویر با استفاده از برابری هیستوگرام افزایش داده می‌شود.

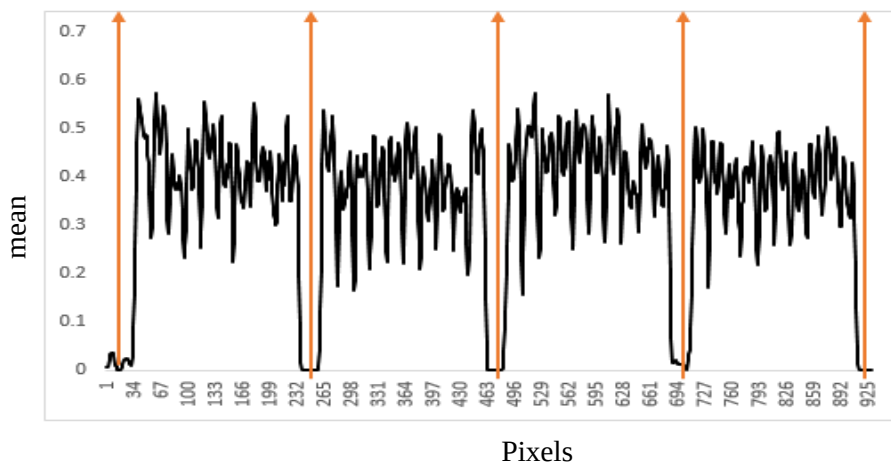
ب) فیلتر وینر را اعمال کرده تا نویزی که در پس‌زمینه ایجاد شده تا حدودی برطرف گردد.

پ) شدت روشنایی پیکسل‌های تصویر در سطر / ستون محاسبه می‌شود.

ج) هیستوگرام مربوطه رسم می‌گردد.

چ) کمترین مقدار مربوط به فضای خالی بین دو بلوک است، بنابراین با محاسبه حداقل مقدار محلی، فضای مابین بلوک‌ها محاسبه می‌شود.

نتیجه بدست آمده در نمودار (۳-۱) نمایش داده شده است. در این نمودار محور عمودی مقدار شدت روشنایی پیکسل‌ها را در هر ستون نمایش داده و علاوه بر آن خطوط عمودی نشان دهنده‌ی خط جداکننده بلوک‌ها می‌باشد.



نمودار ۳ - ۱- تعیین محدوده‌ی بلوک‌ها با استفاده از هیستوگرام تصویر.

برای بلوک‌بندی افقی تصویر کافی است تصویر را ترانهاده و تمامی عملیات تکرار شود. بدین ترتیب با استفاده از پارامترهای بدست آمده، بلوک‌ها به صورت مستطیلی از هم جدا می‌شوند. در نهایت، موقعیت و اندازه هر یک از مستطیل‌ها در یک ماتریس ذخیره و در صورت نیاز با استفاده از آن به هر بلوک فراخوان می‌شود. در شکل (۳-۵) یک تصویر ریزآرایه استاندارد و مصنوعی نمایش داده می‌شود. تمامی مراحل بلوک‌بندی بر روی تصویر اعمال شده و نتیجه‌ی حاصل از روش پیشنهادی و هیستوگرام مربوط به آن نشان داده شده است. خطوط رسم شده در هیستوگرام‌های شکل (۳-۵-ب، ج) مختصات مابین هر دو بلوک را مشخص می‌کند. با محاسبه‌ی مختصات افقی و عمودی در هر دو هیستوگرام، موقعیت و تعداد بلوک‌ها در تصویر ریزآرایه شناسایی شده است.

شبکه‌بندی هر بلوک

در یک تراشه ریزآرایه مولکول‌های تک رشته‌ای دی‌ان‌ای در مکان‌های ثابتی در ابعاد کوچک قرار می‌گیرند. به هر یک از این مکان‌ها یک خال گفته می‌شود. هر خال نماینده یک ژن در دی‌ان‌ای می‌باشد. تراشه‌های ریزآرایه متشکل از هزاران خال است. هر کدام از خال‌ها شامل تعداد زیادی مولکول‌های تک رشته‌ای دی‌ان‌ای یکسان برای شناسایی یک ژن خاص می‌باشد. نماینده‌ی هر ژن منطقه‌ای کوچکی را به صورت دایره‌ای شکل در تراشه اشغال کرده است. شناسایی و تعیین موقعیت هر خال در یک زیر

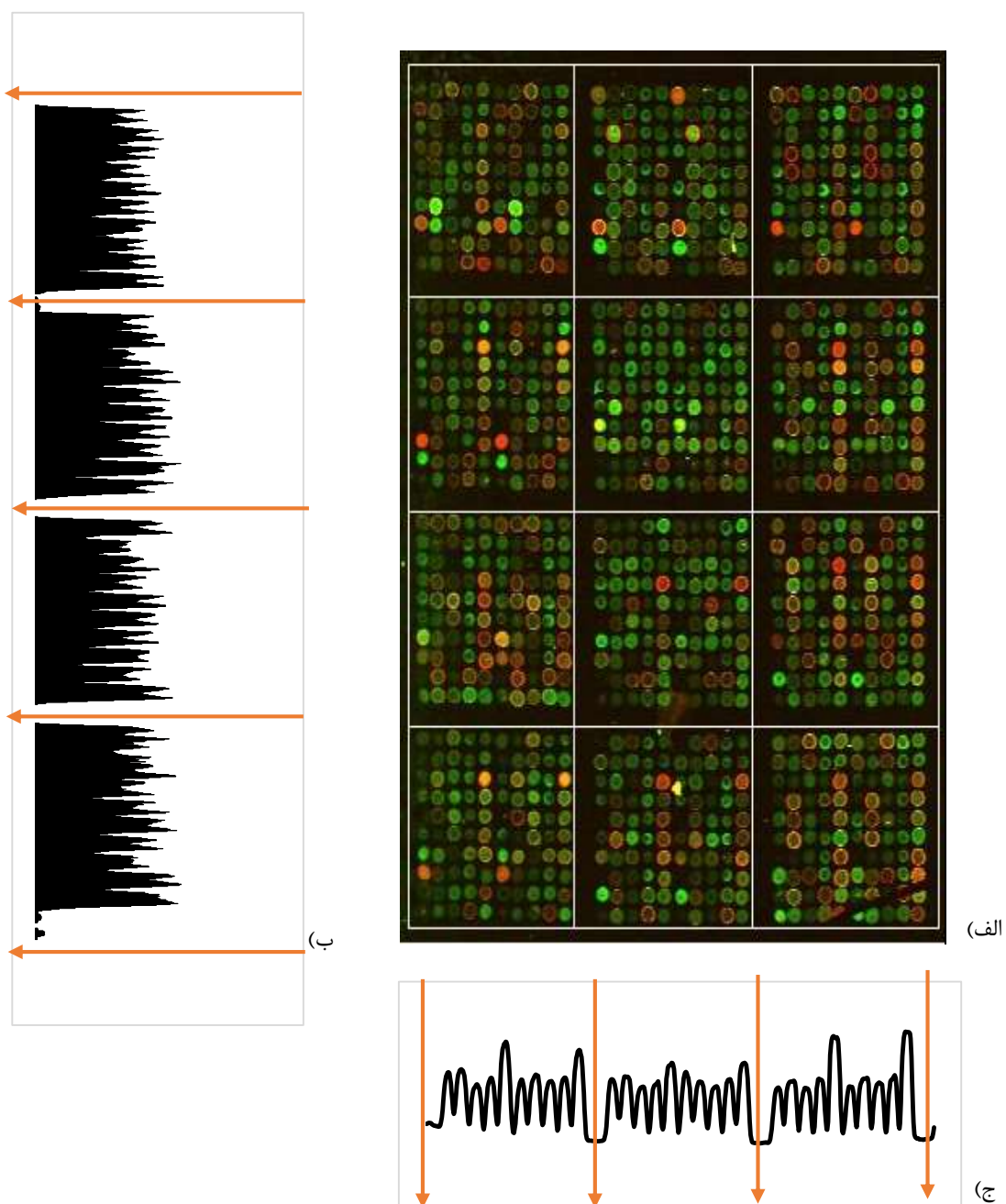
بلوک، هدف اصلی این قسمت می‌باشد. به عبارت دیگر هدف، تقسیم‌بندی خال‌ها به سلول‌های مجزا در

تصویر ریزآرایه است. یک تصویر ریزآرایه کامل و بی‌نقص دارای ویژگی‌های زیر می‌باشد:

- همه بلوک‌ها در تصویر باید به یک اندازه باشد و فاصله بلوک‌ها از هم به صورت یکسان و منظمی می‌باشد.
- تمامی خال‌ها در هریک از بلوک‌ها به صورت منظم و در راستای یک خط قرار می‌گیرند.
- شکل همه خال‌ها باید به صورت کاملاً گرد و دارای شعاع مشخصی باشد.
- تصاویر بدون نویز هستند.
- شدت روشنایی پس‌زمینه در همه قسمت‌ها دارای یک مقدار خاصی است.
- تصویر چرخیده نباشد.
- شدت روشنایی خال‌ها پایین نباشد.
- میزان بیان اکثر ژن‌ها در محدوده‌ی صفر تغییر کند.

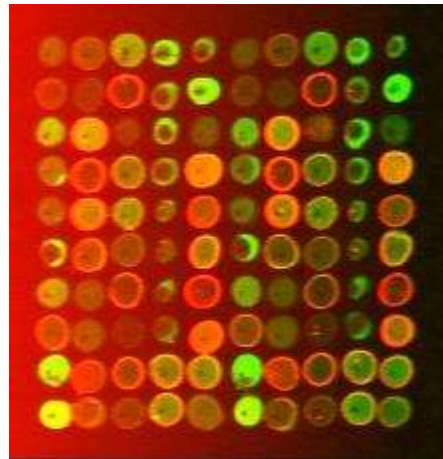
در دنیای واقعی تقریباً هیچ کدام از تصاویر دارای خصوصیات بالا نمی‌باشند. در حقیقت، شکل و محل قرارگیری خال‌ها با هم متفاوت و به صورت منظمی نیستند و تصویر دارای نویزهای متفاوتی می‌باشد. از آنجایی که اندازه بلوک‌ها و محل قرارگیری نقاط در هر تصویر با هم متفاوت می‌باشد. استفاده از یک ماسک خاص برای جداسازی ممکن نیست و باید بر روی هر تصویر به صورت مجزا عملیات شبکه‌بندی را اجرا کنیم. مراحل شبکه‌بندی تصاویر ریزآرایه به شرح زیر می‌باشد:

۱. رفع نویز سراسری از پس‌زمینه تصویر؛
۲. افزایش کنتراست تصویر؛
۳. حذف نویز با استفاده از عملگرهای ریخت‌شناسی؛
۴. اعمال روش آستانه‌گذاری برای تعیین موقعیت خال‌ها؛



شکل ۳-۵- تعیین موقعیت بلوک‌ها در یک تصویر ریزآرایه. الف) شناسایی موقعیت بلوک‌ها در تصویر ریزآرایه. ب) نمایش هیستوگرام افقی و رسم خطوط مربوط به تعیین موقعیت بلوک‌ها. ج) رسم هیستوگرام عمودی و رسم خطوط مربوط به تعیین موقعیت بلوک.

برای درک بهتر روش پیشنهادی یک تصویر ریزآرایه مصنوعی استاندارد به همراه نویز از دانشگاه استنفورد در شکل (۳-۶) تهیه شده است. این تصویر مربوط به هیچ یک از پایگاه داده‌ها نمی‌باشد. در ادامه تمامی عملیات شبکه‌بندی بر روی تصویر اجرا می‌شود.



شکل ۳-۶- تصویر ریزآرایه‌ی مصنوعی به همراه نویز بالا.

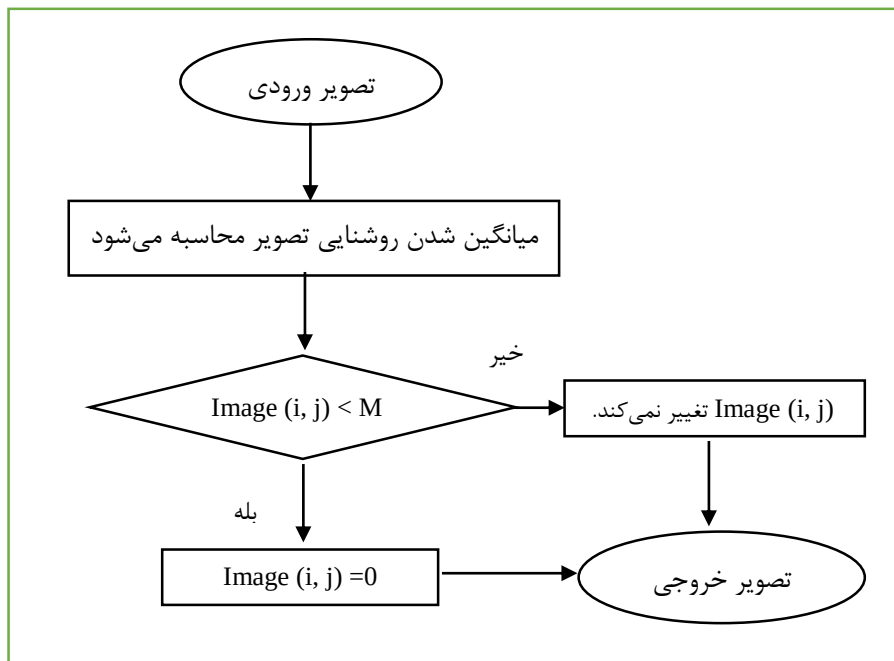
۱- رفع نویز سراسری از پس‌زمینه تصویر

برخی از خال‌های روشن در اطراف تصویر به دلیل نشستن گرد و غبار بر روی اسلاید است که هنگام اسکن، این گرد و غبار به صورت نویز در تصویر ظاهر می‌شود. با تبدیل تصویر رنگی به خاکستری این‌گونه نویزها در پس‌زمینه تصویر مقادیر خاکستری به خود می‌گیرند. برای حذف این نوع نویز لازم است ابتدا میانگین شدت روشنایی تصویر محاسبه و سپس مقدار حاصل با تمامی مقادیر پیکسل‌ها مقایسه شود. اگر مقدار پیکسلی از مقدار میانگین کمتر باشد آن را برابر با صفر قرار داده، در غیر این صورت مقدار تغییری نخواهد کرد. شکل (۳-۷) مراحل حذف نویز را نمایش می‌دهد.

۲- افزایش کنتراست تصویر

در بعضی از تصاویر ریزآرایه، خال‌ها شدت روشنایی پایینی دارند به طوریکه از پس‌زمینه به خوبی تشخیص داده نمی‌شوند. بنابراین در گام دوم با استفاده از یکنواخت کردن هیستوگرام، کنتراست تصویر بین خال‌ها و پس‌زمینه را افزایش می‌دهیم. متأسفانه، با افزایش کنتراست نویزهایی همانند لکه‌های

کوچک سفید رنگ بر روی پس‌زمینه تصویر ظاهر می‌شود. با اعمال فیلتر وینر تا حدودی این نویز برطرف می‌شود.



شکل ۳-۷- شمای کلی تصحیح نویز سراسری پس‌زمینه در تصویر ریز‌آرایه.

۳- حذف نویز با استفاده از عملگرهای ریخت‌شناسی

زمانی که تراشه داخل دستگاه اسکن گذاشته می‌شود، ممکن است هنگام اسکن نویز مشعل در تصویر همانند شکل (۳-۶) ایجاد شود. نویز مشعل ناشی از بازتاب و پراکندگی نور در لنز دستگاه اسکن می‌باشد. نویز مشعل^۱ گاهی باعث ایجاد خرابی‌های جبران‌ناپذیری در تصویر می‌شود به گونه‌ای که ممکن است نیمی از تصاویر با نور مشعل محو و اطلاعات مهم آن پاک شود. این نوع نویز تاثیر بسیار منفی برای تعیین موقعیت خال‌ها دارد. پس حذف این گونه نویز امر حیاتی برای شبکه‌بندی می‌باشد. بنابراین با استفاده از عملگرهای ریخت‌شناسی نویز مشعل حذف و خال‌های موجود در تصویر شناسایی می‌شوند. در ادامه نحوه حذف نویز به صورت کامل شرح داده می‌شود.

^۱ Flare Noise

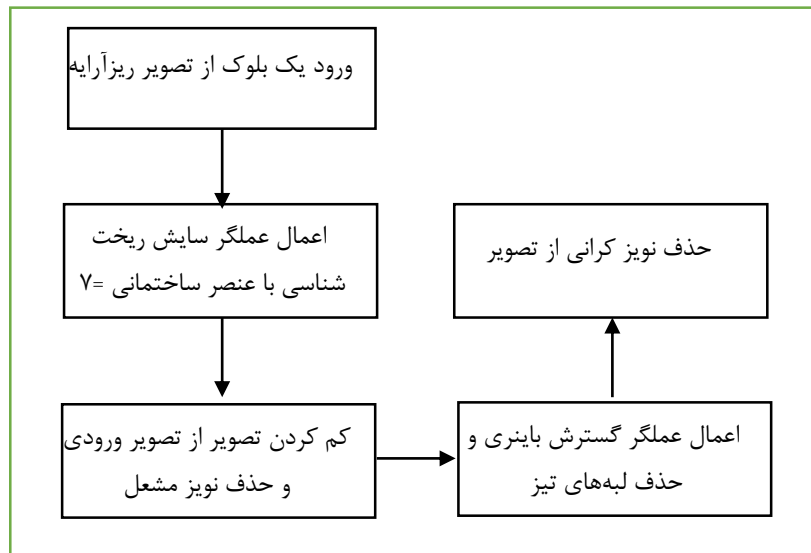
برای حذف خال‌ها در تصویر، از عملگر سایش مربوط به ریخت‌شناسی استفاده می‌شود. همانطور که از نام این عملگرد پیداست، این عملگر باعث سایش نقاط در تصویر می‌شود. از آنجایی که خال‌ها، پیش‌زمینه ی تصویر را تشکیل می‌دهند، ساختار عنصری تصویر به صورت صفحه دایره‌ای با مقدار ۷ تعریف می‌شود، که این مقدار میانگین تقریبی شعاع خال‌ها در تصویر می‌باشد. بعد از اعمال سایش، تصویر حاصل از تصویر اصلی کم شده و بدین ترتیب نویز مشعل حذف می‌گردد. برای منظم سازی تصویر و حذف لبه‌های تیز از عملگر سایش استفاده می‌شود. ساختار عنصری آن با استفاده از روش آزمایش و خطا به صورت صفحه گرد با مقدار ۴ بدست می‌آید. تصویر حاصل، تصویری هموار بدون لبه‌های تیز می‌باشد. اما هنوز مقداری نویز کرانی یا مرزی موجود است. برای حذف آن‌ها کافی است نویزها در محدوده مرزی هر خال را شناسایی و حذف کنیم. شکل (۳-۸) مراحل کار را به صورت کلی نمایش می‌دهد.

۴- اعمال روش آستانه‌گذاری برای تعیین موقعیت خال‌ها

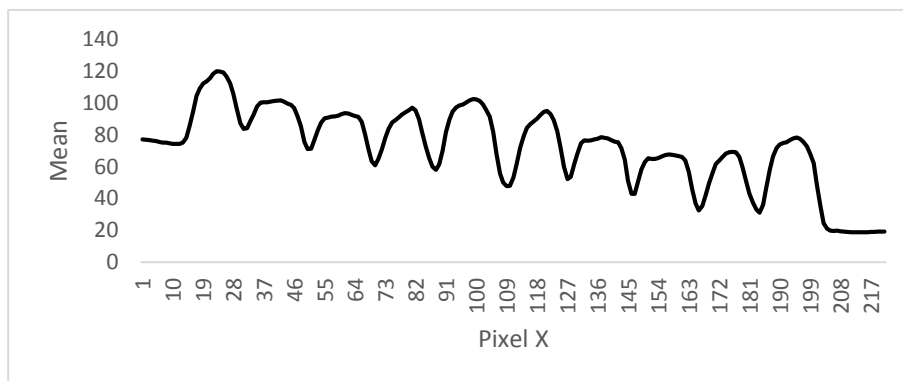
همان طور که قبلاً توضیح داده شد، روش آستانه‌گذاری تصویری یک روش بسیار ساده برای تعیین موقعیت اشیاء در تصویر می‌باشد. این روش مجدداً برای تعیین موقعیت خال‌ها استفاده می‌شود با این تفاوت که با استفاده از روش همبستگی و اتسو حد آستانه خال‌ها در تصویر تعیین می‌گردد. برای شروع میانگین شدت روشنایی پیکسل‌های افقی / عمودی تصویر (با ابعاد x, y) با استفاده از معادله (۳-۵) محاسبه می‌شود.

$$MH(y) = \frac{1}{n} \sum_{x=x_1}^{x=x_n} f(x, y) \quad (۳-۵)$$

با استفاده از تابع گسسته، هیستوگرامی بدست می‌آید (نمودار ۳-۲)، که محور عمودی مجموع پیکسل‌ها سطر / ستون و محور افقی موقعیت آن‌ها را نمایش می‌دهد.



شکل ۳-۸- شمای کلی حذف نویز مشعل در تصویر ریزآرایه.

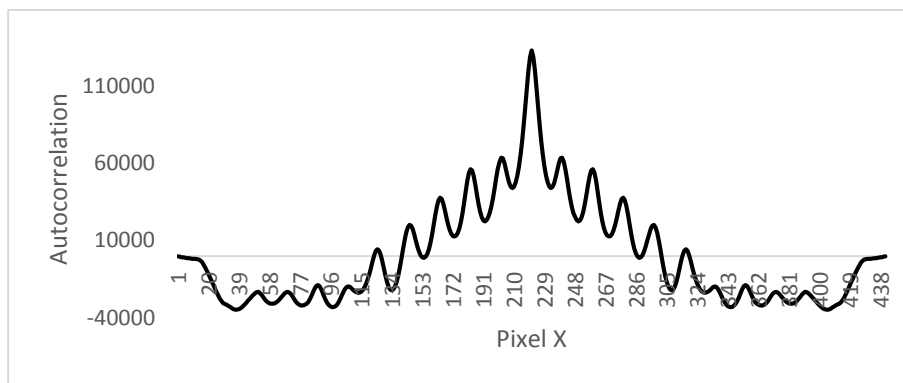


نمودار ۳-۲- هیستوگرام عمودی مربوط به تصویر ریزآرایه.

