ژنتیک نوین
دوره هفدهم، شماره ۲، تابستان ۱۴۰۱
صفحه ۱۴۷ - ۱۳۷

تشخیص لوسمی لنفوسیتی و میلوئیدی حاد با استفاده از انتخاب ژن‌های مؤثر ریزآرایه و الگوریتم‌های داده‌کاوی

Detecting Acute Lymphocytic and Myeloid Leukaemia by Selecting Effective Microarray and Data-Mining Algorithms

علی‌رضا حاجی‌اسکندر[۱]، مجید خلیلیان*[۱]، جواد محمدزاده[۱]، علی نجفی[۲]

۱- به‌ترتیب دانشجوی دکتری، استادیار، استادیار، گروه کامپیوتر، واحد کرج، دانشگاه آزاد اسلامی، کرج، ایران
۲- دانشیار، مرکز تحقیقات بیولوژی مولکولی، پژوهشکده سیستم بیولوژی مسمویت‌ها، دانشگاه علوم پزشکی بقیة الله (عج)، تهران، ایران

Hajieskandar AR[1], Khalilian M*[1], Mohammadzadeh J[1], Najafi A[2]

1- PhD Student, Assistant Professor, Assistant Professor, Department of Computer Engineering, Karaj Branch, Islamic Azad University, Karaj, Iran
2- Associate Professor, Molecular Biology Research Center, Systems Biology and Poisonings Institute, Baqiyatallah University of Medical Sciences, Tehran, Iran

* نویسنده مسئول مکاتبات، پست الکترونیکی: khalilian@kiau.ac.ir

(تاریخ دریافت: ۱۴۰۰/۰۶/۲۷ - تاریخ پذیرش: ۱۴۰۱/۰۱/۲۴)

چکیده

سرطان یکی از مهم‌ترین علت‌های مرگ و میر در جهان است. در بیشتر موارد اگر این بیماری زودتر شناسایی شود قابل درمان است. یکی از روش‌های تشخیص سرطان استفاده از داده‌های ریزآرایه است که بر خلاف روش تصویربرداری، اشعه‌های مضر برای انسان در آن وجود ندارد. ریزآرایه‌ها دارای ژن‌های بسیار زیادی هستند که باعث پیچیده و زمان‌بر شدن تحلیل می‌شوند بنابراین انتخاب ژن‌های مفید یکی از گام‌های اساسی در تشخیص این بیماری است. روش پیشنهادی در این مقاله دارای دو فاز اصلی است که فاز اول آن انتخاب ژن‌های مؤثر است. در فاز بعدی، عمل تشخیص بیماری از ژن‌های انتخاب شده توسط فاز اول انجام می‌گیرد. در گذشته الگوریتم‌های زیادی مانند Ridge برای این منظور ارایه شده است که با توجه به نتایج به‌دست آمده از آزمایش‌ها، دقت روش پیشنهادی این مقاله نسبت به آن‌ها برتری دارد. در این مقاله از مجموعه داده لوکیمیا و سرطان روده به‌عنوان ورودی و ارزیابی روش پیشنهادی استفاده شده است. دقت روش پیشنهادی جهت تعیین مکان ژن‌ها و تشخیص نوع سرطان لوکیمیا و سرطان روده به‌ترتیب ۹۷/۶۲٪ و ۹۲/۳۱٪ است. نتایج به‌دست آمده از این روش نسبت به دیگر روش‌های موجود از نظر دقت در مجموعه داده لوکیمیا ۲/۳۸٪ و در مجموعه داده سرطان روده ۵/۶۱٪ بهبود یافته است؛ همچنین نسبت به آزمایش‌های بیولوژی از دسترسی آسان‌تر و هزینه کمتری برخوردار است.

واژه‌های کلیدی

استخراج ویژگی
ریزآرایه
سرطان روده
شبکه عصبی LSTM
لوکیمیا

مقدمه

سرطان یک بیماری ژنتیکی است که معمولاً زاییده اثرات عوامل محیطی است (Bariamis et al. 2008; Labib et al. 2012) و بعد از بیماری‌های قلبی ـ عروقی، دومین علت اصلی مرگ و میر در جهان به‌شمار می‌رود. اگر سرطان‌ها در مراحل اول تشخیص داده شوند، معمولاً قابل معالجه هستند (Harikiran et al. 2014). سرطان خون یا لوسمی، بیماری پیش‌رونده و بدخیم اعضای خون‌ساز بدن است و یکی از مهم‌ترین سرطان‌هایی است که جامعه بشری با آن درگیر است. در بیماری لوسمی، مغز استخوان به‌صورت غیرعادی، مقدار بسیار زیادی سلول خونی تولید می‌کند. این سلول‌ها با سلول‌های خون طبیعی متفاوت بوده و درست عمل نمی‌کنند، در نتیجه، تولید طبیعی گلبول‌های سفید خون را متوقف و توانایی فرد در مقابله با بیماری‌ها را از بین می‌برند (Toloie and Taheri 2010).