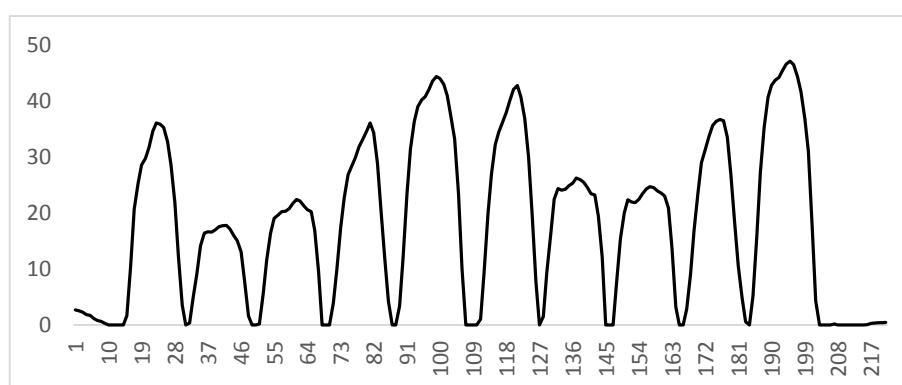
برای یافتن ناحیه هر یک از خال‌ها نمی‌توان به کمینه‌های محلی در نمودار اکتفا کرد و آن‌ها را ملاکی برای تعیین مختصات در نظر گرفت، زیرا در برخی از تصاویر ریزآرایه ممکن است برخی خال‌ها به هم چسبیده باشند. علاوه بر آن به خاطر وجود لکه‌های سفیدی که در حین آزمایش و اسکن تصویر به وجود می‌آید، شاید نواحی کمینه نشان‌گر مرز جدا کننده خال‌ها نباشند.

برای بهبود و منظم سازی، افزایش شباهت و حذف متغیرهای نامربوط در هیستوگرام از رابطه خود همبستگی^۱ استفاده می‌شود. این مفهوم، یک ابزار ریاضی برای یافتن الگوهای تکراری مانند حضور یک سیگنال تناوبی است، که تحت نویز تصویر قرار گرفته‌اند. خود همبستگی را می‌توان معیاری از میزان شباهت سیگنال با انتقال یافته‌ی خودش دانست. با اعمال رابطه‌ی خودهمبستگی بر روی نمودار (۲-۳) تابع یک بعدی (نمودار ۳-۳-الف) حاصل می‌شود. در این تابع نقاطی که دارای بیشینه محلی هستند مشخص کرده و میانگین آن‌ها تخمین زده می‌شود. میانگین این مقادیر، نشان دهنده فاصله تقریبی بین بیشترین و کمترین مقادیر متوالی محلی در هیستوگرام نمودار (۲-۳) است. در ادامه برای حذف نویز و منظم سازی نمودار اولیه از عملگر ریخت‌شناسی top-hat استفاده می‌شود. از آنجایی که داده‌های مورد بررسی یک بعدی هستند، ساختار عنصری آن‌ها به صورت خطی و با مقدار میانگین تخمین زده شده به کمک نمودار (۳-۳-الف)، در نظر گرفته می‌شود. با اعمال عملگر ریخت‌شناسی بر روی نمودار (۲-۳) هیستوگرام منظمی همانند نمودار (۳-۳-ب) حاصل می‌شود. برای شبکه‌بندی و شناسایی موقعیت خال‌ها بر روی هیستوگرام منظم بدست آمده، از روش تعیین حد آستانه اتسو استفاده می‌گردد [۶۳].

¹ Autocovariance



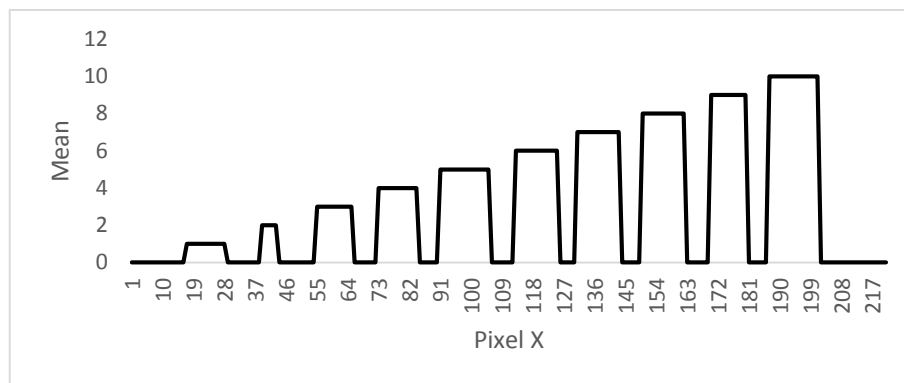
(الف)



(ب)

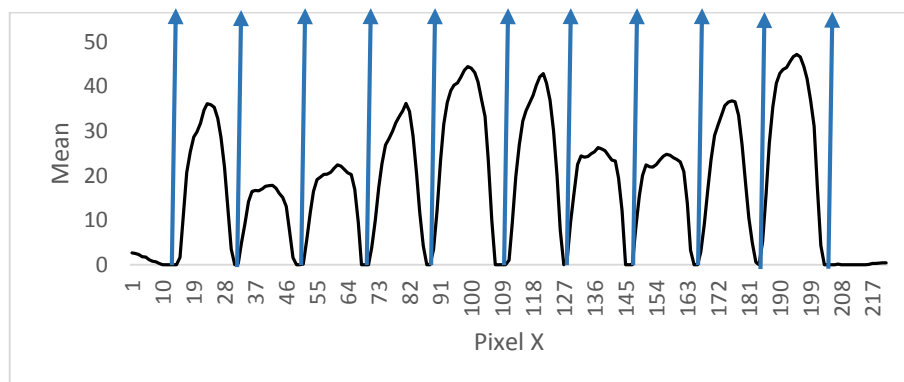
نمودار ۳ - ۳- منظم سازی و حذف نویز هیستوگرام به کمک تابع خود همبستگی. (الف) رسم تابع خود همبستگی حاصل از هیستوگرام عمودی مربوط به تصویر ریزآرایه. (ب) حذف نویز به کمک تابع خود همبستگی و با اعمال عملگر ریخت‌شناسی بر روی هیستوگرام عمودی اولیه تصویر و ترسیم منظم هیستوگرام.

هدف از استفاده‌ی این روش به حداقل رساندن پراکندگی داده‌ها در دو کلاس سیاه و سفید است. بنابراین نمودار به یک نمودار باینری تبدیل شده (نمودار ۳-۴) که در آن یک‌ها نمایان‌گر محدوده خال‌ها در تصویر ریزآرایه بوده و صفرها فاصله بین دو ستون از تصویر را نمایش می‌دهد. بعد از باینری کردن نمودار، برای مجزا شدن هر یک از نواحی آن‌ها را به ترتیب در اعدادی بزرگتر از صفر ضرب می‌کنیم که به این عمل برچسب‌گذاری گفته می‌شود. این عملیات باعث می‌شود که محدوده‌ی هر یک از خال‌ها بصورت مجزا تعیین گردد. نتیجه بدست آمده در نمودار (۳-۴) نمایش داده شده است.



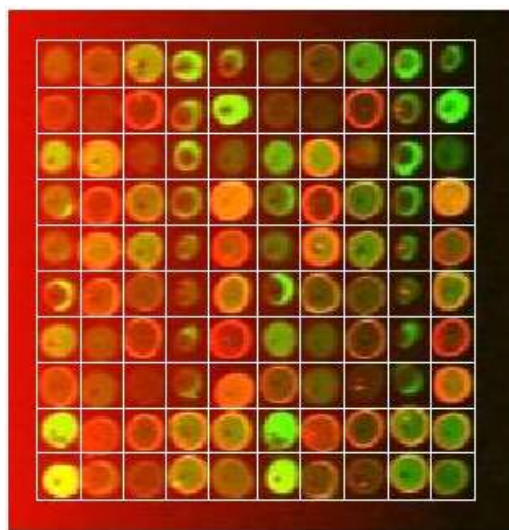
نمودار ۳ - ۴- نحوه‌ی برجسب گذاری نواحی خال‌ها به صورت مجزا.

در نمودار رسم شده مقادیر صفر فاصله‌ی مابین دو خال را بازگو می‌کند. برای محاسبه‌ی آن، ابتدا مرکز ثقل محدوده‌ی هر خال در نمودار بدست آورده، سپس فاصله‌ی مرکز ثقل هر دو قله متوالی را محاسبه و بر دو تقسیم می‌شود. مقدار بدست آمده مربوط به خط جدا کنند بین دو خال می‌باشد. در نمودار (۳-۵) محدوده‌ی خال‌ها با خطوط عمودی نمایش داده شده است.



نمودار ۳ - ۵- تعیین محدوده خال‌ها.

تا اینجا تصویر فقط به صورت عمودی پردازش شده است. برای شبکه‌بندی افقی خال‌ها کافی است تصویر را ترانهاد کرده و تمامی عملیات بالا را بر روی آن انجام دهیم. بدین ترتیب با استفاده از پارامترهای بدست آمده، خال‌ها در سلول‌های مجزایی از هم قرار می‌گیرند. نتیجه در شکل (۳-۹) نشان داده شده است. در پایان این مرحله، موقعیت هر یک از خال‌ها را در یک فرم Excel ذخیره کرده تا در صورت لزوم خال مورد نظر را فراخوانی شود.



شکل ۳-۹- نمایش شبکه‌بندی تصویر ریزآرایه.

قطعه‌بندی خال‌ها

در اکثر روش‌های موجود ابتدا با استفاده از روش‌های خوشه‌بندی یا برخی از رویه‌های پیشنهادی دیگر ابتدا خال‌ها بدون حذف پس‌زمینه از تصویر جدا می‌شوند [۴۷، ۵۰]، سپس با اجرای الگوریتم‌هایی دیگر همانند ریخت‌شناسی مقدار پس‌زمینه‌ی آن‌ها تشخیص و تصحیح می‌گردد. در حالی که با استفاده از یک روش قوی تخمین پس‌زمینه، نیازی به اجرای هر دو مرحله نیست. با اجرای الگوریتم تخمین دقیق پس‌زمینه بر روی هر یک از خال‌ها به صورت مجزا و حذف مقدار پس‌زمینه از تصویر، هر دو عملیات همزمان صورت می‌گیرد. در این قسمت سعی شده هر دو مرحله را با هم ادغام کنیم.

در طی مطالعات صورت گرفته مقدار پس‌زمینه در هر یک از اسلایدها شامل برخی از رشته‌های دی‌ان‌ای است که بر روی تراشه ریزآرایه، جفت مورد نظر خود را یافت نکرده و به ژن‌های مربوطه داخل تراشه متصل نشده است یا مربوط به مواد شیمیایی است که در زمان آزمایش بر روی تراشه باقی مانده‌اند. هر دوی این موارد با شستشوی تراشه که آخرین مرحله قبل از اسکن آن است به صورت کامل از بین نرفته و منجر به ایجاد یک پس‌زمینه با مقادیر غیر ثابت شده‌اند [۶۴].

بنابراین یکی از مهم‌ترین مشکلات تشخیص پس‌زمینه، ناهموار بودن سطح زمینه می‌باشد. در یک اسلاید ریزآرایه میزان شدت روشنایی هر یک از خال‌ها، ترکیبی از مقدار پس‌زمینه و مقدار هریک از

ژن‌ها است که توسط مکمل دی‌ان‌ای و آران‌ای هیبرید شده‌اند. تشخیص و حذف پس‌زمینه برای تعیین مقدار واقعی هیبریداسیون از مکمل دی‌ان‌ای امری ضروری است ولی ناهمگن بودن پس‌زمینه همواره این کار را بسیار دشوار کرده است. در صورتی که مقدار پس‌زمینه حذف نگردد موجب مشکلاتی در هنگام محاسبه بیان ژن می‌شود. از جمله مقدار بیان ژن ممکن است به شدت کاهش یابد یا پراکندگی داده‌ها افزایش یابد.

روش‌های متفاوتی برای تخمین و حذف پس‌زمینه در مقالات مختلف در نظر گرفته شده است. از جمله می‌توان به پس‌زمینه محلی، عملگر ریخت‌شناسی، پس‌زمینه ثابت و در نظر نگرفتن پس‌زمینه اشاره کرد.

روش ریخت‌شناسی بعد از سال ۱۹۹۹ در بسته نرم‌افزاری خال (spot) [۶۵] و در پژوهش دیگری در [۶۶] نیز بکار گرفته شد. عملکرد این روش نسبت به دو روش قبلی در برابر نویزهای موجود در تصویر و ناهمگنی آن برتری دارد. همچنین این روش در مقالات [۶۷، ۶۸] به ترتیب در سال‌های ۲۰۰۳ و ۲۰۰۲ به عنوان بهترین روش معرفی شد. به هر حال این روش دارای معایبی نیز می‌باشد. یکی از مهم‌ترین محدودیت‌ها هموارسازی لبه‌ها در پس‌زمینه می‌باشد. در نتیجه لبه‌های پس‌زمینه در خروجی پیش‌زمینه ظاهر خواهند شد. علاوه بر آن ریخت‌شناسی اثر بیش‌گشایش^۱ یا بیش‌فرسایش^۲ را به صورت محلی بر روی خال‌ها ایجاد می‌کند. سطح این اثر بستگی به اندازه پنجره عملگرهای ریخت‌شناسی دارد. هرچه اندازه این پنجره کوچکتر باشد، این اثرات نیز به مراتب کمتر می‌شوند. اما در برنامه‌های کاربردی باید اندازه پنجره بزرگتر از خال‌ها باشد. در این پژوهش برای حل این مشکلات تلفیق روش منظم‌سازی تغییرات کلی^۳ و حداقل‌سازی نرم L1 را پیشنهاد داده می‌شود.

^۱ Over Dilation

^۲ Over Erosion

^۳ Total Variation Regularization (TV)

۱- روش TV+L1

این روش در سال ۲۰۰۴ برای تجزیه تصویر به ویژگی‌های متفاوت معرفی شد [۶۹]. همچنین در سال ۲۰۰۵ از این الگوریتم برای تخمین پس‌زمینه تصاویر ریزآرایه مورد استفاده قرار گرفت [۲۶]. در این روش تصویر خاکستری دو بعدی به صورت تابع f نمایش داده می‌شود. این تابع متشکل از دو متغیر u و v می‌باشد. u شامل رنگ پس‌زمینه و مرزهای مهم است و مابقی تصویر که به عنوان بافت شناخته می‌شوند در v گنجانده شده است. فرمول TV+L1 در معادله (۶-۳) نمایش داده شده است.

$$\min_{u \in BV(\Omega)} \int_{\Omega} |\nabla u(x)| + \lambda |f(x) - u(x)| dx \quad (6-3)$$

قسمت اول فرمول مربوط به منظم‌سازی تغییرات کلی (TV) و قسمت دوم آن حداقل سازی نرم L1 می‌باشد. Ω دامنه تصویر و BV محدوده تغییرات را در دامنه تصویر مشخص می‌کند. بنابراین تابع f و u بر روی Ω تعریف می‌شود. زمانی که این مدل را برای تخمین پس‌زمینه بر روی تصاویر ریزآرایه اعمال می‌کنیم، f تصویر ریزآرایه ورودی است و u به عنوان پس‌زمینه، خروجی مسئله می‌باشد. پس می‌توان نتیجه گرفت $f-u$ فقط سیگنال خال را برمی‌گرداند. در معادله ذکر شده λ ضریب لاگرانژ است. حل صحیح معادله (۶-۳) بستگی به ضریب لاگرانژ دارد. در تصاویر ریزآرایه مقدار ضریب برابر $2/r_{max}$ انتخاب می‌شود [۲۶]. r برابر شعاع خال‌ها در تصاویر ریزآرایه می‌باشد.

برای حل معادله (۶-۳) از روش‌های مختلفی از جمله معادلات مشتقات جزئی و برنامه نویسی مخروط مرتبه دوم استفاده می‌شود [۶۹، ۷۰]. برای این کار، تصاویر ریزآرایه به عنوان تصاویر دو بعدی $m \times n$ درون ماتریس $R^{m \times n}$ نمایش داده می‌شوند. پس تصویر ورودی $f \in R^{m \times n}$ می‌باشد. همچنین f شامل پس‌زمینه ناهمگن $u \in R^{m \times n}$ است، که در این روند باید تخمین زده شود. فرم کلی معادله مشتق جزئی معادله (۶-۳) به صورت زیر معادله (۷-۳) می‌شود.

$$\nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) + \lambda \frac{f-u}{|f-u|} = 0 \quad (7-3)$$

در عمل برای جلوگیری از تقسیم صفر، یک مقدار خیلی کوچک $\varepsilon > 0$ به $|\nabla u|$ و $|f - u|$ اضافه می‌شود. برای حل معادله (۷-۳) از روش پله‌های زمان^۱ استفاده شده است. راه حل اثباتی معادله (۳-۳) در مقاله [۷۰] تکامل یافته و معادله (۸-۳) بدست می‌آید.

$$u_{i,j}^{n+1} = u_{i,j}^n + \delta_t \ell \partial_x^- \left(\frac{\partial_x^+ u_{i,j}^n}{M} \right) + \delta_t \ell \partial_y^- \left(\frac{\partial_y^+ u_{i,j}^n}{M} \right) + \delta_t \ell \lambda \frac{f_{i,j}^n - u_{i,j}^n}{\sqrt{(f_{i,j}^n - u_{i,j}^n)^2 + \varepsilon}} \quad (۸-۳)$$

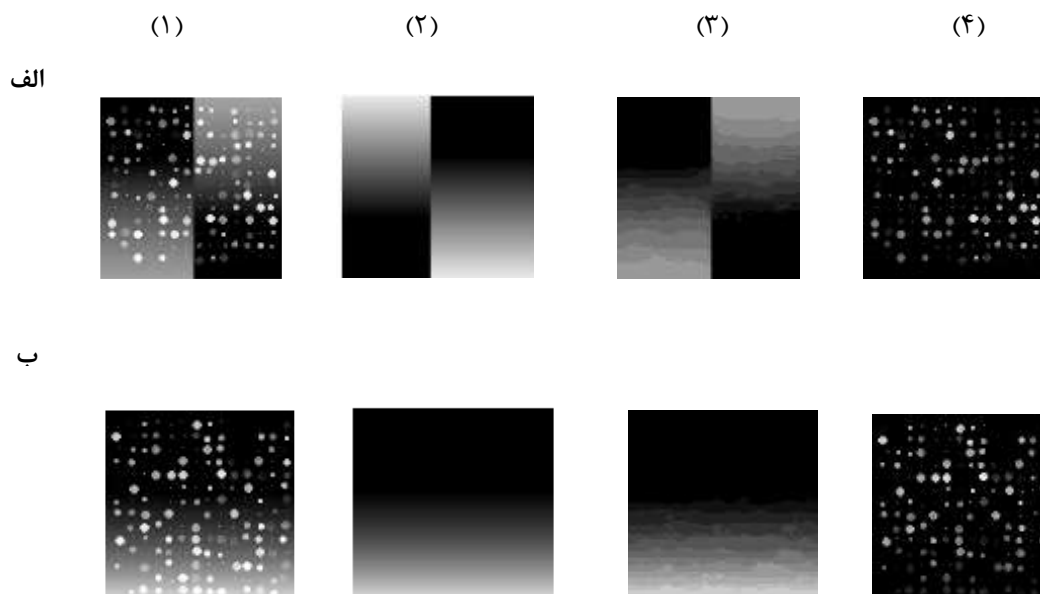
M در فرمول برابر (معادله ۹-۳):

$$M := \sqrt{(\partial_y^+ u_{i,j}^n)^2 + (\partial_x^+ u_{i,j}^n)^2 + \varepsilon} \quad u_{i,j}^0 = f_{i,j} \quad (۹-۳)$$

در معادله (۸-۳) $\delta_t \ell$ پله زمانی و ∂_x^- و ∂_y^+ اختلاف جزئی محدود^۲ به ترتیب رو به عقب و رو به جلو است. در هر تکرار فقط مشتقات گسسته مرتبه اول محاسبه می‌شود. در آزمایشات مختلف تعداد تکرار زیاد در معادله (۸-۳) برای رسیدن به u پایدار مورد نیاز است. زمانی که سیستم به حالت پایدار نزدیک می‌شود مقدار $\delta_t \ell$ خیلی کوچک و هنگام بروز رسانی معادله (۸-۳) تغییرات مهم و اساسی در u به وجود می‌آید. افزایش مقدار پله زمان موجب تشدید خطا، ناپایداری و کمتر شدن زمان محاسبه و کاهش آن علاوه بر مشکلات ذکر شده منجر به افزایش زمان محاسبه می‌گردد. اکنون با استفاده از داده‌های مصنوعی و ساختگی، عملکرد روش پیشنهادی در شکل (۳-۱۰) نمایش داده می‌شود. در این روش ضریب لاگرانژ برای تخمین پس‌زمینه با مقدار تقریبی $\frac{0}{3}$ انتخاب شده است. در تمامی داده‌های تست شده در شکل (۳-۱۰)، روش TV+L1 تقریباً می‌تواند، لبه‌های تصویر را حفظ و از ایجاد اعوجاج در پس‌زمینه جلوگیری کند.

^۱ Time Stepe

^۲ Finite Difference



شکل ۳-۱۰- حذف پس‌زمینه از تصاویر ریزآرایه مصنوعی- الف و ب در هر سطر مثالی جداگانه از حذف پس‌زمینه در داده‌های مصنوعی نشان می‌دهد. ستون‌ها: (۱) نمایش تصاویر اصلی، (۲) نمایش پس‌زمینه حقیقی تصویر، (۳) تخمین پس‌زمینه با استفاده از روش TV+L1، (۴) بازیابی تصویر اصلی با استفاده از روش TV+L1.

محاسبه بیان ژن

یکی از مهم‌ترین عوامل ایجاد سرطان، جهش‌های ژنتیکی در بیمار است. به بیان ساده فرض کنید دنباله‌ی ژن مرجع در فرد سالم به صورت زیر است.

...CTGATGATCTCGAATAGCGTGCGCGAATGAT...

در آزمایش ریزآرایه، نمونه بافت سالم با رنگ سبز و نمونه بافت سرطانی با رنگ قرمز برچسب گذاری می‌شود. بنابراین با فرض اینکه بیان ژن در لایه سبز به صورت زیر است. (این لایه نمونه آزمایشی سالم است)

ATGATCTCGAATAGCGTGCGCGAAT

و بیان ژن در لایه قرمز به صورت زیر است. (این لایه نمونه آزمایشی تومور است)

ATGATCTCGAATGCGTGCGCGAAT

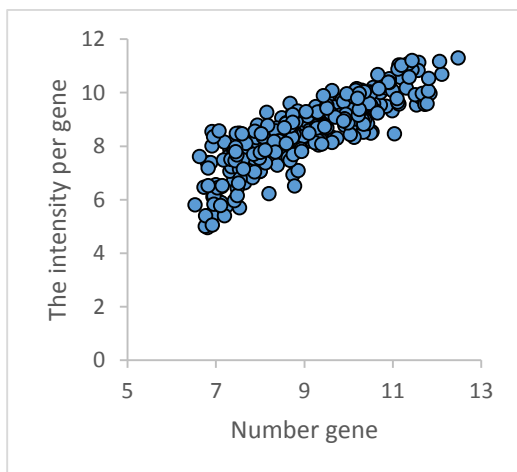
با مقایسه نمونه سالم و توموری متوجه ایجاد جهش در لایه قرمز می‌شویم. بنابراین برای تشخیص جهش کافی است دو نمونه سالم و توموری را با هم مقایسه و تفاوت آن‌ها را محاسبه کنیم. برای مقایسه شباهت‌ها و تفاوت‌ها بیان ژن معرفی شده است. تاکنون، تصاویر ریزآرایه بلوک‌بندی و سپس هر کدام بصورت مجزا شبکه‌بندی شده‌اند. بعد از بدست آوردن موقعیت خال‌ها، مقدار پس‌زمینه تخمین زده و از تصویر اصلی کم می‌شود. اکنون شدت روشنایی هر یک از خال‌ها را در تصویر ریزآرایه با توجه به معادله (۳-۱۰) محاسبه می‌گردد.

$$\text{intensity_spot} = \text{sum}(\text{rows}) + \text{sum}(\text{columns}) \quad (۳-۱۰)$$

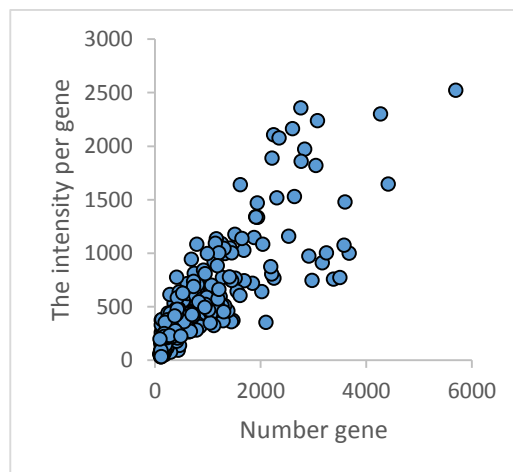
با توجه به اینکه هر یک از تصاویر ریزآرایه شامل دو کانال رنگی قرمز و سبز می‌باشد. معادله (۳-۱۰) به تمامی خال‌ها در هر دو کانال اعمال می‌شود. بعد از استخراج داده‌های خام عموماً تبدیلات لگاریتم آن‌ها محاسبه می‌شود [۷۱، ۷۲]. این تبدیلات توزیع داده‌های خام را بهبود و پراکندگی آن‌ها را کاهش می‌دهد. در نمودار (۳-۶) از ۳۰۰ داده ساختگی که نماینده ژن‌ها در کانال سبز و قرمز هستند، استفاده می‌شود. در این نمودار کاملاً مشهود است که تبدیلات لگاریتمی باعث کاهش پراکندگی داده‌ها می‌شود. داده‌ها با بازه تغییرات [۳۱-۵۶۹۸] به بازه تقریبی ۵-۱۳ کاهش یافته است. با لگاریتم گرفتن میزان شدت روشنایی استخراج شده از خال‌ها، درجه هیبرید شدن تک رشته‌های دی‌ان‌ای با نمونه‌های توموری و سالم بدست می‌آید [۷۳].

برای مقایسه داده‌های مختلف ریزآرایه، باید میزان جهش یا فعالیت هر یک از خال‌ها یا به عبارت دیگر میزان بیان ژن‌ها محاسبه شود. در اکثر مقالات برای تخمین این مقدار معادله (۳-۱۱) پیشنهاد شده است [۷۳، ۷۴].

$$\text{Gene} - \text{expression} = \log_2 \left(\frac{\text{int}(\text{cy5})}{\text{int}(\text{cy3})} \right) = \log_2(\text{int}(\text{cy5})) - \log_2(\text{int}(\text{cy3})) \quad (۳-۱۱)$$



(ب)



(الف)

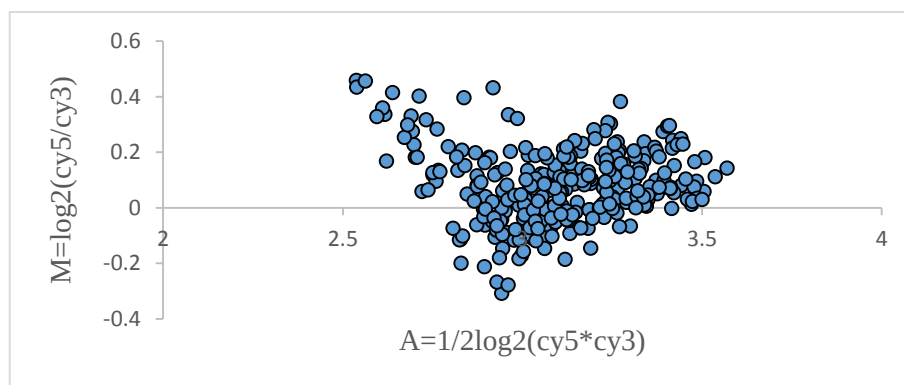
نمودار ۳ - ۶- پراکندگی داده‌های ریزآرایه. الف) استخراج داده‌های خام از دو کانال رنگی و نمایش میزان توزیع داده‌ها. ب) نمایش محاسبه تبدیلات لگاریتمی برای تمامی خال‌ها در دو کانال رنگی و کاهش میزان توزیع داده‌ها.

در معادله (۳-۱۱)، $int(cy3)$ و $int(cy5)$ به ترتیب نماینده درجه هیبرید شدن ژن‌های توموری و سالم است. اگر مقدار محاسبه شده تقریباً عدد صفر بدست آید به معنی برابری میزان فعالیت ژن‌ها می‌باشد. اگر حاصل آن یک عدد بزرگتر از صفر (یا کمتر از صفر) باشد میزان هیبرید شدن در نمونه $cy5$ بیشتر از $cy3$ (کمتر از $cy3$) همانند نمودار (۳-۷) می‌باشد. در این نمودار محور عمودی نمایانگر بیان ژن و محور افقی نماینده روشنایی خال‌ها در دو کانال است که با استفاده از معادله (۳-۱۲) محاسبه می‌شود.

$$Gene - expression = \frac{\log_2(int(cy5)) + \log_2(int(cy3))}{2} \quad (۳-۱۲)$$

این دو معادله (۳-۱۱) و (۳-۱۲) مکرراً در مقالات مختلف برای بیان گزارش نتایج آزمایش‌های ریزآرایه استفاده می‌شوند [۷۵، ۷۶].

برای شناسایی نوع تومور در یک سرطان فقط تعداد محدودی از ژن‌ها تغییر می‌کنند. بنابراین توزیع میزان بیان ژن‌ها در اطراف صفر باید متمرکز باشد. اما داده‌ها در نمودار (۳-۷) به صورت نامنظمی قرار گرفته‌اند. پس نرمال‌سازی داده‌ها یک امر حیاتی برای نظم‌دهی به داده‌های ریزآرایه می‌باشد. در بخش بعد، راه‌حلی برای این مشکل پیشنهاد می‌دهیم.



نمودار ۳ - ۷ بیان ژن‌ها با محاسبه میزان شدت روشنایی هر خال در تصویر ریزآرایه (با محاسبه میزان روشنایی خال‌ها در دو کانال رنگی تصویر). A میانگین شدت روشنایی خال‌ها و M میزان بیان ژن‌ها را نشان می‌دهد.

نرمال‌سازی

آزمایش‌های ریزآرایه از ابتدا تا تولید داده‌های خام در مراحل مختلفی نظیر استخراج دی‌ان‌ای، برچسب‌گذاری نمونه‌ها، اسکن تراشه و پردازش داده‌ها در معرض تغییرات ناخواسته‌ای بوده که ناشی از ناهم‌هنگی آزمایشگاهی هستند. مانند تفاوت خواص فیزیکی رنگ‌ها، بازده‌های برچسب‌گذاری و اسکن دو رنگ با تنظیمات متفاوت اسکنر، که باعث عدم تعادل بین رنگ‌ها^۱ می‌شوند. در اکثر موارد نسبت شدت روشنایی دو کانال مقدار دقیق بیان ژن را منعکس نمی‌کند. هدف از نرمال‌سازی اصلاح تفاوت و منظم‌سازی میزان شدت روشنایی بین خال‌ها می‌باشد. جهت حذف یا حداقل کردن این تغییرات، روش‌های آماری تحت عنوان تکنیک‌های نرمال‌سازی با فرضیات آماری و بیولوژیکی متفاوت پیشنهاد شده است [۷۷-۷۹].

۱- نرمال‌سازی^۲ Lowess

برای نرمال‌سازی داده‌های ریزآرایه با توجه به نوع تکنولوژی بکار رفته، روش مناسب انتخاب می‌شود. نرم‌افزار Bioconductor به صورت رایگان برای نرمال‌سازی داده‌ها در دسترس عموم قرار دارد. در این نرم‌افزار از الگوریتم GC-RMA برای ریزآرایه‌های تک کانال یا تک رنگ [۸۰] و از الگوریتم Lowess

^۱ Dye Bias

^۲ Locally weighted linear regression (Lowess)

برای ریزآرایه‌های دو کانال استفاده می‌شود [۷۷]. از آنجایی که تصاویر ما دو کانال می‌باشد، روش انتخابی ما نیز Lowess است.

در این قسمت فرض می‌شود اکثر ژن‌ها در دو کانال تصویر یکسان بوده و بیان ژنی یا به عبارت دیگر نسبت لگاریتمی آن‌ها در محدوده مشخص می‌باشد. بنابراین پراکندگی داده‌ها روی نمودار باید به شکل متقارن باشد. در حالیکه داده‌ها استخراج شده در اغلب تصاویر ریزآرایه همانند نمودار (۳-۷) در راستای صفر نیستند. برای حل این مشکل در ادامه روش نرمال‌سازی داده‌ها به روش Lowess به صورت کامل شرح داده می‌شود.

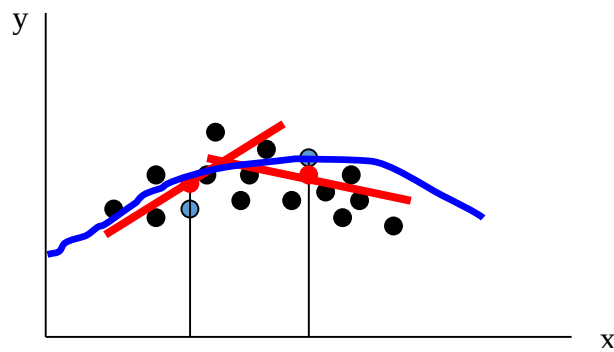
مطابق معادله (۳-۱۳) داده‌ها براساس منحنی برازش رگرسیون غیرخطی مبتنی بر داده یعنی $f(x)$ نرمال می‌شوند. به عبارت دیگر، در این معادله $f(x)$ تخمین میزان وابستگی $\log_2(R/G)$ نسبت به $1/2\log(R*G)$ است.

$$\text{Normalization} = \log_2 \left(\frac{R_i}{G_i} \right) - f(x) = \log_2 \left(\frac{R_i}{G_i} \right) - \log_2 (k(A)) = \log_2 \frac{R_i}{k(A)*G_i} \quad (۳-۱۳)$$

برای محاسبه منحنی برازش $f(x)$ ابتدا رگرسیون خطی محلی هر داده براساس داده‌های نزدیک به خود به صورت مجزا تخمین و رسم می‌شود. رگرسیون خطی هر داده مطابق معادله (۳-۱۴) محاسبه می‌شود.

$$y_i = a + f(x_i)x_i + \varepsilon \quad (۳-۱۴)$$

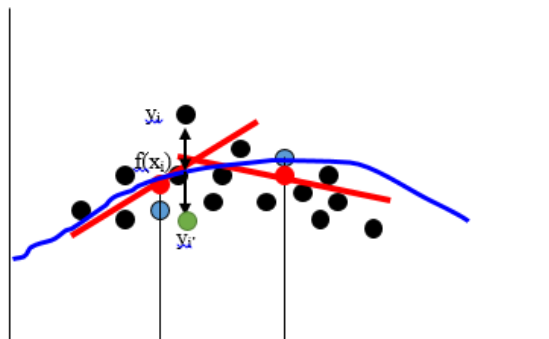
برای هر داده y_i , a عرض از مبدا و $f(x_i)$ شیب خط رگرسیونی و ε مقدار خطای تقریبی است. اکنون با توجه به رگرسیون‌های خطی رسم شده منحنی برازش رگرسیون غیرخطی با معادله (۳-۱۵) بدست می‌آید. برای درک بهتر به نمودار (۳-۸) مراجعه کنید.



نمودار ۳-۸- تخمین منحنی برازش بر روی داده‌های ریزآرایه.

$$Y = f(x) + \varepsilon \quad (3-15)$$

در نهایت اختلاف داده‌ها Y با منحنی برازش رسم شده در نمودار برابر مقدار نرمال داده‌ها می‌باشد. نمودار (۳-۹) نحوه محاسبه نرمال‌سازی برای یک داده را نمایش می‌دهد. در این نمودار y_i مقدار متغیر Y برای نمونه i ام است، مقدار برآورد رگرسیون آن بر روی برازش منحنی برابر $f(x_i)$ و اختلاف این مقدار برابر y_i' می‌باشد.



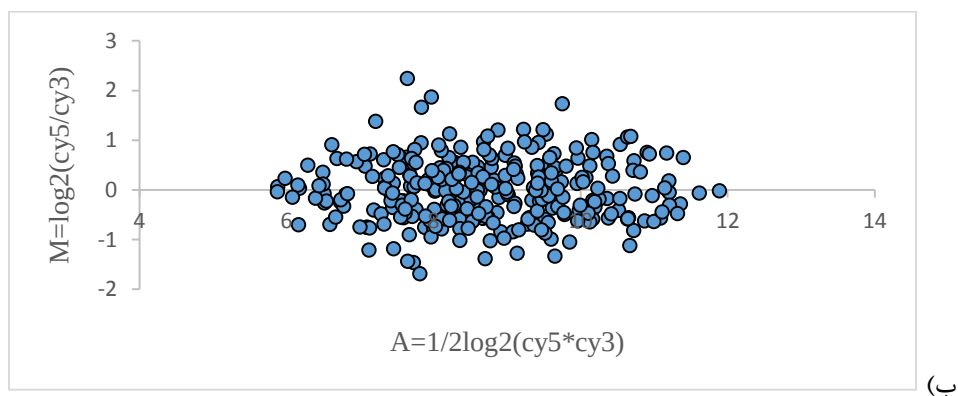
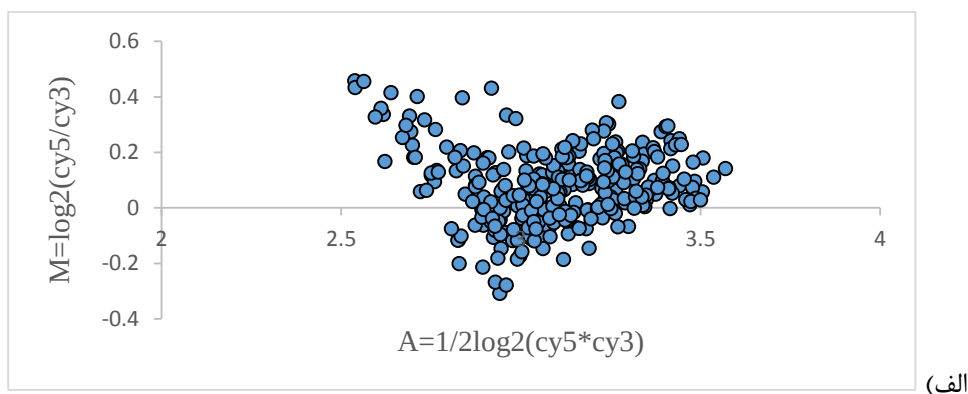
نمودار ۳-۹- محاسبه داده‌های نرمال با استفاده از منحنی برازش.

تمامی مراحل بالا بر روی ۳۰۰ داده‌ی ساختگی نیز اعمال می‌شود. نتیجه نرمال‌سازی داده‌ها کاهش پراکندگی و منظم‌سازی نمودار است. آثار آن در نمودار (۳-۱۰) کاملاً مشهود است.

انتخاب ویژگی

به دلیل حجم انبوهی از داده‌های بدست آمده از ریزآرایه، بررسی و کشف اطلاعات مفید بدون استفاده از روش‌های آماری و تکنولوژی اطلاعات ممکن نیست. علاوه بر آن اطلاعات مشابه در برخی از ژن‌ها

باعث عدم اعتبار بعضی از روش‌های تحلیل داده می‌شود. وجود ابعاد بالا در داده‌های ریزآرایه همچنین باعث افزایش هزینه محاسباتی و پیچیدگی روش‌های طبقه‌بندی می‌شود و از همه مهم‌تر، تفسیر ژن‌های عامل بیماری مشکل خواهد شد. همچنین از نظر زیست‌شناسان، مجموعه‌ای کوچکی از ژن‌ها عامل بیماری هستند، بنابراین حضور ژن‌های بسیار زیاد باعث کم‌رنگ شدن اثر ژن‌های عامل بیماری خواهند شد.



نمودار ۳ - ۱۰ - مقایسه داده‌های مصنوعی ریزآرایه قبل و بعد از عمل نرمال‌سازی بر روی نمودار M_A . (الف) پراکندگی داده‌ها قبل از عمل نرمال‌سازی. (ب) پراکندگی داده‌ها بعد از عمل نرمال‌سازی.

این نوع داده‌ها به دلیل داشتن صفات بسیار زیاد و همچنین نمونه‌های محدود مشکلاتی را برای الگوریتم‌های یادگیر و دسته‌بندی کننده ایجاد می‌کنند. در نتیجه باید قبل از عمل دسته‌بندی، نوعی ساده‌سازی و کاهش اطلاعات یا به عبارتی انتخاب ژن‌های مفید و تاثیرگذار از بین هزاران به دست آمده صورت گیرد. پس به طور خلاصه انتخاب ویژگی شامل پیدا کردن زیرمجموعه اصلی است که دسته‌بندی کننده با این زیرمجموعه، عملکرد بهتری نسبت به دسته‌بندی کننده با ویژگی‌های مجموعه ورودی

داشته باشد. در این بخش از روش بهره اطلاعات برای انتخاب ژن‌های مفید استفاده می‌شود. این روش به صورت گسترده در چندین مقاله استفاده و نتایج خوبی حاصل شده است [۸۱, ۸۲].

۱- بهره اطلاعات

اگر ویژگی هدف دارای C مقدار مختلف باشد. آنتروپی S نسبت به این دسته‌بندی C گانه، بصورت معادله (۱۶-۳) تعریف می‌شود.

$$Entropy(S) = \sum_{i=1}^C (-p_i \log_2 p_i) \quad (۱۶-۳)$$

که در آن p_i مطابق نسبتی از مقدار S است که به دسته i تعلق دارند (معادله ۱۷-۳).

$$p_i = \frac{|c_i|}{|\sum c|} \quad (۱۷-۳)$$

بهره اطلاعات یک ویژگی عبارت است از مقدار کاهش آنتروپی که بواسطه جداسازی مثال‌ها از طریق این ویژگی حاصل می‌شود. به عبارت دیگر بهره اطلاعات $Gain(S,A)$ برای یک ویژگی نظیر A نسبت به مجموعه مثال‌های S بصورت معادله (۱۸-۳) تعریف می‌شود.

$$Gain(S,A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|s_v|}{|S|} Entropy(s_v) \quad (۱۸-۳)$$

که در آن $Values(A)$ مجموعه‌ی همه مقدار ویژگی‌های A بوده و s_v زیرمجموعه‌ای از S است که برای آن A دارای مقدار v است. در تعریف فوق عبارت اول مقدار آنتروپی داده‌ها و عبارت دوم مقدار آنتروپی بعد از جداسازی داده‌هاست [۸۱]. هرچه مقدار بهره اطلاعات یک ویژگی بیشتر باشد، ویژگی A به عنوان ویژگی تاثیرگذار انتخاب می‌شود.

دسته‌بندی داده‌ها

پیش‌بینی و دسته‌بندی صحیح نوع تومور سرطانی در داده‌های ریزآرایه با استفاده از تحلیل میزان بیان ژن، برای تعیین مرحله درمانی بیمار، یا تجویز دارو نقش به‌سزایی دارد. دسته‌بندی دارای تکنیک‌های متنوعی است که در تحقیقات مختلف از آن‌ها استفاده می‌شود. از تکنیک‌های رایج در طبقه‌بندی

می‌توان به مواردی از جمله، شبکه عصبی، ماشین بردار پشتیبانی، طبقه‌بندی بیزین، سیستم‌های خبره، الگوریتم‌های ژنتیک و درخت تصمیم‌گیری اشاره کرد. درخت تصمیم‌گیری به علت سادگی و قابلیت فهم بالا از محبوبیت بالایی برخوردار هستند [۸۳، ۸۴].

درخت تصمیم‌گیری تنها مقادیر ویژگی‌های گسسته را پیشگویی می‌کند. این پیشگویی توسط متغیر کلاس به نام ویژگی هدف یا ویژگی وابسته صورت می‌گیرد. درخت تصمیم‌گیری دارای الگوریتم‌های مختلفی هستند که برخی از آن‌ها شامل ID3^۱، C4.5، C5، CART^۲، CHAID^۳ است. در این پایان‌نامه از الگوریتم درخت تصمیم‌گیری J48 به منظور طبقه‌بندی داده‌های ریزآرایه استفاده می‌شود. در ادامه با نحوه کار این الگوریتم بیشتر آشنا خواهید شد.