لوسمی بر اساس طیف، شدت و سرعت پیشرفت روند بیماری به دو نوع حاد و مزمن و بر اساس نوع گلبول سفید درگیر، به لنفوئیدی و میلوئیدی تقسیم می‌شود (Instruments 2013) (Bozinov and Rahnenführer 2002). بسیاری از مطالعه‌ها روند بدخیمی لوسمی را به ناهنجاری‌های ژنتیکی نسبت می‌دهند و مطالعه‌های زیادی پیرامون کشف عوامل مولکولی درگیر در این بیماری صورت گرفته است (Wang et al. 2015). به‌کمک پیشرفت‌های فناوری در بیوانفورماتیک و روش‌های مولکولی، داده‌های زیادی به‌دست آمده که در شناخت زودرس بیماری سرطان کمک خواهد کرد. همچنین غربال‌گری به موقع برای بعضی از سرطان‌ها، کمک مؤثری در تشخیص زودرس آن می‌نماید (Labib et al. 2012). با توجه به این که گرفتن تصمیم مناسب برای درمان انواع لوسمی از مهم‌ترین فعالیت‌ها بعد از تشخیص نوع سرطان است، هدف از انجام این مقاله، تشخیص لوسمی میلوژنیک و لنفوسیتیک حاد با استفاده از انتخاب ژن‌های مؤثر از داده‌های ریز آرایه و الگوریتم‌های داده کاوی است (Harikiran et al. 2014).

یکی از حوزه‌های جدید دانش در بیان ژن‌ها، استفاده از فن‌آوری ریزآرایه[1] است (Tarca et al. 2006). این فن‌آوری امکان بررسی هم‌زمان بسیاری از فعل و انفعالات زیستی را فرآهم می‌کند بنابراین می‌توان با تحلیل آماری تغییرات بیان ژن‌ها به‌طور هم‌زمان، ژن‌های مؤثر در سرطان را شناسایی کرد (Medigue 2015; Wallack 2015; Golub et al. 1999; Ramaswamy et al. 2001; Halder et al 2015; Wang et al. 2007). از مهم‌ترین کاربردهای ریزآرایه، بررسی بیان ژن و تغییرات آن در اثر عواملی مانند درمان، عوامل بیماری‌زا، آسیب سلول، تعیین محتوای ژنوم موجودات زنده، مقایسه آن‌ها با یکدیگر، شناسایی چند شکلی‌های تک نوکلئوتیدی، تشخیص بیماری و طبقه‌بندی سرطان است (Sreedevi and Jangamashetti 2009).

با توجه به این که ابعاد داده‌های ریزآرایه بسیار زیاد است، تحلیل آن‌ها پیچیده بوده و هزینه‌ی محاسباتی بالایی دارد، از این رو، اولین گام مهم در آنالیز داده‌های ریز آرایه، کاهش تعداد ژن‌ها یا به‌عبارتی انتخاب ژن‌های مؤثر است و انجام این فرآیندها بدون کمک آنالیز آماری و روش‌های هوشمند تحلیل اطلاعات ممکن نیست (Yin et al. 2005). الگوریتم‌های مختلف داده کاوی و یادگیری ماشین[2] می‌توانند در خوشه‌بندی و دسته‌بندی ژن‌ها مورد استفاده قرار گرفته و در نتیجه در تشخیص بیماری مؤثر باشند (Toloie and Taheri 2010).

به‌کمک پیشرفت‌های فن‌آوری در بیوانفورماتیک و روش‌های مولکولی، داده‌های زیادی به‌دست آمده که با استفاده از آن‌ها می‌توان بیماری سرطان را زودتر تشخیص داد. مطالعه‌های متعددی توسط محققان بر روی مجموعه داده‌های بیان ژن لوسمی با روش‌های مختلف انجام گرفته است (Labib et al. 2012; Harikiran et al. 2014).

مطالعه‌های متعددی توسط محققان بر روی مجموعه داده‌های بیان ژن لوسمی با روش‌های مختلف انجام گرفته است. در پژوهشی که مربوط به سال ۲۰۱۸ است اهمیت MA[3] نشان داده شده است. انتخاب زیر مجموعه کوچکی از ژن‌ها که به ساختن یک مدل طبقه‌بندی مناسب برای پیش‌بینی بیماری بر روی داده‌های ریزآرایه کمک می‌کند یک مسأله بهینه‌سازی است. الگوریتم ژنتیک (GA[4]) یک الگوریتم بهینه‌سازی مبتنی بر جمعیت است که

[1] Microarray

[2] Machine learning
[3] Memetic Algorithm
[4] Genetic Algorithm

کاربردهای زیادی در زمینه زیست‌شناسی مولکولی دارد. اما هم‌گرایی زودرس یکی از محدودیت‌های GA است. الگوریتم ممتیک، احتمال چنین هم‌گرایی زودرس را کاهش می‌دهد. برتری MA نسبت به GA، تبرید شبیه‌سازی شده (SA[1]) و جستجوی ممنوعه (TS[2]) اثبات شده است. آزمایش‌ها روی سه مجموعه داده مشهور، یعنی DLBCL، لوسمی و سرطان پروستات، نشان می‌دهد که MA نتایج امیدوار کننده‌تری نسبت به GA، SA و TS کلاسیک دارد (Begum et al. 2018).