۱- درخت تصمیم‌گیری J48

درخت J48 درخت هرس شده‌ی C4.5 می‌باشد. الگوریتم استنتاج C4.5 در سال ۱۹۹۳ توسط کوئینلن تهیه شد [۸۵]. این الگوریتم متغیرهای پیوسته و گسسته را در محاسبات خود لحاظ کرده و مقادیر مفقود را در الگوریتم خود در نظر می‌گیرد و قادر به دسته‌بندی داده‌های نویزی نیز می‌باشد. این الگوریتم لزوماً دودویی نیست. برای انتخاب یک جداکننده بهینه در طول مسیر درخت تصمیم‌گیری از شاخص کسب اطلاعات یا کاهش آنتروپی استفاده می‌کند. برای درک مفاهیم کسب اطلاعات (بهره اطلاعات) و آنتروپی به فصل سوم مراجعه کنید. در این الگوریتم درخت تصمیم‌گیری از بالا به پایین ساخته می‌شود و از نوعی قانون هرس استفاده می‌کند. این الگوریتم با این سوال شروع می‌شود، کدام ویژگی باید در ریشه درخت مورد آزمایش قرار گیرد؟

برای یافتن جواب از یک آزمون آماری استفاده می‌شود تا مشخص گردد هر کدام تا چه حد قادر است به تنهایی نمونه‌های آزمایشی را دسته‌بندی کند. به این منظور در این جا از آنتروپی و بهره اطلاعات استفاده می‌شود. با انتخاب این ویژگی برای هر یک از مقادیر ممکن است یک شاخه ایجاد شود و نمونه‌های

^۱ Iterative Dichotomiser 3

^۲ Classification And Regression Tree

^۳ CHi-squared Automatic Interaction Detection

آموزشی براساس ویژگی هر شاخه مرتب می‌شوند. بدنبال آن عملیات فوق برای مثال‌های قرارگرفته در هر شاخه تکرار می‌شود تا بهترین ویژگی برای گره بعدی انتخاب شود. این الگوریتم یک جستجوی حریصانه است که در آن انتخاب‌های قبلی هرگز مورد بازبینی قرار نمی‌گیرند. در این جا حتی اگر هیچ خطایی در داده‌های آموزشی وجود نداشته باشد، هرس انجام می‌شود. که باعث عام‌تر شدن درخت شده و وابستگی آن به داده‌های آموزشی کم‌تر می‌شود.

بعد از اتمام کار قواعد مورد نظر از بررسی ریشه تا برگ‌ها بدست می‌آیند. هر مجموعه، ترکیبی از IF- THEN هاست که اساس این قواعد را عبارت شرطی بدست آمده از درخت 48j تشکیل می‌دهند. 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 آستانه برای ژن‌های $gene(i) \dots gene(k)$ است. C_j کلاس خروجی یا به عبارت دیگر نوع تومور را نشان می‌دهد.

۲-۳ نتیجه‌گیری

در این فصل، سیستمی برای تحلیل و دسته‌بندی تصاویر و داده‌های ریزآرایه معرفی شد. این سیستم به طور کلی از سه گام مهم و اساسی پردازش تصویر، داده‌کاوی و یادگیری ماشین تشکیل می‌شود. هر گام به صورت مجزا دارای چندین زیر مجموعه می‌باشد. در سیستم پیشنهادی برای پیاده‌سازی هر مرحله، الگوریتمی پیشنهاد داده می‌شود. مهم‌ترین مراحل سیستم پیشنهادی در گام پردازش تصویر شامل شبکه‌بندی و قطعه‌بندی می‌باشد. در مرحله شبکه‌بندی موقعیت هر خال در تصویر شناسایی و در مرحله قطعه‌بندی مقدار هر خال از تصویر تخمین زده می‌شود. در مرحله قطعه‌بندی تخمین صحیح و دقیق هر خال از اهمیت بالایی برخوردار است. بعد از استخراج داده از تصاویر ریزآرایه به دلیل اینکه حجم بالایی از اطلاعات در اختیار سیستم قرار می‌گیرد، اطلاعات موثرتر با روش بهره اطلاعات یافت می‌شوند. در نهایت داده‌های ریزآرایه به منظور تشخیص نوع سرطان دسته‌بندی می‌شوند. در فصل بعدی نتایج حاصل سیستم پیشنهادی با مقایسه دیگر سیستم‌ها بیان خواهد شد.

فصل چهارم

بررسی و تحلیل نتایج

۱-۴ مقدمه

در این پژوهش یک سیستم تمام خودکار جهت استخراج و آنالیز اطلاعات از تصاویر ریزآرایه، برای تشخیص نوع سرطان ارائه گردید. سیستم طراحی شده بر روی سه دسته از تصاویر ریزآرایه سرطانی آزمایش می‌شود. بدین منظور ابتدا تصاویر با کمک تکنیک‌های پردازش تصویر در طی چهار مرحله: تشخیص زاویه چرخش تصویر، تشخیص موقعیت بلوک، تشخیص موقعیت خال‌ها و قطعه‌بندی آن‌ها، تبدیل به داده‌های خام می‌شوند، سپس به کمک تکنیک‌های داده‌کاوی ابعاد داده کاهش و با کمک یادگیری ماشین داده‌های ریزآرایه طبقه‌بندی می‌شوند. در این بخش عملکرد قسمت‌های مختلف روش پیشنهادی بر روی تصاویر و داده‌های ریزآرایه مورد ارزیابی قرار می‌گیرد.

۲-۴ پایگاه داده

در این بخش مقدمه کوتاهی در مورد مجموعه داده‌های جمع‌آوری شده شرح داده می‌شود. مطالعات صورت گرفته در این پایان‌نامه مربوط به تصاویر ریزآرایه سرطان سینه^۱، سرطان خون^۲ و سرطان غدد لنفاوی^۳ است. هر دسته سرطان شامل دو زیر مجموعه مجزا می‌باشد. در این پژوهش مجموعه تصاویر ریزآرایه مورد استفاده از پایگاه داده دانشگاه استنفورد تهیه شده است [۸۶]. هر تصویر رنگی ریزآرایه، به دو تصویر خاکستری مربوط به کانال قرمز و سبز تقسیم می‌شود. کانال قرمز نماینده ژن‌های بیمار و کانال سبز نماینده ژن‌های فرد سالم می‌باشد. هر تصویر خاکستری شامل ۴۸ بلوک است و هر بلوک بسته به نوع تصویر ریزآرایه مابین ۸۷۰ تا ۹۶۱ خال دارد. این تصاویر در قالب عدد صحیح دو بایتی نمایش داده می‌شوند. در ادامه شرح مختصری از هر پایگاه داده بیان می‌شود.

۱-۲-۴ پایگاه داده سرطان سینه

این پایگاه داده شامل ۷۱ تصویر ریزآرایه است. از این تعداد ۲۲ نمونه آن مربوط به سرطان سینه خوش‌خیم^۴ می‌باشند [۸۷]. متخصصان ژنتیک برای تهیه این پایگاه داده، حدود ۷۰-۸۰٪ سلول‌های

^۱ Breast Cancer

^۲ Myeloid Leukemia

^۳ Lymphoma Cancer

^۴ Primary Breast Tumor

سرطانی خوش‌خیم را کشت و سپس تمام دی‌ان‌ای آن را برداشت کردند. مابقی داده‌ها مربوط به نمونه‌های سرطانی است که تومور به صورت موضعی و حاد پیشرفت داشته است و افراد بیمار تحت شیمی درمانی نئوآدجوانت^۱ می‌باشند [۸۸]. هر نمونه از تصاویر درون این مجموعه شامل ۴۱۷۶۰ ژن می‌باشد. در سیستم طراحی شده داده‌های سرطان سینه به دو دسته آموزش و آزمایش تقسیم می‌شوند. از بین تمامی آن‌ها ۴۹ (تقریباً ۷۰٪) نمونه آزمایش و ۲۲ (تقریباً ۳۰٪) نمونه تست در اختیار سیستم قرار می‌گیرد.

۲-۲-۴ پایگاه داده سرطان غدد لنفاوی

این پایگاه داده شامل ۴۶ نمونه است. از این تعداد ۱۳ نمونه مربوط به بافت لنفاوی بدخیم می‌باشد. برای تهیه این پایگاه داده، ابتدا طی شش سال بیماران در کنار شیمی درمانی خود داروی ریتوکسی‌مب^۲ نیز مصرف می‌کردند. سپس متخصصان مقداری از تومور بدخیم را برش و مورد آزمایش خود قرار داده‌اند که نتیجه آن ۱۳ نمونه مربوط به بافت بدخیم لنفاوی می‌شود [۸۹]. مابقی نمونه‌های دیگر در لنفوم فولیکار^۳ مربوط به انتشار سلول بزرگ تهاجمی^۴ B در بافت لنفاوی می‌باشد [۹۰]. هر نمونه از تصاویر موجود در این مجموعه شامل از ۳۷۶۳۲ ژن می‌باشد. در سیستم طراحی شده داده‌های سرطان غدد لنفاوی به دو دسته آموزش و آزمایش تقسیم می‌شوند. از بین تمامی نمونه‌ها ۳۲ (تقریباً ۷۰٪) نمونه آزمایش و ۱۴ (تقریباً ۳۰٪) نمونه تست در اختیار سیستم قرار می‌گیرد.

۳-۲-۴ پایگاه داده سرطان خون

این پایگاه داده شامل ۴۷ نمونه است. از این تعداد ۱۰ نمونه آن مربوط به ۱۷ رده‌ی سلول‌هایی^۵ است که انحرافات اولیه کروموزمی در لایه‌های زیرین مغز استخوان را نشان می‌دهند [۹۱]. مابقی مربوط به سرطان خون حاد در مغز استخوان می‌باشد. این افراد حدود چهار سال تحت درمان دانشگاه Ulm آلمان

^۱ neoadjuvant

^۲ Rituximab

^۳ Follicular Lymphoma

^۴ Diffuse Large B Cell Lymphoma (DLBCL)

^۵ Cell Line

بوده‌اند [۹۲]. هر نمونه از تصاویر این مجموعه نیز دارای ۴۳۲۰۰ ژن می‌باشد. این مجموعه نیز همانند دو مجموعه قبل به دو دسته آموزش و آزمایش تقسیم می‌شود. از بین نمونه‌های موجود ۳۰ (تقریباً ۷۰٪) نمونه آزمایش و ۱۷ (تقریباً ۳۰٪) نمونه تست در اختیار سیستم قرار می‌گیرد.

۳-۴ بررسی و تحلیل نتایج

در این بخش عملکرد سیستم به صورت مجزا بر روی پایگاه داده‌های موجود بررسی و در هر قسمت نتیجه‌ی عملکرد بخش‌های مختلف سیستم بر روی تصاویر و داده‌های ریزآرایه مورد تجزیه و تحلیل قرار می‌گیرد.

۱-۳-۴ نتایج تنظیم چرخش تصویر

برای تعیین میزان صحت اثر تبدیلات رادن برای تصحیح چرخش تصاویر ریزآرایه، ابتدا دو تصویر با زاویه چرخش صفر درجه انتخاب می‌شود. سپس تصاویر به صورت دستی در جهت عقربه ساعت و عکس آن با زاویه انحراف 0° ، 25° ، 50° و 75° چرخانده می‌شوند. با استفاده از روش پیشنهادی مقدار زاویه چرخش θ_{est} تخمین زده می‌شود. محاسبه میزان دقت الگوریتم، با مقایسه دو زاویه واقعی θ_{real} و تخمینی θ_{est} در معادله (۱-۴) امکان پذیر می‌باشد.

$$Accuracy = \frac{|\theta_{real} - \theta_{est}|}{\theta_{real}} \times 100 \quad (1-4)$$

در جدول (۱-۴) میزان کارایی روش پیشنهادی بر روی دو تصویر چرخیده شده، نمایش داده می‌شود. این نتایج نشان می‌دهد که روش پیشنهادی قادر است با دقت بالایی زاویه چرخش تصویر را در هر دو جهت به خوبی تشخیص دهد و تصحیح کند.

جدول ۴-۱- محاسبه میانگین دقت روش پیشنهادی برای تصحیح زاویه چرخش بر روی دو تصویر.

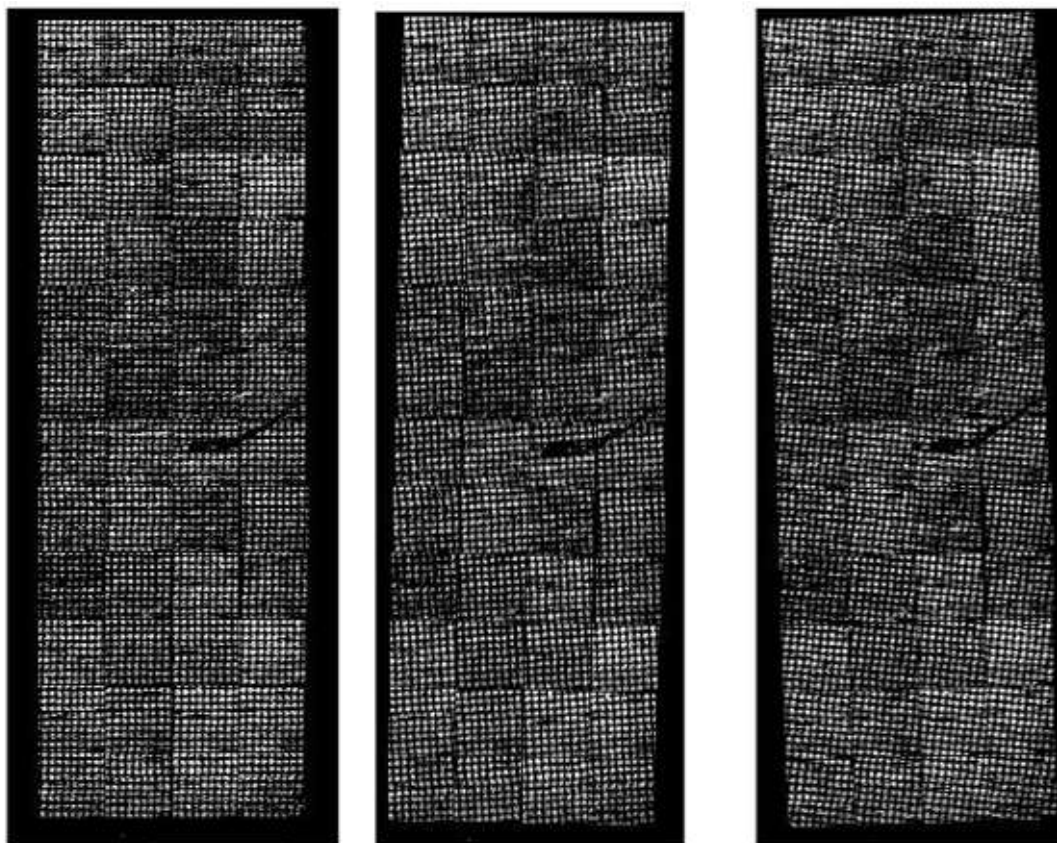
زاویه چرخش	درصد درستی تصویر ۲	درصد درستی تصویر ۱
۱°	٪۱۰۰	٪۹۹
۰٫۷۵°	٪۹۸٫۶۷	٪۹۹
۰٫۵°	٪۹۸	٪۹۹
۰٫۲۵°	٪۹۸	٪۹۸
-۱°	٪۹۹	٪۱۰۰
-۰٫۷۵°	٪۹۸	٪۹۹٫۳
-۰٫۵°	٪۱۰۰	٪۹۹
-۰٫۲۵°	٪۹۸	٪۹۸٫۶

برای درک بهتر اثر تبدیلات رادن بر روی تصاویر ریزآرایه مثالی در شکل (۴-۱) نمایش داده می‌شود. در این نمونه ابتدا تصویری با زاویه چرخش صفر درجه انتخاب، سپس به صورت دستی در دو زاویه $۱/۵^{\circ}$ و $-۱/۵^{\circ}$ ، در جهت عقربه ساعت و عکس آن چرخانده (شکل ۴-۱-الف-ب)، سپس به کمک الگوریتم تبدیلات رادن میزان انحراف تصویر تشخیص و با زاویه‌ی مناسب چرخش تصحیح می‌گردد (شکل ۴-۱-ج).

۲-۳-۴ نتایج بلوک‌بندی و شبکه‌بندی تصاویر ریزآرایه

برای تعیین تعداد و شناسایی محل بلوک‌ها در تصاویر ریزآرایه همانند شکل (۴-۲) از روش آستانه‌گذاری هیستوگرام استفاده شد. در این روش با کمک مجموع پیکسل‌ها در هر سطر و ستون تابع یک بعدی حاصل می‌شود. با رسم هیستوگرام آن به راحتی نواحی هر بلوک مشخص می‌گردد. در ادامه نیز، برای شناسایی موقعیت هر بلوک‌ها ابتدا عملیات حذف نویز صورت گرفت، سپس با میانگین‌گیری پیکسل‌ها در هر سطر و ستون تابع یک بعدی حاصل گردید. با رسم هیستوگرام مربوط به تابع یک بعدی و به کمک تکنیک‌های آستانه‌گذاری تصویر، محل دقیق هر خال مشخص می‌شود.

در شکل (۳-۴) موقعیت هر خال در یک بلوک از تصویر ریزآرایه شناسایی شده است. همان‌طور که قبلاً توضیح داده شد برای تعیین موقعیت هر خال از روش آستانه‌گذاری اتسو و ضریب خود همبستگی استفاده می‌شود. در این تصویر تمامی خال‌ها به وسیله‌ی خطوط سفید رنگی از هم مجزا شده‌اند.

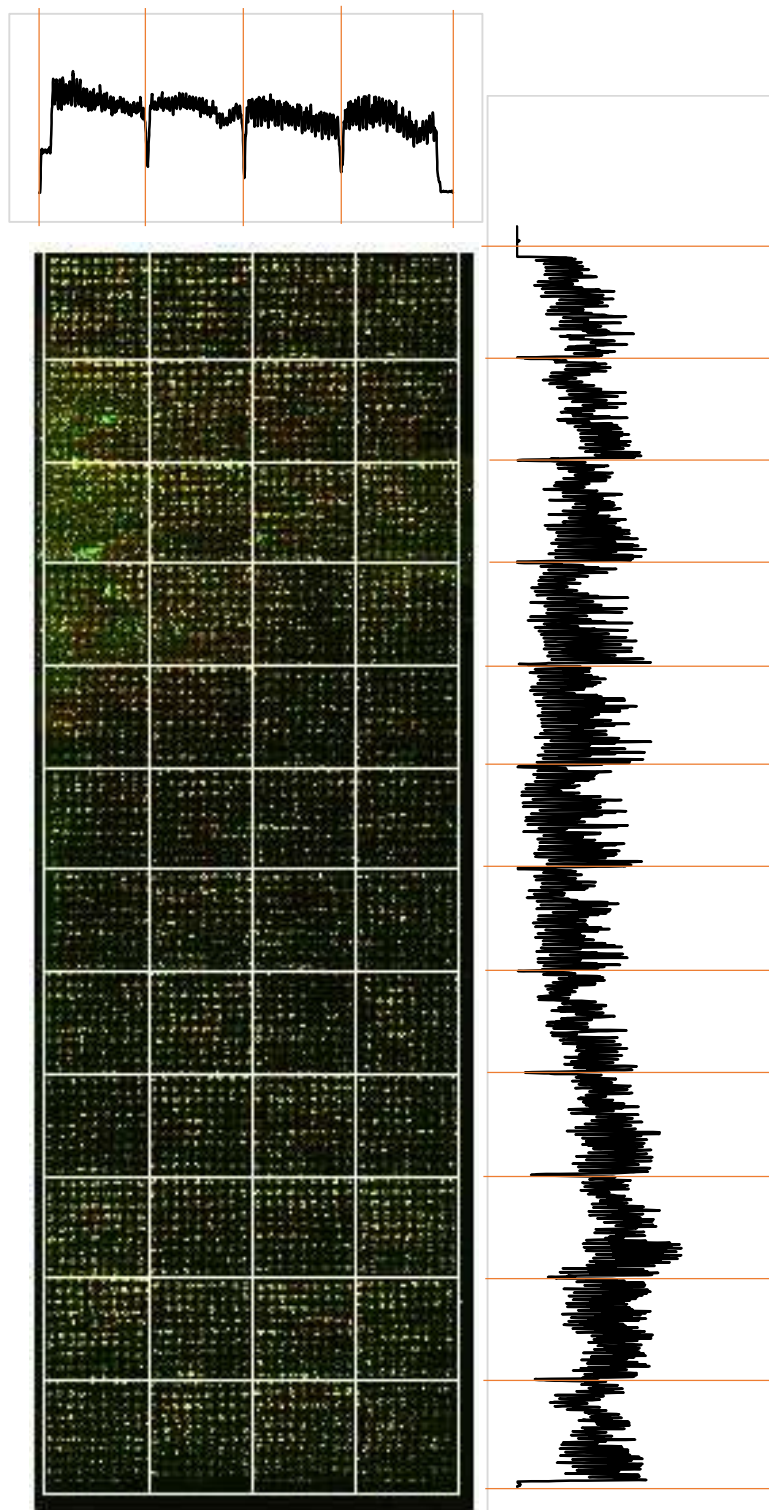


(ج)

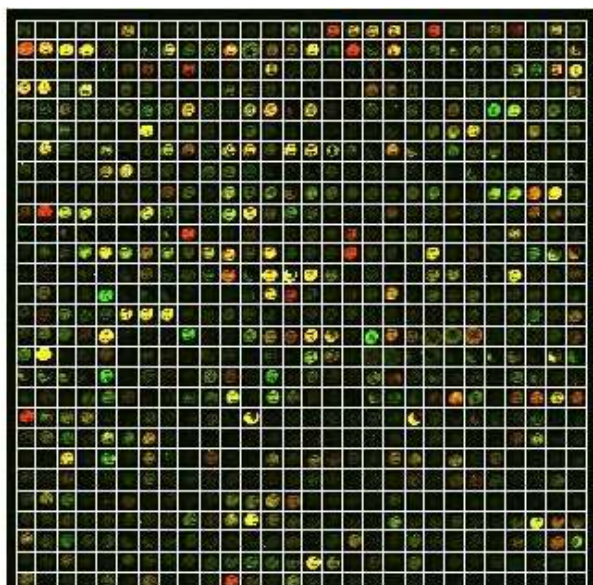
(ب)

(الف)

شکل ۴-۱- 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.



شکل ۴-۲- تعیین موقعیت بلوک‌ها در یک تصویر ریزآرایه.



شکل ۴-۳- شبکه‌بندی تصویر ریزآرایه.

برای تعیین دقت شبکه‌بندی هر تصویر از معادله (۲-۴) استفاده می‌شود.

$$\text{Accuracy} = \left(\frac{N(\text{correctspot})}{N(\text{total spot})} \right) * 100 \quad (۲-۴)$$

در معادله بالا تعداد خال‌هایی که به صورت صحیح شبکه‌بندی شده‌اند $N(\text{correctspot})$ با تعداد کل خال‌ها، $N(\text{total spot})$ ، مقایسه می‌شوند. برای ارزیابی روش پیشنهادی به منظور تعیین موقعیت بلوک‌ها و شناسایی مکان خال‌ها در هر بلوک، الگوریتم بر روی سه تصویر با کیفیت‌های متفاوت آزمایش شده است. این نتایج در جدول (۲-۴) ذکر شده و بیانگر دقت روش پیشنهادی بر روی سه تصویر می‌باشد. در بسیاری از تصاویر، عملکرد روش پیشنهادی بیش از ۹۸٪ می‌باشد و فقط تعداد کمی از تصاویری که شامل نویز و نقص قابل توجهی هستند دقت آن‌ها به طور متوسط به ۹۸٪ کاهش می‌یابد.