در سال ۲۰۱۸ تحقیقی در زمینه مدیریت بهینه لوسمی میلوئید حاد انجام شده است. این تحقیق به دنبال شناسایی ویژگی‌های متمایز سلول‌های AML برای نظارت جهانی MRD[3] است که در آن ژن گسترده سلول‌های AML از ۱۵۷ بیمار با میلووبلاست طبیعی مقایسه شده است. نشان‌گرهای رمزگذاری شده توسط ژن‌های بیان شده به‌طور غیرقابل انفعال، از جمله برخی از سلول‌های بنیادی لوسمی، که قبلاً با سلول‌های بنیادی لوسمی مرتبط بودند، با استفاده از فلوسیتومتری در ۲۴۰ بیمار مبتلا به AML و در میلووبلاست‌های غیرلوسمی از ۶۳ نمونه مغز استخوان مورد مطالعه قرار گرفته و نتایج ۲۲ نشان‌گر به‌طور نامحسوس در AML بیان شده‌اند. اطلاعات مربوط به لوسمی تعریف شده توسط این نشان‌گرها به سلول‌های نابالغ AML کامل شده و نشان داده شده است که در طول درمان پایدار است. MRD روی ۱۲۹ بیمار به‌صورت پیاپی انجام شده و با استفاده از یک الگوریتم یادگیری ماشین نتایج قابل توجهی به‌دست آمده است که می‌تواند یک سلول لوسمی را در بین بیش از یک میلیون سلول طبیعی تشخیص دهد (Coustan-Smith et al. 2018).

سرطان‌هایی که از نظر پاتولوژیک مشابه هستند، اغلب به رژیم‌های دارویی یکسان پاسخ‌های متفاوت نشان می‌دهند. در روشی امیدوار کننده برای شناسایی مارکرهای مولکولی قوی برای درمان هدفمند لوسمی میلوئید حاد ارائه شده است (Lee et al. 2018). با معرفی داده‌ها از ۳۰ بیمار AML شامل پروفایل بیان ژن گسترده و حساسیت آزمایشگاهی به ۱۶۰ داروی شیمی درمانی، یک روش محاسباتی برای شناسایی مارکرهای بیان ژن برای حساسیت به دارو با ترکیب اطلاعات قبلی چندتایی مربوط به پتانسیل هر ژن برای پیشگیری از سرطان ارائه شده است که این روش در شناسایی مارکرهای مولکولی تکثیر شده در داده‌های اعتبار سنجی و پیش‌بینی دقیق حساسیت به دارو موفق عمل می‌کند. در این مقاله از داده‌های مربوط به ۳۰ نمونه بیمار AML و ۱۶۰ داروی موجود در پانل داروهای سفارشی و کلاس‌های مکانیسم عملکرد آن‌ها در داده‌های تکمیلی استفاده شده است.

در از الگوریتم RMA[4] برای کاهش تعداد ویژگی‌ها و در ادامه پیش‌بینی بیماری‌های سرطانی استفاده شده است (Ghosh et al. 2019). از آن‌جایی که تعداد نشان‌گرهای زیستی در داده‌ها بسیار کم است، بنابراین الگوریتم‌های تکاملی سنتی نمی‌توانند نتایج خوبی ایجاد کنند و همچنین به‌دلیل کاوش در فضای جستجوی بزرگ نیازمند زمان زیادی است و کاهش فضای جستجو با استفاده از روش‌های فیلتر اگرچه متداول است اما نتیجه خوبی را تضمین نمی‌کند، یعنی اگر فضای جستجو تا حد زیادی توسط این‌گونه الگوریتم‌ها کاهش داده شود، ممکن است ویژگی‌های مهم از دست بروند. این امر باعث می‌شود که کاهش فضای ویژگی مورد توجه اصلی قرار گیرد. بنابراین در این مقاله با استفاده از الگوریتم RMA تعداد ویژگی‌ها کاهش یافته است. این الگوریتم در موارد AMLGSE2191، Colon، DLBCL، لوسمی، پروستات، MLL و SRBCT استفاده شده و نتایج خوبی به‌دست آمده و در مقایسه با GA و MA اصلی عملکرد بهتری داشته است. پتانسیل RMA در شناسایی نشان‌گرهای زیستی و طبقه‌بندی نمونه‌های بیان ژن با دقت بالا بسیار خوب است.

با توجه به روند رو به رشد استفاده از تکنیک‌های ML[5] در پزشکی، در یک روش مبتنی بر یادگیری ماشین ارائه شده است که احتمال بهبودی کامل در بیماران مبتلا به لوسمی میلوئید حاد را پیش‌بینی می‌کند (Gal et al. 2019). در این مطالعه، این سؤال بررسی می‌شود که آیا این الگوریتم‌ها می‌توانند از طریق توالی RNA احتمال بهبودی کامل در کودکان AML که تحت درمان القایی قرار دارند را به‌صورت دقیق پیش‌بینی کنند. برای شناسایی مدلی که بهترین عملکرد را در چارچوب این مطالعه دارد،

[1] Simulated Annealing
[2] Tabu Search
[3] Minimal Residual Disease
[4] Recursive Memetic Algorithm
[5] Machine learning

منحنی‌های مشخصه عملیاتی گیرنده رسم شده است. بهترین مقدار ناحیه زیرمنحنی با استفاده از ۷۵ ژن برتر برای الگوریتم K-NN، ۰/۸٤ به‌دست آمده است. اطلاعات به‌دست آمده از این مقاله ممکن است برای پزشکان در انتخاب دوره‌های درمانی مناسب و یا یک درمان جایگزین جدید کمک زیادی داشته باشد.