جدول ۴-۲- 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	٪۹۹٫۷۶	٪۹۸٫۲۳

۳-۳-۴ نتایج قطعه‌بندی خال از تصویر

برای تخمین پس‌زمینه، معروف‌ترین روش استفاده از عملگرهای ریخت‌شناسی می‌باشد. اما در این پایان‌نامه روشی پیشنهاد شده که نتایج خروجی آن همواره از روش‌های ریخت‌شناسی بهتر و دقیق‌تر است. اکنون با استفاده از داده‌های ساختگی، روش پیشنهادی با عملگر ریخت‌شناسی در شکل (۴-۴) مقایسه می‌شود. در این روش برای عملگرهای ریخت‌شناسی از پنجره 7×7 استفاده می‌شود. به همین دلیل ضریب لاگرانژ برای تخمین پس‌زمینه با مقدار تقریبی 0.3 انتخاب شده است. در تمامی داده‌های تست شده در شکل (۴-۴)، روش $TV+L1$ مقدار پس‌زمینه صحیح‌تری را نسبت به روش ریخت‌شناسی تخمین می‌زند. در هر دو حالت روش $TV+L1$ لبه‌ها را حفظ در حالی که روش ریخت‌شناسی اعوجاج زیادی در لبه‌ها ایجاد می‌کند، همچنین مقداری از شدت پس‌زمینه در شکل (۴-۴) الف) هنوز باقی‌مانده است.

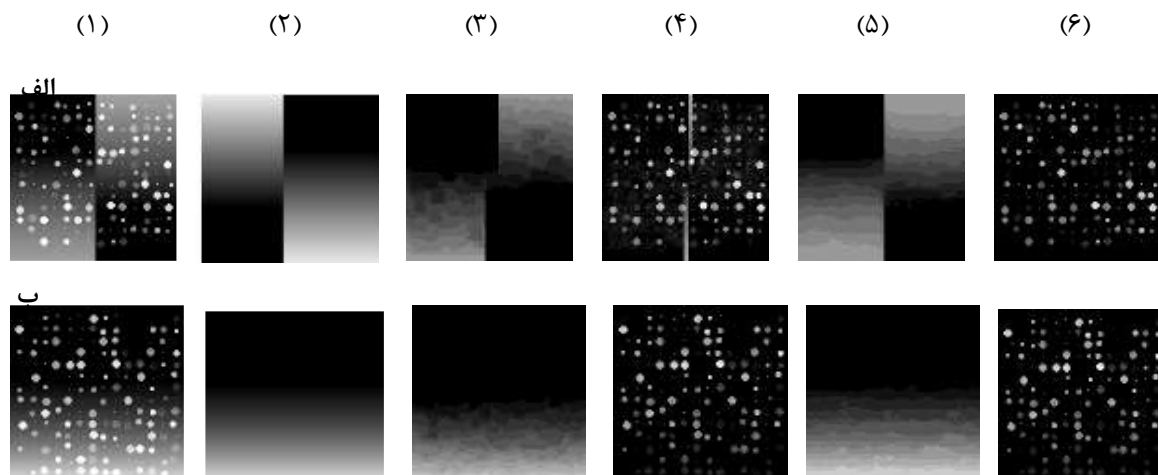
برای ارزیابی دقیق‌تر سیستم با استفاده از معادله (۴-۳)، میانگین خطای متوسط شدت روشنایی هر تصویر محاسبه می‌شود.

$$Accuracy = \sum_{i=1}^m \frac{\sum_{j=1}^n |u_{i,j}^{true} - u_{i,j}^{est}|}{m \times n} \quad (3-4)$$

$u_{i,j}^{est}$ و $u_{i,j}^{true}$ مقادیر مربوط به شدت روشنایی واقعی و تخمینی پس‌زمینه حاصل از اعمال الگوریتم پیشنهادی در ماتریس $m \times n$ می‌باشد. مقدار متوسط خطا در جدول (۴-۳) نمایش داده می‌شود. با توجه به شکل (۴-۴) و جدول (۴-۳) روش پیشنهادی همواره از روش ریخت‌شناسی بهتر عمل می‌کنند. در شکل (۴-۵) ماتریسی از یک خال در یک تصویر واقعی ریزآرایه استخراج شده است. با اجرای روش پیشنهادی بر روی این خال مقدار پس‌زمینه تخمین زده از ماتریس اصلی حذف شده است. همانطور که در شکل (۴-۵) ب) مشاهده می‌شود، شدت روشنایی پس‌زمینه صفر و به طور تقریبی مقادیر پیکسل‌های پیش‌زمینه مربوط به خال باقی‌مانده است. بنابراین مجموع مقادیر پیکسل‌ها بیانگر شدت روشنایی خال بدون احتساب پس‌زمینه خواهد بود.

۴-۳-۴ نتایج محاسبه بیان ژن از تصاویر ریزآرایه

همانطور که قبلاً ذکر شد یکی از مهم‌ترین عوامل ایجاد سرطان، جهش‌های ژنتیکی در بیمار است. بدین منظور درجه هیبرید شدن ژن‌های توموری و سالم مشخص شده و بعد از آن میزان بیان هر ژن محاسبه می‌شود. برای محقق شدن این امر ابتدا شدت روشنایی هر خال بدست می‌آید، سپس برای بهبود توزیع داده‌های خام و کاهش پراکندگی عموماً تبدیلات لگاریتم آن‌ها محاسبه می‌شود. در نمودار (۴-۱) داده‌های واقعی ریزآرایه خام و تبدیلات لگاریتمی آن‌ها در نمودار رسم شده است. در این دو نمودار واضح است که پراکندگی داده‌های به شدت کاهش یافته است.



شکل ۴-۴ - مقایسه روش پیشنهادی و روش ریخت‌شناسی برای حذف پس‌زمینه از تصاویر ریزآرایه - الف و ب در هر سطر مثالی جداگانه از حذف پس‌زمینه در داده‌های مصنوعی نشان می‌دهد. ستون‌ها: (۱) نمایش تصاویر اصلی، (۲) نمایش پس‌زمینه حقیقی تصویر، (۳) تخمین پس‌زمینه با استفاده از روش ریخت‌شناسی، (۵) بازیابی تصویر اصلی با استفاده از روش ریخت‌شناسی، (۶) تخمین پس‌زمینه با استفاده از روش TV+L1، (۷) بازیابی تصویر اصلی با استفاده از روش TV+L1.

جدول ۴-۳ - محاسبه میانگین خطا برای مقایسه دو روش ریخت‌شناسی و TV+L1 برای حذف پس‌زمینه.

تصاویر در شکل (۴-۴-الف) سطر دوم تصاویر در شکل (۴-۴-ب)		
۴/۶۱	۱/۱۴	میانگین خطا در روش ریخت‌شناسی
۱/۷۶	۰/۱۸	میانگین خطا در روش TV+L1



(الف)

143	400	740	203	218	310	335	52	44	257	49	151	214	166	216
215	188	222	142	145	118	231	164	180	84	316	63	453	201	432
106	9	272	116	311	73	227	50	351	63	35	136	197	357	254
403	148	79	294	304	410	850	1028	1101	1235	132	95	220	144	182
421	256	124	16	304	1159	1171	1287	1071	1090	1288	600	98	257	135
220	477	162	731	1224	917	1015	968	1596	1186	921	744	253	212	72
157	150	421	1400	1918	1526	1024	1349	746	522	1062	494	347	54	222
245	81	402	987	597	1642	1062	1014	975	772	723	1104	869	153	220
191	30	191	774	1515	1361	1086	760	1140	1126	1870	1488	643	359	191
264	275	284	845	2024	1143	944	1494	780	1093	1632	1582	208	680	50
233	150	138	220	666	1375	1482	1614	1180	1025	1969	616	208	152	236
522	171	406	147	361	802	1780	909	1001	1179	558	269	349	227	149
92	164	34	218	396	116	156	435	311	307	253	288	165	164	41
151	280	578	355	201	317	224	265	97	259	248	255	128	409	177

(ب)

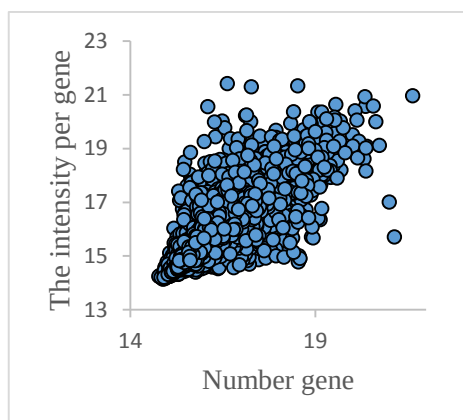
0	96	136	0	0	0	31	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	12	0	49	0	128
0	0	0	0	7	0	0	0	47	0	0	0	0	53	0
99	0	0	0	0	106	546	724	797	931	0	0	0	0	0
117	0	0	0	0	855	867	983	767	786	984	296	0	0	0
0	173	0	427	920	613	711	664	1292	882	617	440	0	0	0
0	0	117	1096	1614	1222	720	1045	442	218	758	190	43	0	0
0	0	98	683	293	1338	758	710	671	468	419	800	565	0	0
0	0	0	470	1211	1057	782	456	836	822	1566	1184	339	55	0
0	0	0	541	1720	839	640	1190	476	789	1328	1278	101	376	0
0	0	0	0	362	1071	1178	1310	876	721	1665	312	101	0	0
218	0	54	0	57	498	1476	605	697	875	254	0	45	0	0
0	0	0	0	92	0	0	131	7	3	0	0	0	0	0
0	0	140	51	0	13	0	0	0	0	0	0	0	105	0

(ج)

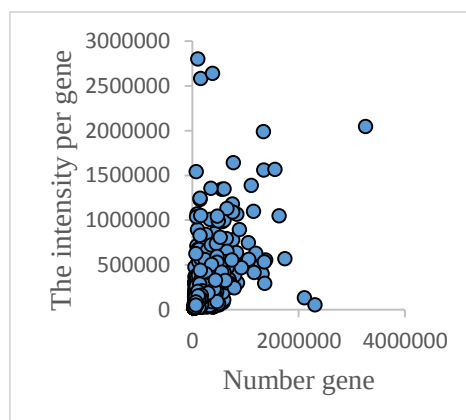
شکل ۴-۵- مقادیر پیکسل‌های یک خال در تصویر ریزآرایه و حذف پس‌زمینه با روش پیشنهادی. (الف) خال استخراج شده از تصویر ریزآرایه، (ب) نمایش ماتریس خال بدون حذف پس‌زمینه، (ج) نمایش ماتریس خال با اعمال روش پیشنهادی و حذف پس‌زمینه.

در نمودار (۴-۱) محور افقی نماینده درجه هیبرید شدن داده‌های توموری و محور عمودی نماینده درجه هیبرید شدن داده‌های سالم می‌باشد. میزان بیان هر ژن در یک تصویر ریزآرایه مربوط به سرطان سینه در نمودار (۴-۲) نمایش می‌دهد. همانطور که در نمودار (۴-۲) مشاهده می‌شود، داده‌های ریزآرایه اصطلاحاً به شکل موزی^۱ در می‌آیند. از آنجایی که انتظار می‌رود، بیان ژن‌ها در حول محور صفر تغییر کنند، نرمال‌سازی برای داده‌ها امری حیاتی است. در ادامه نتایج نرمال‌سازی داده‌ها شرح داده می‌شود.

¹ Bnana-Shaped

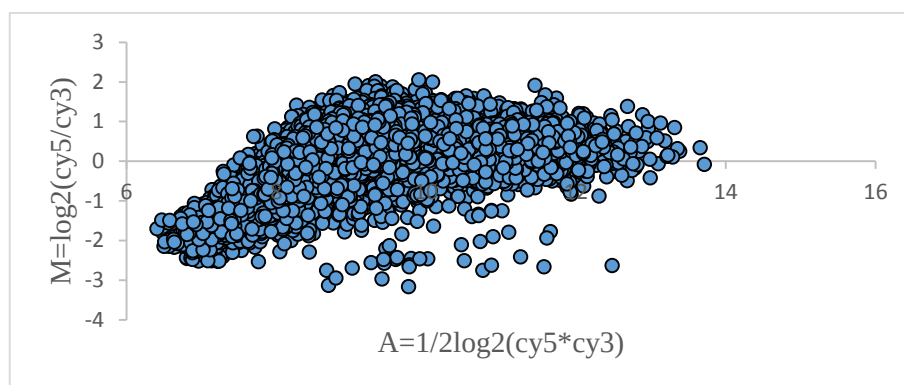


(ب)



(الف)

نمودار ۴-۱- مقایسه پراکندگی داده‌های خام استخراج شده از تصویر ریزآرایه با تبدیلات لگاریتمی آن‌ها (محور افقی و عمودی به ترتیب نماینده شماره ژن‌ها و میزان سطح خاکستری هر یک از ژن‌ها می‌باشد). الف) استخراج داده‌های خام از دو کانال رنگی و نمایش میزان توزیع داده‌ها. ب) نمایش محاسبه تبدیلات لگاریتمی برای تمامی خال‌ها در دو کانال رنگی و کاهش میزان توزیع داده‌ها.

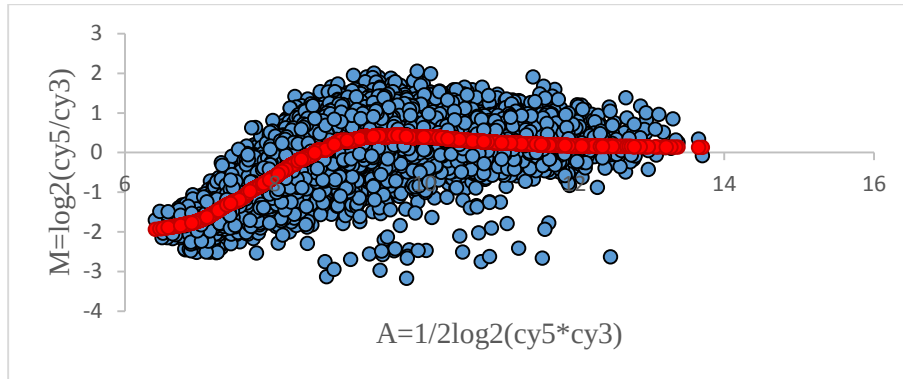


نمودار ۴-۲- بیان ژن‌های محاسبه شده از تصویر ریزآرایه (محاسبه میزان روشنایی خال‌ها در دو کانال رنگی تصویر). A میانگین شدت روشنایی خال‌ها و M میزان بیان ژن‌ها را نشان می‌دهد.

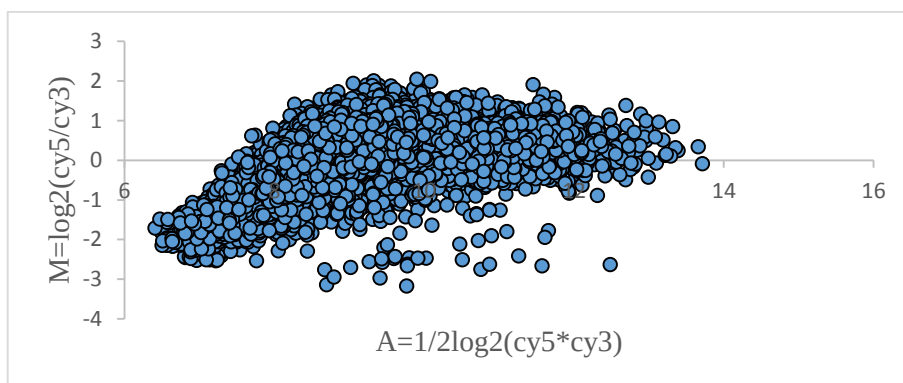
۴-۳-۵ نرمال‌سازی داده‌های ریزآرایه

همانطور که قبلاً شرح داده شد، نرمال‌سازی با استفاده از الگوریتم Lowess صورت می‌گیرد. در این روش ابتدا مطابق نمودار (۴-۳) منحنی برازش رگرسیون بر روی داده‌های ریزآرایه واقعی سرطان سینه محاسبه می‌شود. اختلاف تمامی داده‌ها با منحنی برازش رسم شده در نمودار برابر مقدار نرمال داده‌ها می‌باشد. در نمودار (۴-۴) داده‌ها قبل و بعد از نرمال‌سازی با هم مقایسه می‌شوند. قبل از نرمال‌سازی بسیاری از ژن‌ها، بیان ژن متفاوتی نسبت به مابقی داده‌ها دارند، اما بعد از نرمال‌سازی این اختلاف

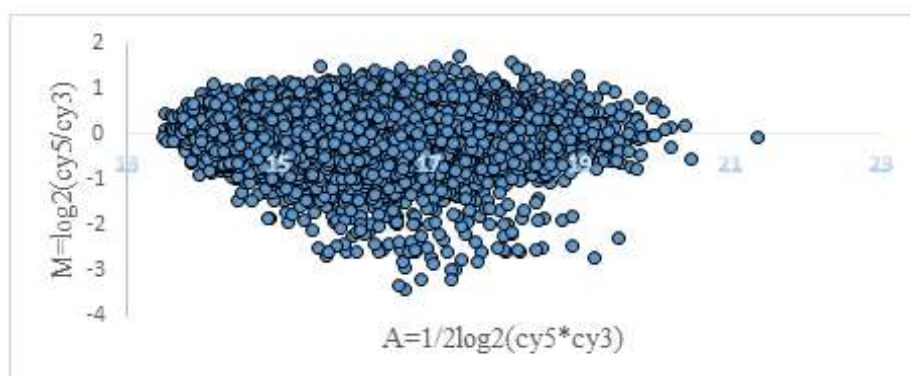
کاهش یافته است و پراکندگی داده‌ها بر روی نمودار تقریباً به شکل متقارن و منظم‌تری شکل گرفته است.



نمودار ۴-۳- تعیین منحنی برازش بر روی داده‌های ریزآرایه با استفاده از روش Lowess.



(الف)



(ب)

نمودار ۴-۴- مقایسه داده‌های ریزآرایه قبل و بعد از عمل نرمال‌سازی در نمودار M_A . (الف) پراکندگی داده‌ها قبل از عمل نرمال‌سازی. (ب) پراکندگی داده‌ها بعد از عمل نرمال‌سازی.

۴-۳-۶ نتایج استخراج ویژگی بر روی پایگاه داده ریزآرایه

با توجه به این حقیقت که تنها چند ژن برای تشخیص بیماری کافی است، با اعمال بهره اطلاعات بر روی داده‌های استخراج شده از تصویر ریزآرایه، ۱۰۰ ژن مفید با بالاترین ضریب IG از پایگاه داده‌ی سرطان سینه، سرطان غدد لنفاوی و سرطان خون مربوط به مغز استخوان به منظور دسته‌بندی داده‌ها استخراج شده است. در ادامه از بین ۱۰۰ ژن، ۱۰ ژن برتر به همراه تمامی ضریب بهره اطلاعاتی‌شان نمایش داده می‌شود.

۱- پایگاه داده سرطان سینه

از بین هزاران ژن در مجموعه داده‌های سرطان سینه، ۱۰۰ ژن با بالاترین ضریب بهره اطلاعاتی انتخاب و تنها ۱۰ ژن موثرتر در جدول (۴-۴) نمایش داده می‌شود.

جدول ۴-۴ 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

۲- پایگاه داده سرطان غدد لنفاوی

در مجموعه داده‌های سرطانی غدد لنفاوی، ۱۰۰ ژن با بالاترین ضریب بهره اطلاعاتی انتخاب و ۱۰ ژن برتر در جدول (۴-۵) نمایش داده می‌شود.

جدول ۴-۵ ده ژن برتر استخراج شده از سرطان غدد لنفاوی با اعمال روش بهره اطلاعات.

rank	Gene-number	IG value	rank	Gene-number	IG value
1	Gene-11635	0.859	6	Gene-21502	0.859
2	Gene-21509	0.859	7	Gene-35450	0.850
3	Gene-9827	0.859	8	Gene-35453	0.843
4	Gene-28234	0.859	9	Gene-4749	0.843
5	Gene-15184	0.859	10	Gene-21498	0.842

۳- پایگاه داده سرطان خون در مغز استخوان

در جدول (۴-۶) با استفاده از روش بهره اطلاعات، ۱۰ ژن برتر مربوط به پایگاه داده سرطان خون در مغز استخوان نمایش داده می‌شود.

جدول ۴-۶- ده ژن برتر استخراج شده از سرطان خون با اعمال روش بهره اطلاعات.

rank	Gene-number	IG value	rank	Gene-number	IG value
1	Gene-411	0.747	6	Gene-27679	0.644
2	Gene-26780	0.678	7	Gene-11622	0.644
3	Gene-36757	0.67	8	Gene-33564	0.644
4	Gene-13429	0.659	9	Gene-8099	0.644
5	Gene-28195	0.65	10	Gene-10812	0.644

۴-۳-۶ نتایج دسته‌بندی داده‌های ریز آرایه برای تعیین نوع سرطان

در این پایان‌نامه، از دسته‌بندی درخت تصمیم‌گیری برای تشخیص نوع سرطان توسط تحلیل ۱۰۰ ژن برتر استفاده می‌شود. بدین منظور برای هر پایگاه داده دو نوع سرطان موجود و در هر کدام ۱۰۰ ژن برتر به عنوان داده‌های ورودی به درخت تصمیم‌گیری اعمال می‌شود. خروجی درخت تصمیم‌گیری بیان‌گر نوع سرطان می‌باشد. پس از اتمام کار قواعد با خواندن از ریشه تا برگ نوشته می‌شود. مراحل و نتیجه هر پایگاه داده در ادامه به صورت مجزا شرح داده می‌شود.

برای ارزیابی عملیات مورد نظر، از پارامتر دقت معادله (۴-۴) و روش اعتبارسنجی متقابل^۱ استفاده می‌شود.