در این مقاله، یک روش خودکار جهت استخراج و تحلیل داده‌های ریزآرایه به‌منظور تشخیص بیماری‌های سرطانی ارائه شده است. روش پیشنهادی شامل دو فاز اصلی داده‌کاوی و شناسایی نوع بیماری است. در فاز اول داده‌های استخراج شده، نرمال‌سازی شده و با کمک تکنیک‌های داده‌کاوی، ژن‌های مفید و تاثیرگذار از بین هزاران ژن انتخاب می‌شوند؛ در این فاز به‌منظور کاهش حجم فضای جستجو با استفاده از الگوریتم K_Means داده‌های ریزآرایه خوشه‌بندی می‌شوند و خوشه‌ها در اختیار الگوریتم ReliefF قرار می‌گیرند تا ژن‌های مؤثر آن برای تشخیص سرطان انتخاب شوند. سپس در فاز دوم، عمل شناسایی و تشخیص بیماری با توجه به داده‌های استخراج شده انجام می‌گیرد. داده‌های ریزآرایه جهت تشخیص نوع سرطان، با استفاده از الگوریتم‌های موجود در یادگیری ماشین دسته‌بندی[1] می‌شوند که به‌منظور تعیین مرحله‌ی درمانی بیمار یا تجویز دارو نقش به‌سزایی دارد. در نهایت برای تشخیص نمونه‌های سرطانی از شبکه عصبی LSTM[2] استفاده شده است.

مواد و روش‌ها

در این مقاله، راه‌کاری برای شناسایی و آماده‌سازی داده‌ها، استخراج ویژگی‌های مؤثر در سرطان و مدل‌سازی یادگیری عمیق، با استفاده از زبان پایتون نسخه ۳ و بسته کراس یادگیری عمیق توسعه یافته است. هدف راه‌کار پیشنهادی[3] تشخیص لوسمی و سرطان کلورکتال با استفاده از مجموعه داده‌های ریزآرایه مبتنی بر استخراج ویژگی با متد ReliefF و تشخیص کلاس نهایی با استفاده از شبکه عصبی LSTM است. برای ارزیابی روش پیشنهادی، از دو مجموعه داده لوکیمیا[4] و سرطان روده[5] استفاده شده است. مجموعه داده لوکیمیا شامل ٤۹ نمونه لوسمی لنفوسیتیک حاد[6] (ALL) و ۲۳ نمونه لوسمی میلوژنیک حاد[7] (AML) است و هر نمونه شامل ۷۱۲۹ ژن است. در مجموعه داده سرطان روده ٤۰ نمونه از ٦۲ نمونه موجود سرطانی هستند و بقیه بافت‌های معمولی را تشکیل می‌دهند، تعداد ژن‌های هر نمونه در این مجموعه داده ۲۰۰۰ ژن است. روش پیشنهادی از سه بخش تشکیل شده است که در ادامه هر یک از بخش‌ها توضیح داده می‌شوند.

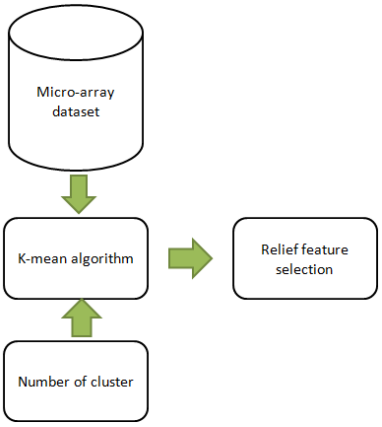
مجموعه داده‌های سرطان روده بزرگ و لوسمی را به‌صورت جداگانه به‌عنوان ورودی دریافت کرده و توسط الگوریتم خوشه‌بندی K-means به ٦ خوشه (تعداد خوشه‌ها با آزمایش و خطا به‌دست آمده است) خوشه‌بندی می‌کند؛ با استفاده از این کار فضای جستجو به ٦ خوشه تقسیم می‌شود و در نتیجه فضای جستجوی محدود شده و منظمی در اختیار متد ReliefF برای استخراج ویژگی‌های مؤثر قرار می‌گیرد. در حقیقت متد ریلیف به‌جای جستجو در کل داده‌های ورودی، در بین خوشه‌ها عمل جستجو را انجام می‌دهد. خوشه‌بندی را می‌توان به‌عنوان مهم‌ترین مسأله در یادگیری بدون نظارت در نظر گرفت. الگوریتم خوشه‌بندی K-means بر اساس حداقل کردن مربع خطاها یا تغییرات درون گروهی که معادل با بیشینه کردن تغییرات بین خوشه‌هاست، بنا نهاده شده است. بنابراین، هدف کلی این الگوریتم به‌دست آوردن خوشه‌هایی است که با مقدار ثابت K به‌طور کلی مربع خطاها را کمینه کنند. این قسمت از روش پیشنهادی در شکل ۱ نمایش داده شده است.

روش پیشنهادی از روش ReliefF استفاده می‌کند که یک روش انتخاب متغیر مبتنی بر معیار فاصله است. این بخش، داده‌های خوشه‌بندی شده را به‌عنوان ورودی دریافت کرده و بعد از انتخاب ویژگی‌های مؤثر در سرطان، آن‌ها را به قسمت بعدی تحویل می‌دهد. در این روش، اگر یک متغیر به ازای نمونه‌های

[1] Classification
[2] Long Short-Term Memory
[3] Relief +LSTM

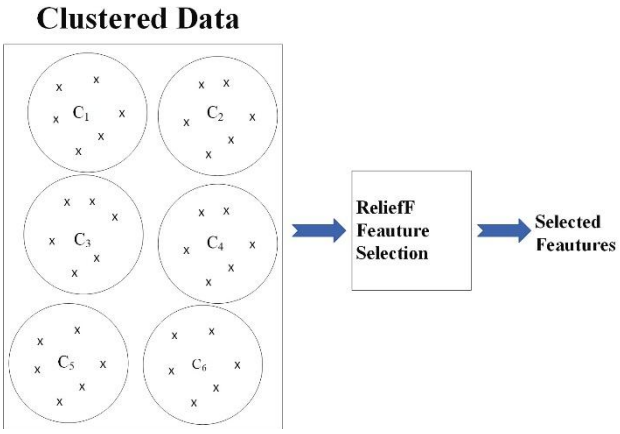
[4] https://rdrr.io/cran/propOverlap/man/leukaemia.html
[5] https://rdrr.io/cran/ShrinkCovMat/man/colon.html
[6] Acute Lymphoblastic leukaemia
[7] Acute Myeloid leukaemia

درون یک طبقه، مقدار یکسان و به ازای نمونه‌های دیگر طبقه‌ها مقادیر مختلفی داشته باشد، وزن بالاتری می‌گیرد.



شکل ۱- بخش اول روش پیشنهادی

ریلیف از بین داده‌های آموزشی، یک نمونه را به‌صورت تصادفی انتخاب می‌کند و سپس فاصله اقلیدسی آن نمونه تا نزدیک‌ترین نمونه از طبقه مشابه و نزدیک‌ترین نمونه از طبقه متفاوت را به‌دست آورده و سپس این فاصله‌ها را برای به روز کردن وزن هر متغیر به‌کار می‌برد. در نهایت، الگوریتم، آن دسته از متغیرهایی را انتخاب می‌کند که وزن آن‌ها از یک حد آستانه از پیش تعریف شده به‌وسیله کاربر، بیشتر باشد، در واقع، رتبه‌ای که ریلیف به هر متغیر می‌دهد بر اساس میزان نقش آن متغیر در جداسازی نمونه‌های متفاوت همسایه است. شکل ۲ بیان‌گر این بخش از روش پیشنهادی است.



شکل ۲- بخش دوم روش پیشنهادی

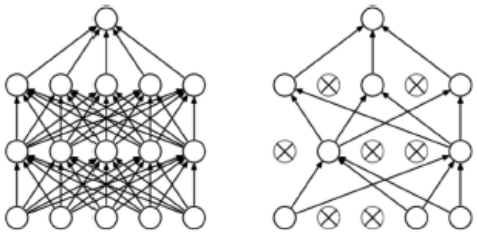
از شبکه عصبی عمیق LSTM برای کلاس‌بندی استفاده می‌کند. این شبکه نوعی شبکه عصبی مکرر است که با استفاده از سلول‌های حافظه اختصاصی، اطلاعات را به مدت طولانی ذخیره می‌کند. پیکره خاص عملیات در این شبکه، جریان اطلاعات را در داخل کنترل می‌کند (Halder et al. 2015). این سلول‌های حافظه، مشابه بردار حالت در یک مدل سیستم‌های دینامیکی سنتی هستند، که باعث می‌شوند LSTMها به‌طور بالقوه کاندیدای ایده‌آل برای مدل‌سازی سیستم‌های دینامیکی باشند (Wang et al. 2007). در روش پیشنهادی پس از مراحل استخراج ویژگی، شبکه عصبی LSTM سعی در انتخاب بهترین کلاس متناظر برای ریزآرایه ورودی می‌کند. شکل ٤ فلوچارت روش پیشنهادی را معرفی می‌نماید. هر مجموعه داده به‌صورت جداگانه وارد روش پیشنهادی می‌شود، در ادامه ۷۰٪ از داده‌ها به‌عنوان داده آموزشی و بقیه به‌عنوان داده آزمایش در نظر گرفته می‌شوند.