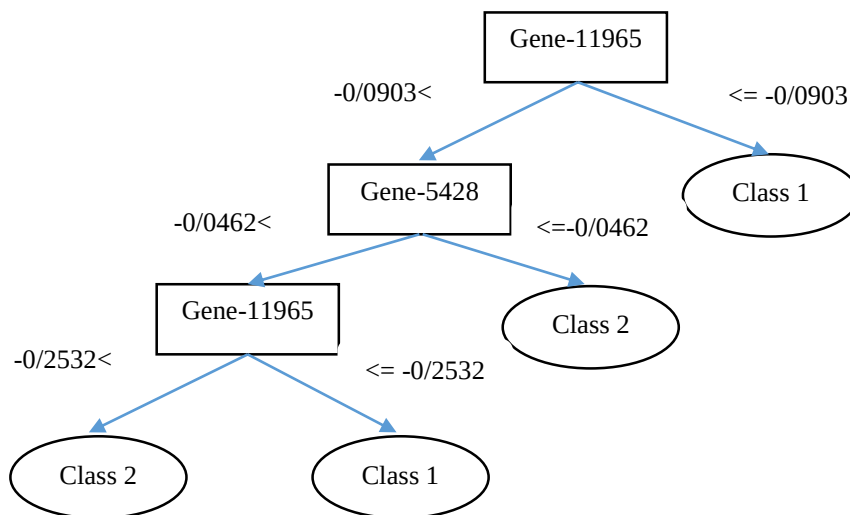
$$Accuracy = \frac{(TP+FN)}{(TP+TF+FP+FN)} = \frac{\text{Number of Correct Classified Instance}}{\text{Number of All Instance}} \quad (4-4)$$

¹ Cross-Validation

درخت تصمیم‌گیری، از نظر آماری اطلاعات کاملی در مورد مثبت صحیح^۱، مثبت کاذب^۲، منفی صحیح^۳ و منفی کاذب^۴ می‌دهد. همراه با این مقادیر دقت دسته‌بندی کننده با استفاده از فرمول بالا محاسبه می‌شود. در روش اعتبارسنجی متقابل در هر بار اجرا هر داده دقیقاً یک‌بار در مرحله آموزش و یک‌بار در مرحله تست قرار می‌گیرد. مقدار دقت نهایی از میانگین جمع دقت‌ها مربوط به آزمایش‌ها بدست می‌آید.

۱- درخت تصمیم‌گیری برای پایگاه داده سرطان سینه

در شکل (۴-۶) درخت تصمیم‌گیری برای تشخیص نوع سرطان سینه رسم شده است. خروجی به صورت دو کلاس نمایش داده می‌شود. کلاس یک نماینده تومور خوش‌خیم و کلاس دو نماینده تومور بدخیم می‌باشد.



شکل ۴-۶- درخت تصمیم‌گیری برای تشخیص نوع سرطان سینه.

با استفاده از چهار قانون برای تشخیص نوع تومور بدست می‌آید. شکل مجموعه قواعد تعیین شده به صورت زیر است.

$R1: \text{if } (gene_{11965} \leq -0.0903) \text{ then class } 1$

¹ True Positive (TP)

² Fals Positive (FP)

³ True Negative (TN)

⁴ Fals Negative (FN)

$R2: \text{if } (gene_{11965} > -0.0903) \text{ and } (gene_{5428} > 0.0467) \text{ and } (gene_{11965} > 0.2532) \text{ then class 2}$

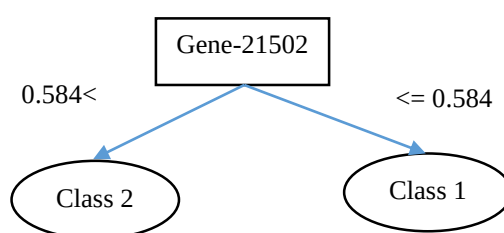
$R3: \text{if } (gene_{11965} > -0.0903) \text{ and } (gene_{5428} \leq 0.0467) \text{ then class 2}$

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

همانطور که قبلاً بیان شده، ۴۹ نمونه از داده‌ها به عنوان آموزش و ۲۲ نمونه به عنوان آزمایش در نظر گرفته شد. میزان صحت شناسایی بیماری برای دو نوع سرطان سینه، به همراه چهار قانون و دو ژن ۹۵/۴۵٪ می‌باشد.

۲- درخت تصمیم‌گیری برای پایگاه داده سرطان غدد لنفاوی

درخت تصمیم‌گیری برای دسته‌بندی سرطان غدد لنفاوی در شکل (۴-۷) ترسیم شده است. کلاس یک مربوط به انتشار سلول بزرگ تهاجمی B و کلاس دو نماینده بافت لنفاوی بدخیم می‌باشد.



شکل ۴-۷- درخت تصمیم‌گیری برای تشخیص نوع سرطان غدد لنفاوی.

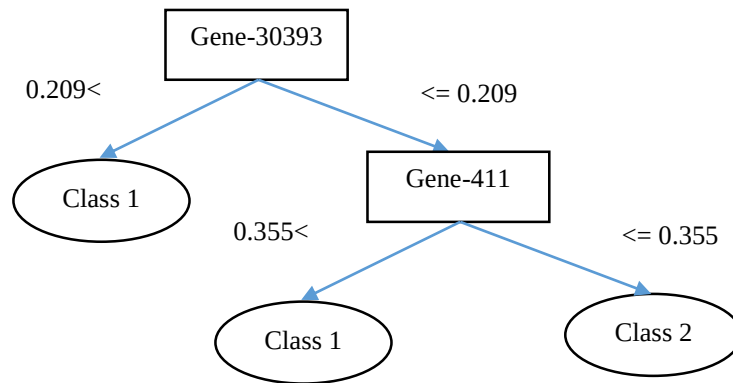
تنها با استفاده از یک ژن و دو قانون، نوع سرطان غدد لنفاوی با میزان صحت ۱۰۰٪ محاسبه می‌شود. مجموعه قواعد تعیین شده به صورت زیر می‌باشد.

$R1: \text{if } (gene_{21502} \leq 0.584) \text{ then class 1}$

$R2: \text{if } (gene_{21502} > 0.584) \text{ then class 2}$

۳- درخت تصمیم‌گیری برای پایگاه داده سرطان خون

نتیجه حاصل از بکارگیری الگوریتم درخت تصمیم‌گیری J48 برای شناسایی نوع سرطان خون در شکل (۴-۸) نمایش داده می‌شود.



شکل ۴-۸- درخت تصمیم‌گیری برای تشخیص نوع سرطان خون.

در شکل (۴-۸) کلاس یک مربوط به سرطان خون حاد در مغز استخوان و کلاس دو مربوط به انحرافات کروموزی در لایه زیرین استخوان می‌باشد. مجموعه قواعد بدست آمده به صورت زیر می‌باشد.

R1: if (gene₃₀₃₉₃ > 0.209) then class 1

R2: if (gene₃₀₃₉₃ ≤ 0.209) and (gene₄₁₁ ≤ 0.355) then class 2

R3: if (gene₃₀₃₉₃ ≤ 0.209) and (gene₄₁₁ > 0.355) then class 1

با استفاده از دو ژن و سه قانون، دو نوع سرطان خون با میزان صحت ۹۴/۱۱٪ شناسایی می‌شوند.

نتایج نشان می‌دهد که سیستم شناسایی کننده تنها با استفاده از چند ژن قادر به دسته‌بندی نوع سرطان با دقت بالایی در هر مجموعه پایگاه داده می‌باشد.

۴-۴ مقایسه سیستم پیشنهادی با نتایج دیگر روش‌ها

سه نرم‌افزار GenePix, ScanAlyze و ImaGene در بخش (۲-۲) معرفی گردید. همان‌طور که ذکر شد این سه نرم‌افزار در زمینه‌ی کمی‌سازی تصاویر ریزآرایه بکار گرفته می‌شوند. به منظور بررسی عملکرد سیستم در زمینه‌ی کمی‌سازی تصاویر ریزآرایه، این سه نرم‌افزار به صورت جامع با سیستم پیشنهادی مقایسه می‌شوند. از مهم‌ترین معایب این نرم‌افزارها نیمه خودکار بودن آنها است. در ادامه برای کمی‌سازی تصویر، هر مرحله در نرم‌افزار با سیستم پیشنهادی مقایسه می‌گردد. برای دسته‌بندی داده‌های ریزآرایه روش پیشنهادی J48 با مابقی روش‌های درخت تصمیم مقایسه می‌شود.

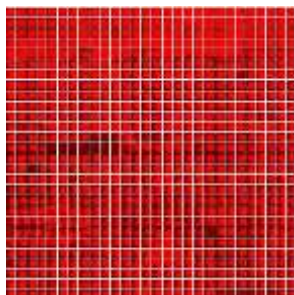
۴-۴-۱ تشخیص و تصحیح زاویه چرخش

نرم‌افزارها قابلیت تشخیص میزان زاویه چرخش و تصحیح زاویه به صورت خودکار ندارد. بدین منظور کاربر ملزم است تا زاویه چرخش تصویر را به روش صحیح و خطا به صورت تقریبی وارد سیستم نماید. سیستم با توجه به زاویه‌ی دریافتی، جهت تصویر ریزآرایه را تغییر می‌دهد. این درحالتی اتفاق می‌افتد که تصویر دارای تنها یک بلوک باشد، در صورتی که تصویر ریزآرایه متشکل از چندین بلوک باشد، نرم‌افزار قابلیت چرخاندن تصویر را به صورت کامل ندارد. در این قسمت نیز کاربر موظف می‌باشد میزان زاویه چرخشی هر بلوک را به صورت مجزا وارد سیستم نماید تا چرخش بلوک‌ها تصحیح گردد. در تصاویری که تعداد بلوک‌ها بالاست این امر می‌تواند بسیار وقت گیر باشد. بعلاوه زاویه چرخش محدودی برای تصاویر ریزآرایه تعریف شده است. به طور مثال محدوده زاویه چرخش در نرم‌افزار GenePix مابین ۲۰ تا ۲۰- تعریف می‌شود. بنابراین اگر بلوک بیشتر از ۲۰ درجه چرخش داشته باشد، زاویه‌ی چرخش قابل تصحیح نخواهد بود. در سیستم پیشنهادی با استفاده از تبدیلات رادن میزان چرخش تصویر در هر دو جهت عقربه‌ی ساعت به صورت تمام خودکار قابل شناسایی و تصحیح می‌باشد. همچنین محدودی چرخش تصاویر تا زاویه ۱۸۰ درجه نیز قابل تصحیح خواهد بود.

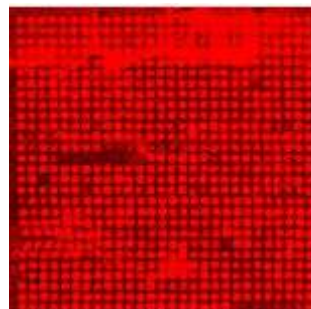
۴-۴-۲ بلوک‌بندی و شبکه‌بندی

در هر سه نرم‌افزار برای شناسایی موقعیت بلوک‌ها و خال‌ها، کاربر باید از تعداد دقیق آن‌ها در هر سطر و ستون آگاهی داشته باشد. علاوه بر این، با توجه با این حقیقت که نرم‌افزار فاصله مابین خال‌ها و بلوک‌ها را برای کاربر اندازه‌گیری نمی‌کند، این فواصل باید به صورت دستی از طریق کاربر وارد سیستم شوند. این قسمت از نرم‌افزارها زمانی به درستی عمل می‌کند که تصاویر دارای خرابی یا به عبارت دیگر نویز پایینی داشته باشند. یکی از معایب بزرگ این نرم‌افزارها، اجرای عملیات بلوک‌بندی و شبکه‌بندی به صورت مجزا در دو کانال رنگی می‌باشد. در حالی که تمام موقعیت‌های بلوک و خال‌ها در هر دو کانال با هم برابر می‌باشد، بنابراین بهتر است برای افزایش دقت این مراحل دو کانال رنگی را با هم ادغام، سپس الگوریتم مربوطه اجرا گردد.

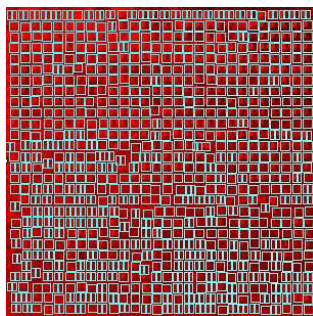
در سیستم ارائه شده، عملیات بلوک‌بندی و شبکه‌بندی با استفاده از روش آستانه‌گذاری تصویر بدون دخالت کاربر به صورت تمام خودکار با دقت بالایی ارائه می‌شود. علاوه بر این در روش پیشنهاد شده به دلیل اعمال پیش‌پردازش برای حذف نویز، در تصاویری که دارای خرابی همانند نویز مشعل باشند به خوبی موقعیت تمامی خال‌ها شناسایی می‌گردد. در شکل (۴-۹) عملیات شبکه‌بندی بر روی یک تصویر ریزآرایه همراه با نویز مشعل به منظور مقایسه با روش پیشنهادی با استفاده از نرم‌افزار GenePix و روش [۴۵] انجام گردیده است. همانطور که در این شکل مشاهده می‌شود، عملیات شبکه‌بندی در نرم‌افزار و روش معرفی شده در برابر نویز بسیار بد عمل می‌کند، در حالی که روش پیشنهادی به خوبی می‌تواند تمامی موقعیت خال‌ها در تصویر را شناسایی کند. هر شاخص مستطیلی در نرم‌افزار GenePix، بیان‌گر موقعیت یک خال بوده و هر خط در روش [۴۵] خط جداکننده خال‌ها است که توسط ماشین بردار پشتیبان رسم شده است.



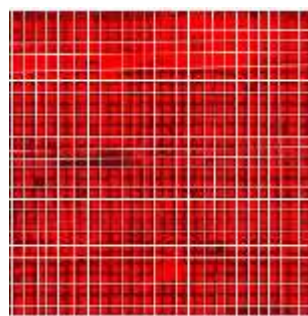
(ب)



(الف)



(د)



(ج)

شکل ۴-۹- مقایسه عملیات شبکه‌بندی با استفاده از نرم‌افزار GenePix با روش پیشنهادی. (الف) نمایش تصویر در لایه قرمز به همراه نویز مشعل. (ب) اعمال شبکه‌بندی با استفاده از روش پیشنهادی. (ج) اعمال روش شبکه‌بندی با استفاده از روش [۴۵]. (د) اعمال شبکه‌بندی با استفاده از نرم‌افزار GenePix.

۳-۴-۴ قطعه‌بندی و تصحیح پس‌زمینه

به منظور قطعه‌بندی، هر سه نرم‌افزار بر اساس دایره‌ی افقی، دایره‌ی ثابت، شکل افقی و هیستوگرام عمل می‌کنند. هر روش در شرایط متفاوت دارای معایب و مزایای خاص خود می‌باشند. برای مثال در روش دایره‌ی ثابت اگر خال‌ها با اندازه‌های یکسان بر روی تصویر قرار نگیرند این روش بسیار ضعیف عمل می‌کند. در روش دایره‌ی افقی اگر کاربر دامنه‌ی شعاع خال‌ها را به درستی وارد سیستم نکند، ممکن است بسیاری از خال‌ها شناسایی نشوند. همچنین در این نرم‌افزارها خال‌هایی که با هم هم‌پوشانی دارند یا شعاع کمتر از پنج پیکسل دارند، قابل تشخیص نمی‌باشند. در روش شکل افقی نیز تعداد و محل شروع نقاط بسیار مهم است. در نرم‌افزارها نقاط شروع برای هر خال توسط کاربر و به صورت دستی انتخاب می‌شود. بعد از قطعه‌بندی، مقدار پس‌زمینه به روش‌های محلی، پس‌زمینه‌ی ثابت و ریخت‌شناسی تعیین می‌شوند. با حذف پس‌زمینه از خال‌ها، مقدار واقعی هر خال که بیانگر میزان هیبرید شدن هر ژن می‌باشد، تخمین زده می‌شود. در روش محلی ممکن است گاهی بدلیل وجود نویز مقدار میانگین شدت روشنایی پیکسل‌ها در منطقه محلی انتخابی، بیشتر از مقدار خال‌ها باشد. بنابراین هنگام تصحیح پس‌زمینه، ممکن است پیکسل‌های مربوط به خال مقادیر منفی به خود بگیرند. این تخمین اشتباه، منجر به ایجاد مشکلاتی در ادامه کار می‌شود. در روش پس‌زمینه‌ی ثابت سه درصد از مقدار شدت روشنایی تمامی خال‌ها به عنوان پس‌زمینه در نظر گرفته می‌شود. از آنجایی که تمامی تصاویر ریزآرایه به صورت ناهمگون هستند، این روش به ندرت در برنامه‌های کاربردی واقعی بکار برده می‌شود و دارای عملکرد نامناسبی می‌باشد. در روش ریخت‌شناسی با استفاده از عملگرهای سایش و سپس گسترش میزان پس‌زمینه تخمین زده می‌شود. عملکرد این روش نسبت به دو روش قبلی در برابر نویزهای موجود در تصویر و ناهمگنی آن برتری دارد. مهم‌ترین معایب این روش شامل، هموارسازی لبه‌ها در پس‌زمینه و اثر بیش‌گشایش یا بیش‌فرسایش به صورت محلی بر روی خال‌ها می‌باشد. سطح این اثر بستگی به اندازه پنجره عملگرهای ریخت‌شناسی دارد. هرچه اندازه این پنجره کوچکتر باشد، این اثرات نیز کمتر می‌شوند. اما در برنامه‌های کاربردی باید اندازه پنجره بزرگتر از خال‌ها باشد.

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

(الف)

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

(ب)

0	0	0	0	0	0	1	0	0	0	0	0	0
0	0	2	0	0	0	0	0	1	0	0	1	0
0	0	0	31	157	124	149	152	85	0	0	0	0
0	0	21	137	137	128	108	165	134	97	0	2	0
0	2	137	148	134	131	147	145	142	160	0	0	1
0	0	170	155	118	149	137	122	132	163	116	0	0
0	0	165	138	145	125	150	160	127	138	115	0	0
0	0	88	141	153	155	133	143	154	144	164	1	0
0	0	120	204	125	158	126	140	137	160	0	0	0
0	0	0	0	88	124	81	212	116	2	0	0	0
0	0	0	0	0	1	1	0	0	0	0	0	0
1	2	0	0	0	1	0	1	0	0	0	0	0

(ج)

5	5	3	4	5	3	7	4	5	4	4	5	0
4	1	8	5	4	4	0	5	7	0	0	7	4
1	2	0	37	163	130	155	158	91	1	0	5	4
4	3	28	143	143	134	114	171	140	103	3	8	1
2	9	144	155	140	137	153	151	148	166	6	0	7
5	1	177	162	125	156	143	128	139	170	123	1	3
4	6	172	145	152	132	157	167	134	145	122	5	2
0	5	95	148	160	162	140	150	161	151	171	8	3
3	4	127	211	132	165	133	147	144	167	6	2	6
5	0	5	0	95	131	88	219	123	9	0	1	2
6	7	5	4	4	8	8	7	2	3	7	4	3
8	9	4	4	1	8	3	8	2	6	6	7	5

(د)

شکل ۴-۱۰- مقایسه‌ی روش پیشنهادی برای تخمین میزان پس‌زمینه از روش ریخت‌شناسی، (الف) ماتریس واقعی
 خال بدون احتساب پس‌زمینه، (ب) ماتریس خال به همراه پس‌زمینه ناهمگون، (ج) تخمین و حذف پس‌زمینه به روش
 (د) تخمین و حذف پس‌زمینه به روش اعمال عملگرهای ریخت‌شناسی. (TV+L1)

در روش پیشنهادی با اجرای الگوریتم $TV+L1$ ، تخمین پس‌زمینه بر روی هر یک از خال در سلول مشخص خود به صورت مجزا صورت می‌گیرد. با حذف پس‌زمینه مقدار واقعی هر خال از تصویر مشخص می‌شود. بنابراین نیازی به اجرای دو الگوریتم متفاوت قطعه‌بندی و تخمین پس‌زمینه نیست. تمامی روش‌های مذکور تخمین پس‌زمینه، بر روی یک خال مصنوعی همراه با پس‌زمینه ناهمگون (شکل ۴-۱۰-ب) اعمال شده و میانگین خطاهای هر کدام در جدول (۴-۷) محاسبه شده است. برای درک بهتر برتری روش پیشنهادی، خروجی الگوریتم ریخت‌شناسی و $TV+L1$ در شکل (۴-۱۰) نمایش داده می‌شود.

جدول ۴-۷- 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

با توجه به شکل (۴-۱۰) و جدول (۴-۷) روش پیشنهادی نسبت به دیگر روش‌های موجود در نرم‌افزارها، مقدار پس‌زمینه را بهتر تخمین زده و همواره دقت بالاتری دارد. این روش در هیچ‌کدام از نرم‌افزارهای مذکور به کار گرفته نشده است.

۴-۴-۴ محاسبه شدت روشنایی خال‌ها

در نرم‌افزارهای موجود پس از استخراج خال‌ها، برای هر کدام مقدار میانگین، میانه و مجموع شدت سطح خاکستری پیکسل‌ها محاسبه می‌شود. فرد خبره با توجه به نیاز خود اطلاعات مربوط را در فرم متنی ذخیره می‌کند. همانطور که قبلاً ذکر شد، ریزآرایه‌ها از ابتدا تا تولید داده‌های خام در مراحل مختلفی در معرض تغییرات نامطلوب قرار می‌گیرند. نرمال‌سازی موجب اصلاح تفاوت و منظم‌سازی میزان شدت روشنایی بین خال‌ها می‌شود. نرم‌افزارهایی که با استفاده از روش Lowess و میانگین‌گیری سراسری داده‌های موجود را نرمال می‌کنند، در سیستم پیشنهادی تنها میانگین خال‌ها از تصاویر ریزآرایه استخراج می‌شوند و با کمک الگوریتم Lowess داده‌های استخراج شده نرمال می‌شوند. این

نرم افزارها قابلیت تحلیل و طبقه بندی چندین تصاویر ریزآرایه را ندارند. از آن ها فقط برای کمی سازی تصاویر ریزآرایه استفاده می شود. بنابراین دو گام، داده کاوی به منظور کاهش ابعاد داده و یادگیری ماشین به منظور دسته بندی تصاویر در این نرم افزارها تعریف نشده است.

۴-۴-۵ طبقه بندی داده های ریزآرایه

در این پژوهش فرد خبره نیازمند به دسترسی مراحل کار و تولید قوانین قابل فهم می باشد. در بین روش های موجود برای دسته بندی داده های ریزآرایه، درخت تصمیم گیری مناسب ترین گزینه می باشد. توصیف شرایط در درخت تصمیم به آسانی با منطق بولی امکان پذیر می باشد و شرایط رسیدن به نتایج قابل درک هستند. به این الگوریتم ها که مراحل رسیدن به نتیجه قابل مشاهده است مدل جعبه سفید^۱ می نامند. در حالی که بیشتر الگوریتم ها بدلیل پیچیدگی در توصیف، نتایج آنها همانند یک جعبه سیاه^۲ می باشند. به منظور مقایسه سیستم شناسایی بیماری ارائه شده با دیگر روش ها و درختان تصمیم گیری، میزان دقت الگوریتم ها برای هر سه نوع سرطان در جدول (۴-۸) محاسبه و گزارش شده است.

جدول ۴-۸- مقایسه عملکرد الگوریتم های درخت تصمیم گیری برای تشخیص نوع تومور.