شبکه عصبی LSTM استفاده شده در روش پیشنهادی از دو لایه حذف تصادفی[1] برای کنترل بیش‌برازش[2]، دو لایه Dense و لایه نهایی برای انتخاب کلاس متناظر تشکیل شده است. لایه حذف تصادفی، مسأله را با عدم داشتن اطلاعات کافی در مورد میزان آموزش، حل می‌کند و مانع از اضافه کاری در آن می‌شود. حذف تصادفی بدین معنی است که برخی از واحدها (پنهان و قابل مشاهده) در شبکه عصبی به طور موقت از بین می‌روند، یعنی از شبکه خارج می‌شوند و در انتقال و ارسال پیشین شرکت نمی‌کنند. این کار باعث می‌شود که شبکه عصبی هر بار یک معماری متفاوت را انتخاب کند. هم‌چنین حذف تصادفی ترکیب سازگاری پیچیده‌ی واحدها در شبکه عصبی را کاهش می‌دهد، از آنجا که یک واحد نمی‌تواند به‌طور کامل بر حضور واحدهای خاص دیگر تکیه کند، ممکن است حذف شود، این باعث می‌شود که شبکه قوی‌تر باشد (Gal and Ghahramani 2016). پس از اعمال حذف تصادفی به یک شبکه عصبی عمیق، شبکه حاصل شبکه‌ای نازک‌تر است که شامل تمام واحدها می‌شود که از انقراض، جان سالم به در برده‌اند. نحوه عمل‌کرد این لایه در شکل ۳ نشان داده شده است.

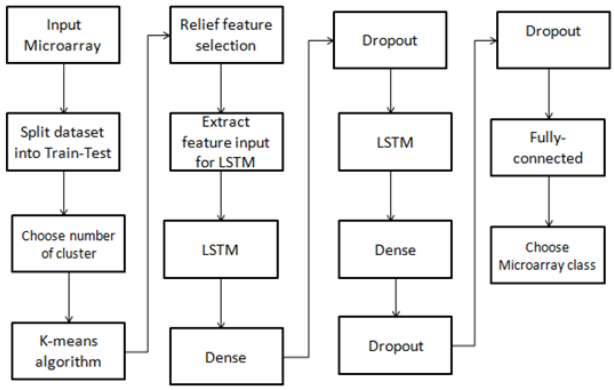
[1] Dropout
[2] Over-fitting

همان‌طور که در شکل ۳ مشاهده می‌شود، انتخاب واحدها برای حذف به‌صورت تصادفی انجام می‌گیرد و در ساده‌ترین حالت این بدان معنی است که هر واحد با یک احتمال ثابت p در شبکه نگه‌داری می‌شود. در این مقاله مقدار این احتمال، ۰/۵ تنظیم شده است که این مقدار با آزمون و خطا به‌دست آمده است. احتمال p، یک ابرپارامتر[1] اضافی است که شدت حذف تصادفی را کنترل می‌کند.

در آخر روش پیشنهادی یک لایه کاملاً متصل[2] که طبقه‌بندی اشیا (کلاس‌بندی) را در خروجی لایه‌های Dense ایجاد می‌کند، استفاده شده است. به‌طور خلاصه، می‌توان گفت که این لایه، یک طبقه‌بند شبکه عصبی استاندارد است که در انتهای یک استخراج‌کننده سطح بالا به‌کار برده می‌شود. مراحل روش پیشنهادی و لایه‌های استفاده شده در آن در شکل ٤ آمده‌اند.



شکل ۳- نحوه عملکرد لایه حذف تصادفی (Gal and Ghahramani 2016)



شکل ٤- فلوچارت روش پیشنهادی

نتایج

نرم‌افزار پیاده‌سازی شده با زبان پایتون ۳ برای روش پیشنهادی و ۹ الگوریتم Ridge، RandomForest، Elastic، Gradinatboosting، Extratree، Bagging، Adaboost، Lasso و Kneighbor بر روی دو مجموعه داده لوکیمیا و سرطان روده که در بخش قبلی توضیح داده شدند اجرا شده و نتایج حاصل با هم مقایسه شده‌اند. به‌منظور ارزیابی عملکرد روش پیشنهادی و مقایسه آن با سایر روش‌ها از ماتریس درهم ریختگی با معیارهای ارزیابی دقت[3]، درستی[4]، بازخوانی و معیار f1، که از رابطه‌های ۱ تا ٤ به‌دست می‌آیند، بهره گرفته شده است. معیارهای درستی، دقت و بازخوانی ارتباط مستقیمی با تشخیص صحیح TP، TN، FN و FP (که در جدول ۱ آمده است) دارد. هر چقدر درست‌ها صحیح تشخیص داده شوند و غلط‌ها در کلاس غلط قرار گیرند مقدار دقت بالاتر می‌رود و این روند برای بازخوانی نیز وجود دارد.

جدول ۱- معیارهای مقایسه‌ای طبقه‌بندی مدل پیشنهادی

TP	تعداد رکوردهایی که به‌درستی، مثبت تشخیص داده می‌شوند.
TN	تعداد رکوردهایی که به‌درستی، منفی تشخیص داده می‌شوند.
FP	تعداد رکوردهایی که به غلط، مثبت تشخیص داده می‌شوند.
FN	تعداد رکوردهایی که به غلط، منفی تشخیص داده می‌شوند.

$$Accuracy = (TP + TN) / All \quad (۱)$$

$$Precision = TP / (TP + FP) \quad (۲)$$

$$Recall = TP / (TP + FN) \quad (۳)$$

$$F1=2* Precision* Recall / Precision+ Recall \quad (٤)$$

همچنین برای ارزیابی خطای حاصل از روش پیشنهادی، معیارهای مجذور میانگین مربعات خطا[5] و میانگین قدر مطلق خطا[6] که به‌ترتیب از رابطه ۵ و ٦ محاسبه می‌شوند، استفاده شده‌اند.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (۵)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(e_i)^2} \quad (٦)$$

[1] Hyper parameter
[2] Fully connected
[3] Precision
[4] Accuracy
[5] Root Mean Square Error
[6] Mean absolute error

مقدار مربوط به e_t از رابطه ۷ به‌دست می‌آید که در آن، y_t مقدار واقعی و f_t مقدار پیش‌بینی شده توسط مدل است.

$$e_t = y_t - f_t \quad (۷$$

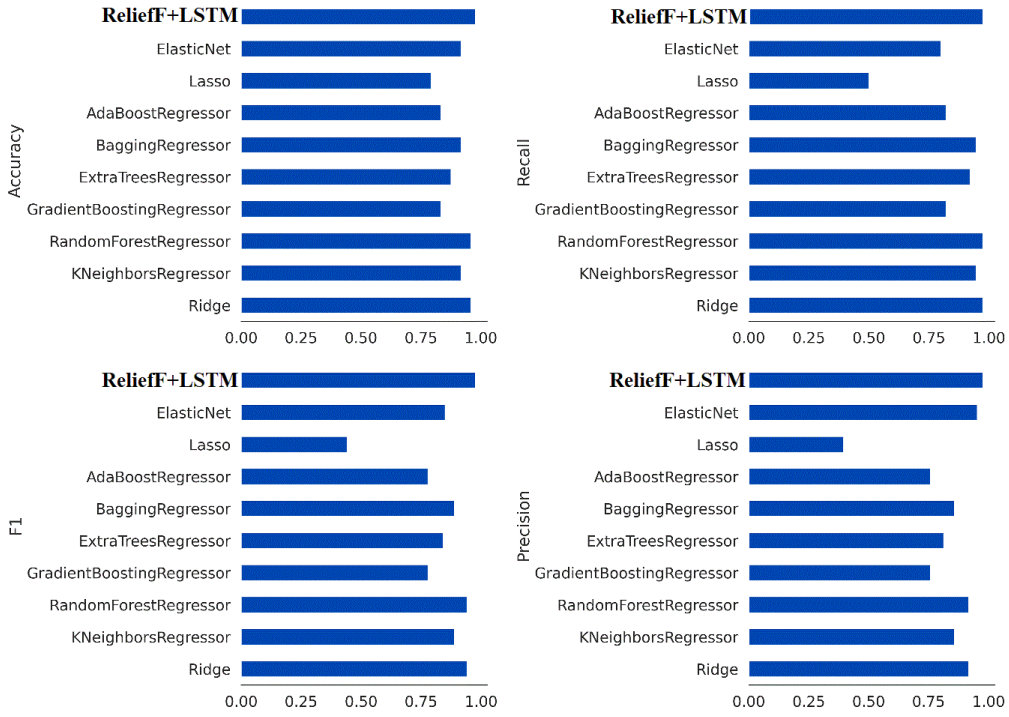
نتایج حاصل از اجرای نرم‌افزار توضیح داده شده برای روش پیشنهادی و ۹ الگوریتم ذکر شده و مقایسه‌ی آن‌ها با نتایج روش‌های قبلی، از لحاظ معیارهای درستی، دقت، بازخوانی، معیار F1 و نیز مجذور میانگین مربعات خطا و میانگین قدر مطلق خطا در جدول ۲ آمده‌اند.

همان‌طور که مشاهده می‌شود معیار درستی برای روش پیشنهادی ۹۷/۶۲ حاصل شده است که ۱/۷۹٪ بیشتر از بهترین مقدار روش‌های قبلی (روش Ridge) است و معیارهای دقت، ۲/۳۸٪، بازخوانی، ۰/۱۳٪ و f1 نیز ۳/٤٪ توسط روش پیشنهادی نسبت به روش‌های قبلی بهبود داده شده‌اند.

مقادیر معیارهای ارزیابی استفاده شده برای ماتریس درهم ریختگی در مجموعه داده لوسمی خون در شکل ۵ به نمایش در آمده‌اند.