نوع درخت تصمیم گیری			دقت دسته بندها برای تشخیص نوع سرطان
	سرطان خون	سرطان غدد لنفاوی	سرطان سینه
J48 Tree	٪ ۹۴٫۱۱	٪ ۱۰۰	٪ ۹۵٫۴۵
Decision Stump	٪ ۹۴٫۱۱	٪ ۹۲٫۸۵	٪ ۶۶
LMT	٪ ۷۰٫۵۸	٪ ۱۰۰	٪ ۹۰٫۴۷
Random Tree	٪ ۹۴٫۱۱	٪ ۱۰۰	٪ ۸۰٫۹۵
Rep Tree	٪ ۸۸٫۲۳	٪ ۹۲٫۸۵	٪ ۸۵٫۷۱
Hybrid method in [۵۳]	٪ ۸۵٫۷۱	٪ ۱۰۰	٪ ۹۰٫۴۷
SMO [۵۱]	٪ ۸۸٫۲۳	٪ ۸۵٫۷۱	٪ ۸۰٫۹۵

^۱ White Box

^۲ Black Box

در جدول (۴-۸) واضح است که کارآیی روش پیشنهادی یعنی درخت J48، برای تمامی مجموعه داده-های موجود بالاتر می‌باشد. این نتایج، علی‌رغم سادگی روش پیشنهادی، بیانگر مطلوبیت نسبی روش ارائه شده در مقایسه با مابقی روش‌ها نیز است. این سیستم به همراه قوانین استخراج شده، می‌تواند ابزار کمکی برای تحلیل و تشخیص بیماری باشد و در درمان بسیاری از بیماری‌ها پزشکان را یاری دهد.

۴-۵ نتیجه‌گیری

در این فصل مراحل مختلف سیستم پیشنهادی ذکر گردید. در این فصل، نتایج حاصل از اعمال سیستم پیشنهادی در مراحل مختلف بر روی تصاویر و داده‌های ریزآرایه‌ی موجود در پایگاه داده، بیان گردید. سپس در مرحله‌ی کمی‌سازی تصاویر ریزآرایه، نتایج سیستم پیشنهادی با نتایج دیگر نرم‌افزارها مقایسه گردید. در انتها مقایسه‌ای برای دقت پیش‌بینی دسته‌بندی داده‌های ریزآرایه در سیستم پیشنهادی با دیگر روش‌های درخت تصمیم صورت گرفته شده است. با توجه به نتایج ذکر شده، سیستم ارائه شده یک سیستم کامل‌تر و دقیق‌تر نسبت به مابقی سیستم‌های ارائه شده در سال‌های اخیر است. همان‌طور که بیان شد، سیستم پیشنهادی یک سیستم تمام خودکار برای تحلیل تصاویر ریزآرایه می‌باشد.

فصل پنجم

نتیجہ گیری و پیشہدات آتی

۵-۱ مقدمه

در این پایان‌نامه سیستمی جدید به منظور استخراج داده‌ها از تصاویر ریزآرایه و طبقه‌بندی داده‌های خام، مربوط به سه نوع سرطان ارائه گردید. این سیستم طی چندین مرحله شامل: شناسایی میزان چرخش تصویر، تعیین موقعیت هر بلوک در تصویر، تعیین محل دقیق خال‌ها، تخمین پس‌زمینه تصویر، محاسبه میزان بیان ژن‌ها، نرمال‌سازی داده‌ها، انتخاب ویژگی و در نهایت دسته‌بندی نمونه‌ها، قادر به تشخیص نوع سرطان در پایگاه داده موجود می‌باشد. در ادامه به صورت مختصر نتایج سیستم به همراه نقاط ضعف و قوت آن مورد بررسی قرار گرفته و برای بهبود نتایج راه‌حل‌هایی پیشنهاد می‌شود.

۵-۲ نتیجه‌گیری

در گذشته بررسی همزمان چند ژن کار بسیار مشکلی بوده است. اما در دهه‌های اخیر با ورود فناوری ریزآرایه امکان بررسی بیان هزاران ژن به طور همزمان فراهم شده است. این فناوری متشکل از صدها یا هزاران ژن در یک سطح کوچک به نام خال می‌باشد. یکی از مهم‌ترین مشکلات پیش‌رو در این فناوری، انبوه داده‌ها و اطلاعات به دست آمده از داده‌های مولکولی و ژنتیکی است که لزوم استفاده از روش‌های آماری و تکنولوژی اطلاعات و همچنین یادگیری ماشین را متذکر می‌شود. از آنجایی که برخی از کارخانه‌ها، سالانه ده‌ها هزار تصویر ریزآرایه با ابعاد مختلف تولید کنند، پردازش خودکار و بلادرنگ این تصاویر و تحلیل داده‌های ریزآرایه از اهمیت بالایی برخوردار می‌باشد.

در فصل اول این پایان‌نامه، فناوری ریزآرایه به صورت کامل معرفی شد. این فناوری متشکل از دو بخش آزمایشگاهی و تحلیل تصاویر ریزآرایه می‌باشد. تصاویر حاصل از آزمون ریزآرایه در سه گام پردازش تصویر، داده‌کاوی و یادگیری ماشین مورد بررسی قرار می‌گیرد. در گام اول الگوریتم‌ها و تکنیک‌های پردازش تصویر به منظور کمی‌سازی تصاویر ریزآرایه یا بعبارت دیگر تعیین میزان بیان هر ژن استفاده می‌شود. پس از استخراج داده‌ها از تصویر ریزآرایه، به دلیل وجود تعداد ژن‌های زیاد، حجم بالایی از اطلاعات در اختیار کاربر قرار می‌گیرد. بنابراین با کمک تکنیک‌های داده‌کاوی حجم داده‌ها را کاهش و با بکارگیری یادگیری ماشین تمامی داده‌های ریزآرایه تحلیل و دسته‌بندی می‌شوند. در فصل دوم به

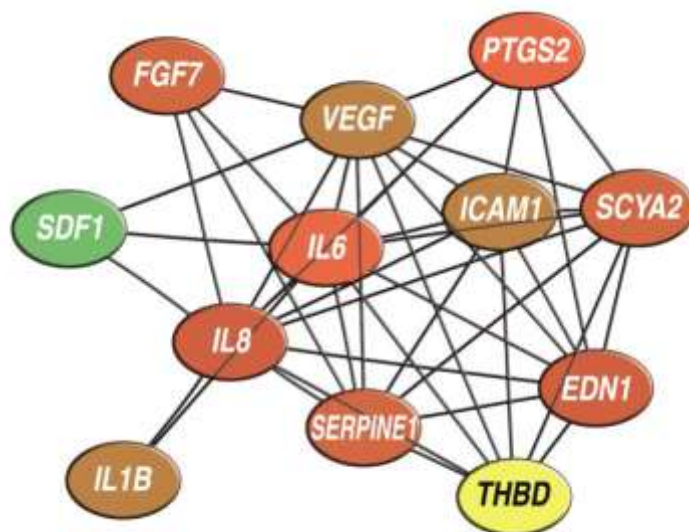
توضیح و بیان نرم افزارها و تحقیقات پیشین برای تحلیل تصاویر ریزآرایه پرداخته شد. همانطور که بیان شد، سیستم تمام خودکار برای دسته‌بندی داده‌های ریزآرایه در اختیار نیست. بنابراین، هدف این پایان-نامه ارائه‌ی سیستمی است که بتواند به صورت خودکار تصاویر ریزآرایه را کمی‌سازی و دسته‌بندی کند. در این پژوهش یک سیستم تمام خودکار برای تحلیل تصاویر و داده‌های ریزآرایه ارائه گردیده است. سیستم پیشنهادی، برای تعیین محل دقیق بلوک‌ها و 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 استفاده از روش تبدیلات رادن با میانگین دقت $98/84\%$ تصحیح می‌گردد، سپس به کمک تکنیک‌های آستانه‌گذاری محل دقیق هر بلوک با دقت 100% شناسایی شده‌اند. در هر تصویر ریزآرایه 48 بلوک به صورت ماتریس 4×12 در کنار هم قرار گرفته‌اند. برای شبکه‌بندی هر بلوک و تعیین موقعیت خال‌ها، ابتدا تأثیرات خرابی‌های ذکر شده توسط عملگرهای ریخت‌شناسی کاهش می‌گردد، سپس به کمک تکنیک‌های آستانه‌گذاری همانند اتسو موقعیت هر خال به صورت دقیق در یک چهارچوب سلولی قرار می‌گیرد. فقط تعداد کمی از تصاویری که شامل نویز و نقص قابل توجهی هستند دقت آن‌ها به طور متوسط به 98% کاهش می‌یابد. بعد از تعیین محل دقیق خال‌ها میزان پس‌زمینه‌ی هر کدام به صورت مجزا با اجرای الگوریتم $TV+L1$ تخمین زده و از تصویر اصلی‌شان کاسته می‌شود. مقدار حاصل نشان-دهنده‌ی میزان هیبرید شدن ژن‌ها می‌باشد. با تعیین میزان شدت روشنایی هر خال و مقایسه ژن‌ها در هر دو کانال رنگی تصویر، بیان ژن محاسبه می‌گردد. داده‌های خام بدست آمده به همراه شماره‌ی خال، شماره بلوک در یک فرم اکسل ذخیره می‌شود. سطرها و ستون‌ها در این فرم به ترتیب نماینده‌ی بیان ژن‌ها و نمونه‌های افراد بیمار می‌باشد. از آنجایی که داده‌ها در زمان آزمایش تا استخراج اطلاعات خام دست خوش تغییراتی می‌باشد، بنابراین برای منظم‌سازی ویژگی‌ها و کاهش پراکندگی، الگوریتم نرمال-سازی Lowess بر روی آن‌ها اعمال می‌شود. برای انجام دسته‌بندی نوع سرطان، از بین ژن‌های موجود 100 ژن مؤثرتر جهت تشخیص سرطان مشخص می‌گردد. جهت تشخیص این ژن‌های مؤثرتر از روش

بهره اطلاعاتی بر روی مجموعه تصاویر ریزآرایه موجود (آموزش) استفاده شده است. در نهایت به کمک درخت تصمیم‌گیری J48 ژن‌های برتر مورد تحلیل قرار می‌گیرند. میزان صحت طبقه‌بندی داده‌های سرطان سینه، سرطان غدد لنفاوی و سرطان خون به ترتیب با تعداد قانون چهار، سه و دو به همراه دقت ۹۵/۴۵٪، ۱۰۰٪ و ۹۴/۱۱٪ تخمین زده شده است.

بررسی‌ها نشان می‌دهد که با استفاده از سیستم پیشنهادی تصاویر ریزآرایه با دقت بالایی تحلیل و دسته‌بندی می‌شوند. مقایسه این سیستم با سایر نرم‌افزارها عرضه شده، بهبود نتایج حاصل از سیستم پیشنهادی را نشان می‌دهد.

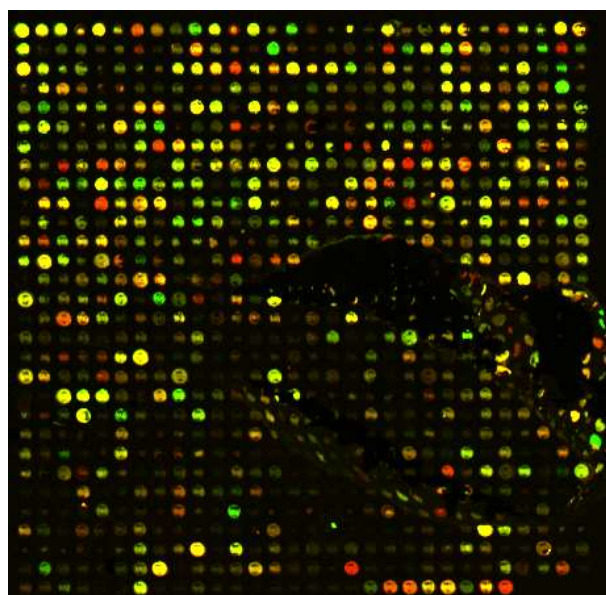
۳-۵ پیشنهاداتی

مهم‌ترین گام برای تحلیل تصاویر ریزآرایه استخراج مقادیر صحیح خال‌ها می‌باشد. زیرا با تشخیص صحیح ژن‌های مسبب بیماری، نوع داروی تجویزی برای بیمار تعیین می‌شود. در برخی آزمایش‌ها تصاویر ریزآرایه شامل نواقص مختلفی از جمله حذف برخی از خال‌ها در حین آزمایش می‌باشند. با توجه به مطالعات صورت گرفته، اکثر ژن‌ها با هم روابط معنادار بیولوژیکی دارند و تمامی آن‌ها در چندین گروه جا داده می‌شوند [۹۳، ۹۴]. با کشف نوع روابط مابین این ژن‌ها قادر به محاسبه‌ی خال‌های از بین رفته خواهیم بود. یکی از مهم‌ترین مشکلاتی که در این زمینه رخ خواهد داد، پیچیدگی بالای محاسباتی است. چراکه تعداد ژن‌ها به صورت میانگین در یک تصویر کامل ریزآرایه تقریباً ۴۵۰۰۰ می‌باشد. کشف روابط مابین آن‌ها کاری سخت، اما شدنی است. در شکل (۵-۱) یک نمونه کوچکی از این روابط مابین ژن‌ها نمایش داده شده است. این ژن‌ها در یک گروه قرار گرفته و به عنوان ژن‌های همکار نامیده می‌شوند. با توجه به میزان فعالیت هر ژن یا به عبارت دیگر مقدار بیان ژن‌ها، هر کدام با رنگ خاصی نمایش داده می‌شوند. رنگ زرد نشان دهنده‌ی عدم فعالیت ژن، رنگ قرمز نشان دهنده‌ی فعالیت زیاد ژن بیمار و رنگ سبز نشان دهنده‌ی فعالیت ژن سالم در مقایسه دو کانال تصویر (نمونه سالم و مریض) می‌باشد.



شکل ۵-۱- روابط مابین ژن‌ها در یک تصویر ریزآرایه- رنگ هر ژن نشان دهنده میزان فعالیت آن ژن یا به عبارتی بیان ژن را در دو کانال رنگی نمایش می‌دهد.

یک نمونه از خرابی‌های ذکر شده در شکل (۵-۲) نمایش داده شده است. در کارهای آتی سعی می‌شود که با کشف روابط مابین ژن‌ها، خال‌هایی که در اثر خرابی تصویر از بین رفته‌اند، کشف و مقادیر آنها تخمین زده شوند.



شکل ۵-۲- نمونه‌ای از خرابی‌های رایج در تصویر ریزآرایه.

همچنین در هر آزمایش تحقیقاتی حجم زیادی از این تصاویر تولید می‌شود. تعداد زیاد این تصاویر، حجم بزرگ آن‌ها و نیاز به نگهداری و ذخیره‌سازی این تصاویر برای مدت طولانی فشرده‌سازی آن‌ها را ضروری می‌کند، بطوری که فشرده‌سازی تصاویر به صورت بلادرنگ از جمله گلوگاه‌های این تکنولوژی می‌باشد. در اکثر نرم‌افزارهای تحلیل تصاویر ریزآرایه از جمله سیستم پیشنهادی در این پایان‌نامه سرعت اجرای سیستم‌ها پایین می‌باشد. بنابراین بهتر است حجم تصاویر با استفاده الگوریتم‌های فشرده‌سازی بدون اتلاف تصاویر کاهش، سپس به عنوان تصویر ورودی وارد سیستم گردند. با اضافه کردن این دو گام به سیستم پیشنهادی، کارایی سیستم افزایش و همچنین یک سیستم کاملاً قابل اعتماد برای تشخیص بیماری‌ها و تحلیل داده‌های ریزآرایه به بازار عرضه خواهد شد.

پوست

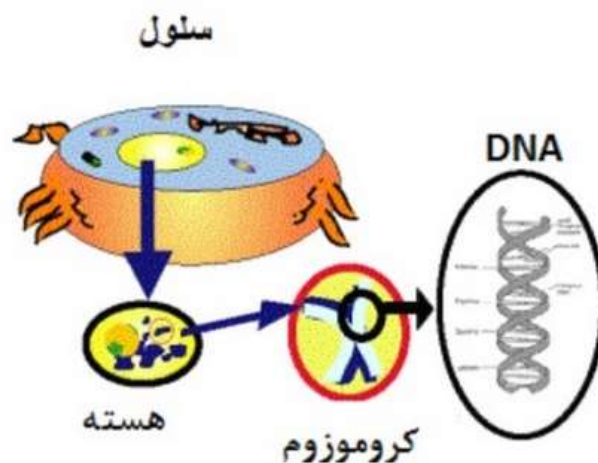
دی‌ان‌ای یا اسید دئوکسی ریبونوکلئیک چیست؟

ماده وراثتی در هسته سلول به شکل کروموزوم سازماندهی شده است. کروموزوم‌ها از کروماتین حاصل می‌شوند و کروماتین از دی‌ان‌ای و آران‌ای ساخته شده است. در ساختار هر مولکول دی‌ان‌ای و آران‌ای، واحدهایی به نام نوکلئوتید وجود دارد که هر کدام از آن‌ها از یک ترکیب فسفردار، نوعی قند و یکی از چهار نوع باز آلی به نام آدنین، تیمین، گوانین یا سیتوزین و در مولکول آران‌ای به جای تیمین باز یوراسیل، که آن‌ها را با علامت‌های A، G، T، C و U شناسایی می‌کنند، تشکیل می‌شوند. دی‌ان‌ای، مولکول بسیار طولانی است که دو رشته دارد.

این دو رشته مانند یک زیپ به هم متصل شده‌اند. در هر رشته، هزاران نوکلئوتید وجود دارد. ترتیب قرار گرفتن نوکلئوتیدها در ساختار دی‌ان‌ای طوری است که همیشه A در مقابل T با دو پیوند هیدروژنی و C به G با سه پیوند هیدروژنی قرار می‌گیرد. اطلاعات زیادی که در دی‌ان‌ای وجود دارد، در واحدهایی به نام "ژن" ذخیره شده است. شکل هر ژن قسمتی از مولکول دی‌ان‌ای است که برای ساختن پروتئین یا آران‌ای مورد استفاده قرار می‌گیرد. یک مولکول دی‌ان‌ای هزاران ژن دارد که مانند واگن‌های یک قطار به دنبال هم قرار گرفته‌اند (شکل ۱).

طول نوار دی‌ان‌ای می‌تواند بیش از چند میلیون مولکول باشد. میلیاردها حالت برای نحوه چیدمان مولکول‌ها در این نوار دی‌ان‌ای وجود دارد. کثرت انواع چیدمان در مولکول دی‌ان‌ای، باعث کثرت تنوع خلق و خو، اندام، رنگ پوست و مو، قد در انسانها و سایر جانداران می‌شود.

اینکه افراد نسبت به داروها و محرک‌ها واکنش‌های متفاوتی نشان می‌دهند، به دلیل تفاوت چیدمان مولکول دی‌ان‌ای در هسته سلول بدن ایشان است. چیدمان مولکول دی‌ان‌ای در هسته سلول بدن دو انسان تا حدود ۹۰ درصد شبیه به هم است، ساختار و وظیفه سلول‌ها تفاوت عمده‌ای ندارند، بلکه رفتار سلول‌ها در شرایط مختلف تغییر می‌کند. مولکول دی‌ان‌ای، به پروتئین چسبیده و تشکیل کروموزوم را می‌دهد (شکل ۱).

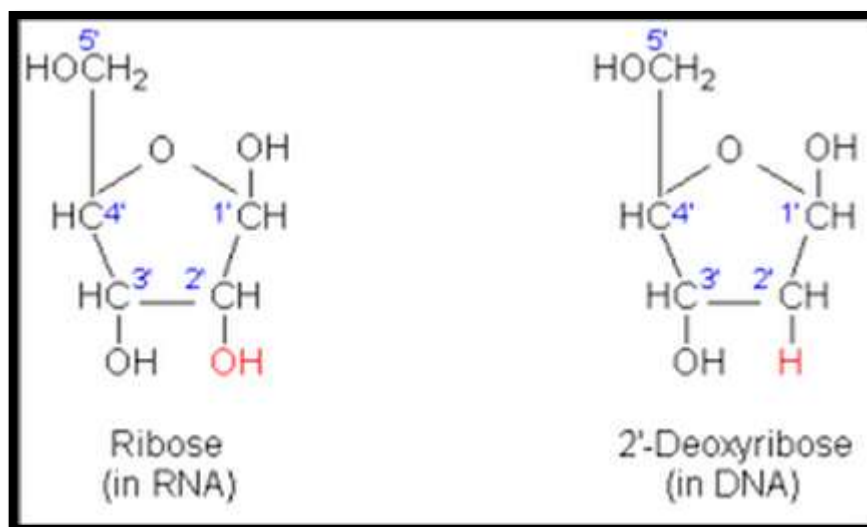


شکل ۱- ساختار دی‌ان‌ای و کروموزوم.

بدن انسان، بیش از ۳۷ تریلیون (۳۷۰۰۰،۰۰۰،۰۰۰،۰۰۰) سلول دارد، در هسته هر یک از این سلول‌ها (سلول پوست، مو، ناخن، دندان)، ۴۶ کروموزوم (۲۳ جفت کروموزوم) وجود دارد، هر انسانی نیمی از کروموزوم‌های خود را از مادر و نیمی دیگر را از پدر دریافت می‌کند. ترکیب چیدمان دی‌ان‌ای موجود در کروموزوم‌ها، منحصر به فرد است و فقط در دوقلوهای همسان کاملاً شبیه به هم است.

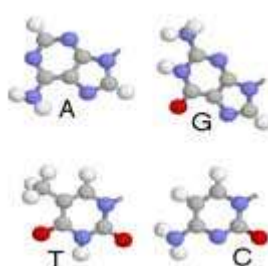
تاریخچه کشف دی‌ان‌ای

در اواخر قرن نوزدهم یک بیوشیمیست آلمانی نشان داد که اسیدهای نوکلئیک (مولکول‌های زنجیری بلند که از واحدهای ساختمانی کوچک تری به نام "نوکلئوتید" تشکیل شده‌اند). دارای قند، اسید فسفریک و چند باز نیتروژن دار می‌باشند. اندکی بعد مشخص شد که قند موجود در اسیدهای نوکلئیک می‌تواند ریبوز یا دئوکسی ریبوز باشد و لذا اسیدهای نوکلئیک به دو دسته (DNA) DeoxyriboNucleic Acid که قند موجود در آن‌ها دئوکسی ریبوز است و (RNA) RiboNucleic Acid که قند موجود در آن‌ها ریبوز است، تقسیم می‌شوند (شکل ۲).



شکل ۲- ساختار دی‌ان‌ای و آر‌ان‌ای

در سال ۱۹۴۸ لینوس پاولینگ کشف کرد که بسیاری از مولکول‌های پروتئینی به شکل یک مارپیچ (helix) هستند، و تقریباً شکلی شبیه فنر دارند. در سال ۱۹۵۰ نیز اروین شارگاف نشان داد که اگرچه آرایش بازهای موجود در ساختار دی‌ان‌ای بسیار متنوع است، اما همواره نسبت باز آدنین (A) و باز تیمین (T) موجود در آن با هم برابر است و همین طور نسبت باز سیتوزین (C) با باز گوانین (G). این دو اکتشاف نقش مهمی را در آشکار شدن ساختمان مولکول دی‌ان‌ای ایفا نمود. ساختار عناصر آدنین، تیمین، سیتوزین و گوانین در شکل (۳) نمایش داده شده است.



شکل ۳- ساختار عناصر آدنین، تیمین، سیتوزین و گوانین

در دهه‌ی ۱۹۵۰ همچنان رقابت برای کشف ساختار دی‌ان‌ای ادامه داشت. در دانشگاه کمبریج کریک واتسون تحت تأثیر کارهای پاولینگ سعی داشتند تا با ارائه‌ی مدل‌های فیزیکی، ساختارهای احتمالی ممکن برای دی‌ان‌ای را محدود کنند تا سرانجام به ساختار صحیح دست یابند. گروه دیگری متشکل از

ویلکینز و فرانکلین نیز در کالج کینگ لندن به طور همزمان مشغول مطالعه دی‌ان‌ای بود. روش کار این گروه با گروه قبلی متفاوت بود. آن‌ها سعی داشتند تا با روش آزمایشگاهی به ویژه با استفاده از تصاویر پراش اشعه X از مولکول دی‌ان‌ای، ساختار آن را معین کنند.

در سال ۱۹۵۱، فرانکلین دریافت که دی‌ان‌ای با توجه به میزان رطوبت هوای محیط، می‌تواند دو شکل متفاوت داشته باشد و بنابراین نتیجه‌گیری کرد که بخش فسفات مولکول در سمت خارجی آن قرار دارد. اندکی بعد او با استفاده از تصاویر اشعه X فهمید که دی‌ان‌ای در حالت "مرطوب (Wet)" از تمامی ویژگی‌های یک مارپیچ (helix) برخوردار است؛ این احتمال که حالت دیگر مولکول دی‌ان‌ای نیز به شکل مارپیچی باشد به ذهن او خطور کرد، اما نمی‌خواست تا زمانی که شواهد قطعی برای این حدس پیدا کند آن را اعلام نماید. در ژانویه ۱۹۵۳ ویلکینز که از به نتیجه رسیدن تحقیقات ناامید شده بود، نتایج تحقیقات فرانکلین را بدون اطلاع و رضایت او، با واتسون در میان گذاشت.

واتسون و کریک با استفاده از این نتایج، مدلی بسیار شگفت‌انگیز را برای ساختار دی‌ان‌ای پیشنهاد نمودند. آن‌ها مولکول را به صورت دو زنجیر مارپیچی متشکل از نوکلئوتیدها تصور کردند که یکی از آن‌ها بالا می‌رفت و دیگری پایین می‌آمد. کریک که به تازگی یافته‌های شارگاف را هم مطالعه کرده بود، سعی کرد با استفاده از آن‌ها نحوه‌ی قرار گرفتن بازها را در مولکول دی‌ان‌ای مشخص کند. او اظهار کرد که بازها در میانه‌ی این مارپیچ دوتایی دو به دو به هم متصل می‌شوند تا فاصله بین دو مارپیچ ثابت بماند. آن‌ها ادعا کردند که هر یک از این دو مارپیچ مولکول دی‌ان‌ای می‌تواند به عنوان قالبی برای ایجاد دیگری استفاده شود.

در تقسیم سلولی، این دو رشته از هم جدا می‌شوند و بر روی هر یک از آن‌ها یک نمونه‌ی جدید شبیه رشته‌ی مقابل قبلی ساخته می‌شود. با این روش بدون اینکه ساختار دی‌ان‌ای عوض شود، یک دی‌ان‌ای شبیه آن تولید می‌شود. در اندک مواردی که در این روند خطایی پیش بیاید، شاهد "جهش" خواهیم بود. مدل آن‌ها چنان با اطلاعات حاصل از آزمایش‌ها مطابقت داشت که بلافاصله مورد قبول همه واقع شد. کشف ساختار دی‌ان‌ای را می‌توان مهم‌ترین اکتشاف زیستی در صد سال اخیر دانست. در سال

۱۹۶۲ واتسون، کریک و ویلکینز موفق به دریافت جایزه نوبل شدند، اما متأسفانه فرانکلین در گذشته

بود.

منابع

1. Cruz JA, Wishart DS (2006) Applications of machine learning in cancer prediction and prognosis. *Cancer informatics* 2 (1):59-77
2. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine* 7 (6):673-679
3. Hanai T, Hamada H, Okamoto M (2006) Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields. *Journal of bioscience and bioengineering* 101 (5):377-384
4. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics* 24 (3):227-235
5. International Human Genome Sequencing Consortium, Finishing the euchromatic sequence of the human genome (2004). *Nature* 431 (7011):931-945
6. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W (2001) Initial sequencing and analysis of the human genome. *Nature* 409 (6822):860-921
7. Sequencing TC, Consortium A (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437 (7055):69-87
8. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, Lozano JA, Armañanzas R, Santafé G, Pérez A (2006) Machine learning in bioinformatics. *Briefings in bioinformatics* 7 (1):86-112
9. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW (2005) Gene selection from microarray data for cancer classification—a machine learning approach. *Computational biology and chemistry* 29 (1):37-46
10. He M, Stoevesandt O, Palmer EA, Khan F, Ericsson O, Taussig MJ (2008) Printing protein arrays from DNA arrays. *Nature methods* 5 (2):175-177
11. Yu-Ping W, Gunampally MR, Wei-Wen C Automated segmentation of microarray spots using fuzzy clustering approaches. In: *IEEE workshop on Machine Learning for Signal Processing*, 2005. pp 387-391
12. Arena P, Fortuna L, Occhipinti L (2002) A CNN algorithm for real time analysis of DNA microarrays. *IEEE Transaction on Circuits and Systems* 49 (3):335-340
13. Sreedevi A, Jangamashetti D Automatically Locating Spots in DNA Microarray Image Using Genetic Algorithm without Gridding. In: *International Association of Computer science and Information Technology – Spring Conference*, 2009. pp 178-181

14. Bariamis D, Maroulis D, Iakovidis DK Automatic DNA microarray gridding based on Support Vector Machines. In: IEEE international conference on bioinformatics and bioengineering, 2008. pp 1-5
15. Labib FEZ, Fouad I, Mabrouk M, Sharawy A (2012) An efficient fully automated method for gridding microarray images. American Journal of Biomedical Engineering 2 (3):115-119
16. Harikiran J, RamaKrishna D, Avinash B, Lakshmi P, KiranKumar R (2014) A New Method of Gridding for Spot Detection in Microarray Images. Computer Engineering and Intelligent Systems 5 (3):25-33
17. Yang YH, Buckley MJ, Speed TP (2001) Analysis of cDNA microarray images. Briefings in bioinformatics 2 (4):341-349
18. Chen W-B, Zhang C, Liu W-L An Automated Gridding and Segmentation Method for cDNA Microarray Image Analysis. In: Computer-Based Medical Systems, 2006. pp 893-898
19. Nagarajan R, Upreti M (2006) Correlation statistics for cDNA microarray image analysis. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 3 (3):232-238
20. Nicolini C, Malvezzi AM, Tomaselli A, Sposito D, Tropicano G, Borgogno E (2002) DNASER. I. Layout and data analysis. IEEE Transactions on NanoBioscience 1 (2):67-72
21. Conte D, Foggia P, Sansone C, Vento M Graph matching applications in pattern recognition and image processing. In: International Conference Image Processing, 2003. pp 21-24
22. Eisen M (2002) Microarray Image Analysis Software, ScanAlyze. <http://www.eisenlab.org>, Accessed December 2015
23. Instruments A (2013) GenePix® Pro 7 Microarray Acquisition & Analysis Software. <http://mdc.custhelp.com>. Accessed December 2015
24. Medigue C (2012) ImaGene® 9.0 – Leading-Edge Microarray Analysis Software. <http://www.biodiscovery.com>. Accessed December 2015
25. Sun S, Huang Y-W, Yan PS, Huang TH, Lin S (2011) Preprocessing differential methylation hybridization microarray data. BioData mining 4 (1):1-15
26. Yin W, Chen T, Zhou SX, Chakraborty A (2005) Background correction for cDNA microarray images using the TV+ L1 model. Bioinformatics 21 (10):2410-2416
27. Pierre S (1999) Morphological image analysis: Principles and Applications. Springer-Verlag Berlin Heidelberg, 5 (3): 223-230

28. Cho S-B, Won H-H Machine learning in DNA microarray analysis for cancer classification. In: The First Asia-Pacific bioinformatics conference on Bioinformatics, 2003. pp 189-198
29. Tsai M-C, Park Y (2007) Cancer classification using Machine Learning Technique on Microarray Data. Machine Learning, 5 (1): 16-22
30. Kumar PG, Victoire TAA, Renukadevi P, Devaraj D (2012) Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm. Expert Systems with Applications 39 (2):1811-1821
31. Kokate H, Nair N, Shete K, Thakur T, Ahirrao S, Detection of Colon Cancer by Classification of Genes and Feature Selection using Microarray Data. International Conference of Advance Research and Innovation 2014. pp: 268-271
32. Abeer M.Mahmoud BAM (2014) A Hybrid Reduction Approach for Enhancing Cancer Classification of Microarray Data. International Journal of Advanced Research in Artificial Intelligence 3 (2):1-10
33. جوادی م , ریحانی ر (۱۳۸۷) روشی جدید برای انتخاب ژن در آنالیز داده‌های ریزآرایه و دسته‌بندی با استفاده از knn. دومین همایش ملی مهندسی کامپیوتر، برق و فناوری اطلاعات – سما – همدان
34. امین‌زاده ف، عصاره ع (۱۳۸۹) تشخیص اتوماتیک بیماری سرطان با استفاده از تحلیل داده‌های ریزآرایه. هجدهمین کنفرانس مهندسی برق ایران، دانشگاه صنعتی اصفهان
35. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science 286 (5439):531-537
36. Wang L, Chu F, Xie W (2007) Accurate cancer classification using expressions of very few genes. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 4 (1):40-53
37. Halder A, Dey S, Kumar A (2015) Active Learning Using Fuzzy k-NN for Cancer Classification from Microarray Gene Expression Data. In: Advances in Communication and Computing, pp 103-113
38. زالی ح، رضایی م، سلیمیان ج، اولاد ر، نژاد ص (۱۳۹۱) شبکه های بیانی ژنی برای آنالیز داده های ماکرواری DNA. مجله علمی دانشگاه علوم پزشکی ایلام
39. Quackenbush J (2006) Microarray analysis and tumor classification. New England Journal of Medicine 354 (23):2463-2472
40. Yokota H (2015) Microarray Resource Core. Boston University. <https://www.scienceexchange.com/>. Accessed November 2015
41. Center UoTMAC (2014) Microarray Services – Support and Pricing. <http://www.mdanderson.org>. Accessed November 2015

42. Corporation A (2015) Services - Microarray Scanning and Data Analysis Services. <http://www.arrayit.com>. Accessed November 2015
43. Instruments A (2013) GenePix Pro 6.0 Microarray Acquisition and Analysis Software for GenePix Microarray Scanners User's Guide & Tutorial. Axon Instruments/ Molecular Devices Corp, Sunnyvale, CA.
44. Wallack D (2002) Data Analysis with ScanAlyze. Muhlenberg College, PA. <http://www.bio.davidson.edu>. Accessed September 2015
45. Bariamis D, Maroulis D, Iakovidis DK (2010) Unsupervised SVM-based gridding for DNA microarray images. *Computerized Medical Imaging and Graphics* 34 (6):418-425
46. Bozinov D, Rahnenführer J (2002) Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering. *Bioinformatics* 18 (5):747-756
47. Giannakeas N, Fotiadis DI (2009) An automated method for gridding and clustering-based segmentation of cDNA microarray images. *Computerized Medical Imaging and Graphics* 33 (1):40-49
48. Uslan V, Bucak O, Cekmece B (2010) Microarray image segmentation using clustering methods. *Mathematical and Computational Applications* 15 (2):240-247
49. Wang T, Li T, Shao G, Wu S (2015) An improved K-means clustering method for cDNA microarray image segmentation. *Genetics and molecular research: GMR* 14 (3):7771-7781
50. Shao G, Wang T, Hong W, Chen Z An improved SVM method for cDNA microarray image segmentation. In: *IEEE International Conference on Computer Science & Education (ICCSE)*, 2013. pp 391-395
51. Kannan SS, John SM, Sundaravadivel P, Ilangovan S, Kumar AVA (2013) Performance Evaluation of Classification Algorithms C4. 5 and SMO on Microarray Gene-set Data. *International Journal of Research in Engineering and Technology* 2 (5):275-279
52. Nagi S, Bhattacharyya DK (2013) Classification of microarray cancer data using ensemble approach. *Network Modeling Analysis in Health Informatics and Bioinformatics* 2 (3):159-173
53. Jaison B, Chilambuchelvan A, Junaid KM (2015) Hybrid Classification Techniques for Microarray Data. *National Academy Science Letters* 38 (5):415-419
54. ابوالقاسمی ر، وثیقی م، صادقی ب، جعفری ل (۱۳۹۲) انتخاب ژن‌های حاوی اطلاعات و طبقه‌بندی داده‌های ریزآرایه با استفاده از الگوریتم ژنتیک و شبکه عصبی خود سازمان‌ده. اولین همایش ملی نانو تکنولوژی مزایا و کاربردها
55. مروت م، عصاره ع، شادگار ب (۱۳۹۳) دسته‌بندی مجموعه داده‌های ریزآرایه براساس تکنیک‌های ترکیبی به منظور تشخیص سرطان. نوزدهمین کنفرانس ملی سالانه انجمن کامپیوتر ایران

56. Yoon Y, Bien S, Park S (2010) Microarray data classifier consisting of k-top-scoring rank-comparison decision rules with a variable number of genes. *IEEE Transactions On Systems, Man, And Cybernetics—Part C: Applications And Reviews* 40 (2):216-226
57. Wang Y, Ma MQ, Zhang K, Shih FY (2007) A hierarchical refinement algorithm for fully automatic gridding in spotted DNA microarray image processing. *Information Sciences* 177 (4):1123-1135
58. Wang Y, Shih F, Ma M Precise gridding of microarray images by detecting and correcting rotations in subarrays. In: *Proceedings of the 8th Joint Conference on Information Sciences*, 2005. pp 1195-1198
59. Fouad IA (2014) a new robust and automatic analytical approach for block indexing of cDNA Microarray image. *International Journal of Computer Engineering and Applications* 5 (1):1-16
60. Rueda L (2007) Sub-grid detection in DNA microarray images. In: *Proceedings of the IEEE Pacific-RIM Symposium on Image and Video Technology*, pp 248-259
61. Rueda L, Rezaeian I (2011) A fully automatic gridding method for cDNA microarray images. *BMC bioinformatics* 12 (1):1-17
62. Patel J, Shah A, Patel H (2015) Skew Angle Detection and Correction using Radon Transform. *International Journal of Electronics, Electrical and Computational System* 4 (1):1-6
63. Otsu N (1975) A threshold selection method from gray-level histograms. *Automatica* 11 (1):23-27
64. Uchida S, Nishida Y, Satou K, Muta S, Tashiro K, Kuhara S (2005) 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 research* 12 (1):1-7
65. Beare RaB, M. (2004) *cDNA Microarray Image Analysis—User’s Guide*. CSIRO Mathematical and Information Sciences
66. Jones AL (2015) Segmenting Microarrays with Deep Neural Networks. *bioRxiv* 2 (1):1-7
67. Smyth GK, Yang YH, Speed T (2003) Statistical issues in cDNA microarray data analysis. In: Brownstein MJ, Khodursky AB (eds) *Functional Genomics: Methods and Protocols*, Humana Press, New Jersey, pp 111-136
68. Yang YH, Buckley MJ, Dudoit S, Speed TP (2002) Comparison of methods for image analysis on cDNA microarray data. *Journal of computational and graphical statistics* 11 (1):108-136
69. Chan TF, Esedoglu S (2005) Aspects of total variation regularized L 1 function approximation. *SIAM Journal on Applied Mathematics* 65 (5):1817-1837

70. Alizadeh F, Goldfarb D (2003) Second-order cone programming. *Mathematical programming* 95 (1):3-51
71. Speed T (2001) Always log spot intensities and ratios. Speed Group Microarray <http://www.stat.berkeley.edu/users/terry/zarray/Html/loghtml> Accessed Octobre 2015
72. Cui X, Kerr MK, Churchill GA (2003) Transformations for cDNA microarray data. *Statistical Applications in Genetics and Molecular Biology* 2 (1):1-22
73. Tarca AL, Romero R, Draghici S (2006) Analysis of microarray experiments of gene expression profiling. *American journal of obstetrics and gynecology* 195 (2):373-388
74. Dudoit S, Yang YH, Callow MJ, Speed TP (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica sinica* 12 (1):111-140
75. Chua S-W, Vijayakumar P, Nissom PM, Yam C-Y, Wong VV, Yang H (2006) A novel normalization method for effective removal of systematic variation in microarray data. *Nucleic acids research* 34 (5):1-7
76. de Resende MDV, e Silva FF, Nascimento M, Azevedo CF, Peternelli LA (2015) Statistical Analysis of Gene Expression and Genomic Data. In: Fritsche R, Borém A (eds) *Phenomics*. Springer International Publishing, pp 33-47
77. Yang B, Bao X (2015) Identification of genes associated with laryngeal squamous cell carcinoma samples based on bioinformatic analysis. *Molecular medicine reports* 12 (1):3386-3392
78. Smyth GK, Speed T (2003) Normalization of cDNA microarray data. *Methods* 31 (4):265-273
79. Welsh EA, Eschrich SA, Berglund AE, Fenstermacher DA (2013) Iterative rank-order normalization of gene expression microarray data. *BMC bioinformatics* 14 (1):153-158
80. Wu Z, Irizarry RA, Gentleman R, Martinez-Murillo F, Spencer F (2004) A model-based background adjustment for oligonucleotide expression arrays. *Journal of the American statistical Association* 99 (468):909-917
81. Wang Z, Palade V, Xu Y (2006) Neuro-fuzzy ensemble approach for microarray cancer gene expression data analysis. In: *IEEE International Symposium on Evolving Fuzzy Systems*. pp 241-246
82. Dai J, Xu Q (2013) Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification. *Applied Soft Computing* 13 (1):211-221
83. Lavanya D, Rani KU (2012) Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services (IJITCS)* 2 (1):17-24

84. Al Snousy MB, El-Deeb HM, Badran K, Al Khilil IA (2011) Suite of decision tree-based classification algorithms on cancer gene expression data. *Egyptian Informatics Journal* 12 (2):73-82
85. Salzberg SL (1994) C4.5: Programs for Machine Learning by J. Ross Quinlan. Morgan Kaufmann Publishers, Inc., 1993. *Machine Learning* 16 (3):235-240
86. Stanford Microarray Database (2002) <http://smd.stanford.edu/>. Accessed December 2015
87. Kao J, Salari K, Bocanegra M, Choi Y-L, Girard L, Gandhi J, Kwei KA, Hernandez-Boussard T, Wang P, Gazdar AF (2009) Molecular profiling of breast cancer cell lines defines relevant tumor models and provides a resource for cancer gene discovery. *PloS one* 4 (7):1-17
88. Bergamaschi A, Kim YH, Wang P, Sørli T, Hernandez-Boussard T, Lonning PE, Tibshirani R, Børresen-Dale AL, Pollack JR (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* 45 (11):1033-1040
89. Bohen SP, Troyanskaya OG, Alter O, Warnke R, Botstein D, Brown PO, Levy R (2003) Variation in gene expression patterns in follicular lymphoma and the response to rituximab. *Proceedings of the National Academy of Sciences* 100 (4):1926-1930
90. Lossos IS, Alizadeh AA, Diehn M, Warnke R, Thorstenson Y, Oefner PJ, Brown PO, Botstein D, Levy R (2002) Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. *Proceedings of the National Academy of Sciences* 99 (13):8886-8891
91. Rücker F, Sander S, Döhner K, Döhner H, Pollack J, Bullinger L (2006) Molecular profiling reveals myeloid leukemia cell lines to be faithful model systems characterized by distinct genomic aberrations. *Leukemia* 20 (6):994-1001
92. Bullinger L, Döhner K, Bair E, Fröhling S, Schlenk RF, Tibshirani R, Döhner H, Pollack JR (2004) Use of gene-expression profiling to identify prognostic subclasses in adult acute myeloid leukemia. *New England Journal of Medicine* 350 (16):1605-1616
93. Fang Z, Tian W, Ji H (2012) A network-based gene-weighting approach for pathway analysis. *Cell research* 22 (3):565-580
94. Fitzpatrick DJ, Ryan CJ, Shah N, Greene D, Molony C, Shields DC (2015) Genome-wide epistatic expression quantitative trait loci discovery in four human tissues reveals the importance of local chromosomal interactions governing gene expression. *BMC genomics* 16 (1):109-122

Abstract:

In recent decades, genetics particularly molecular genetics has been widely developed. The advent of microarray technology as a powerful tool for simultaneous studying and analyzing of the thousands of genes behavior is one of the developments. Images of microarray technology play an important role in the detection, diagnosis and treatment of diseases and analysis of the images has a direct impact on results. But systems to analyze these images are only able to quantify the images (to extract raw data from the image) of microarray and are generally operated manually. This project aims to provide an automated system to extract and analyze microarray images data in order to detect cancer diseases. The proposed system consists of three main phases of image processing, data mining, and disease diagnosis. The image processing includes operations such as noise cancelling, rotation fixing, the gene locating, background deleting and raw data extracting from the image. Data mining involves normalization of the extracted data and selection of the more effective genes. In the third phase, according to the extracted data, identification and diagnosis of the disease are carried out.

In this study, microarray images of breast cancer, leukemia and lymphatic cancer developed by Stanford University are used to evaluate and test the proposed system. The test results show that the system accuracy at genes locating phase on the microarray images is more than 98%. In addition, the system is capable to detect the cancer type in the breast cancer data base, lymphatic cancer and leukemia with an accuracy of 45.95%, 100% and 11.94%, respectively.

Keywords: Microarray; Image Processing; Data Mining; Disease Detection



Shahrood University of Technology
Department of Computer and Information Technology
Computer & Artificial intelligent

DNA microarray data Classification for cancer detection

Nastaran Dehghan Khalilabad

Supervisor:

Dr. Hamid Hassanpour

Adviser:

Dr. Mohammad Reza Abbaszadegan

February, 2016