جدول ۲- مقایسه روش پیشنهادی با ۹ الگوریتم دیگر در مجموعه داده لوسمی خون

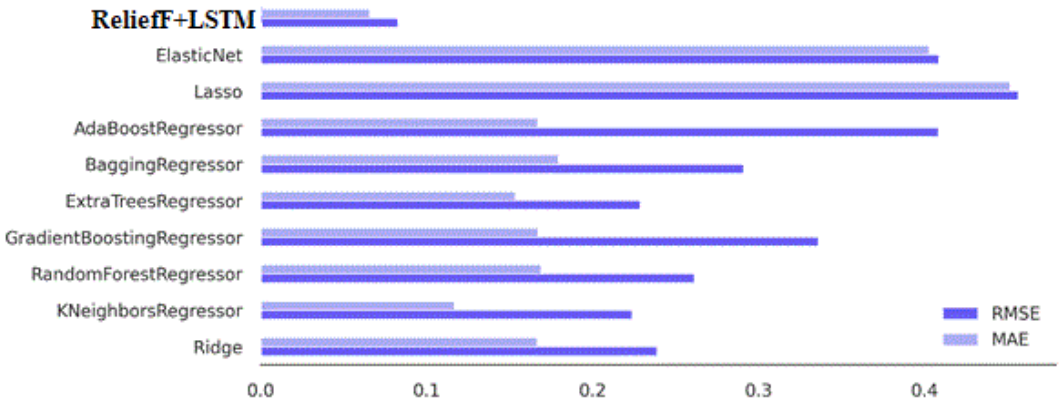
Method	Accuracy	Precision	Recall	F1	RMSE	MAE
Ridge	95.833	91.66	97.3684	94.10	0.238906	**0.1664446**
Random-Forest	91.661	85.7143	94.7368	88.889	0.244583	**0.155417**
Elastic	91.60	95.2381	80	85	0.408590	**0.402683**
Gradiantboosting	83.33	75.6303	82.1053	77.778	0.330755	**0.166845**
Extratree	95.883	91.667	97.3614	94.10	0.212642	**0.145833**
Bagging	87.5	80.5556	84.7368	82.396	0.273861	**0.166667**
Adaboost	91.6	85.71	94.73	88.89	0.288675	**0.083333**
Lasso	79.667	39.5833	50	45.6435	0.456435	**0.451389**
Kneighbor	91.664	85.7143	94.7368	88.889	0.223607	**0.116667**
Relief +LSTM	**97.62**	**97.62**	**97.5**	**97.505**	**0.0824**	**0.0656**



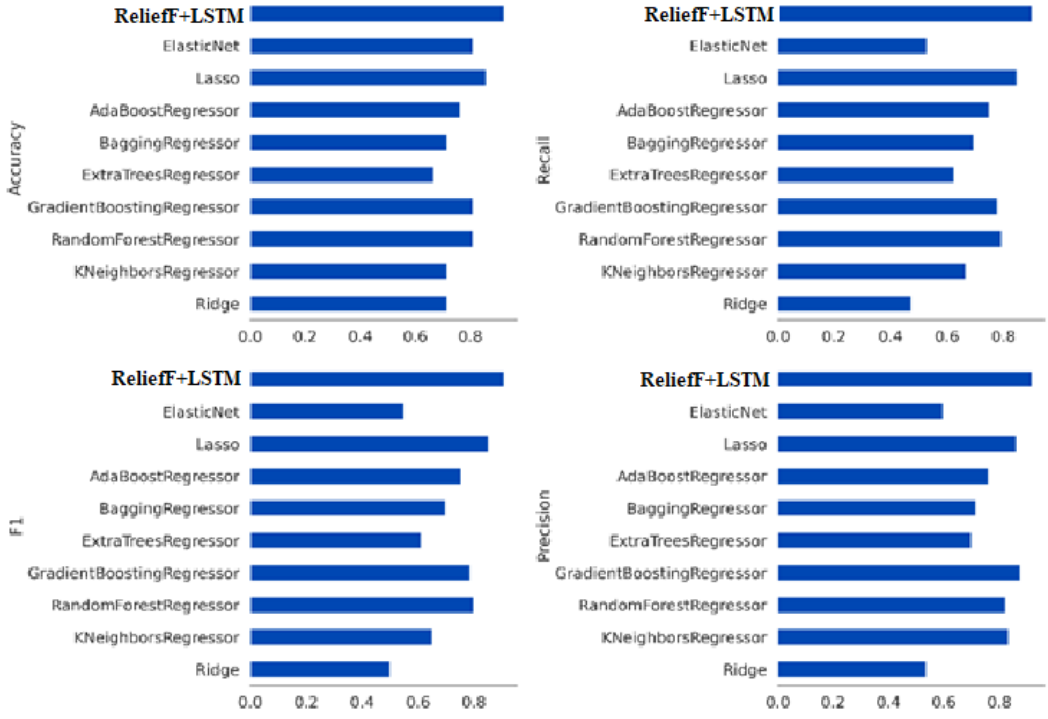
شکل ۵- ارزیابی روش پیشنهادی و مقایسه آن با ۹ الگوریتم دیگر در مجموعه داده لوسمی خون

روش پیشنهادی از نظر تعیین کلاس مثبت و درست واقعی بهبود خوب و با اختلاف قابل قبولی در مقابل ۹ الگوریتم دیگر بر روی مجموعه داده لوسمی خون داشته است. میزان خطای MAE و RMSE برای این روش محاسبه شده و در شکل ٦ نمایش داده شده است. همان‌طور که ملاحظه می‌شود مقدار این خطاها کمتر از ۰/۱ است و این امر به‌دلیل استفاده مناسب روش پیشنهادی از لایه‌های حذف تصادفی dropout و استخراج ویژگی‌های مؤثر است.

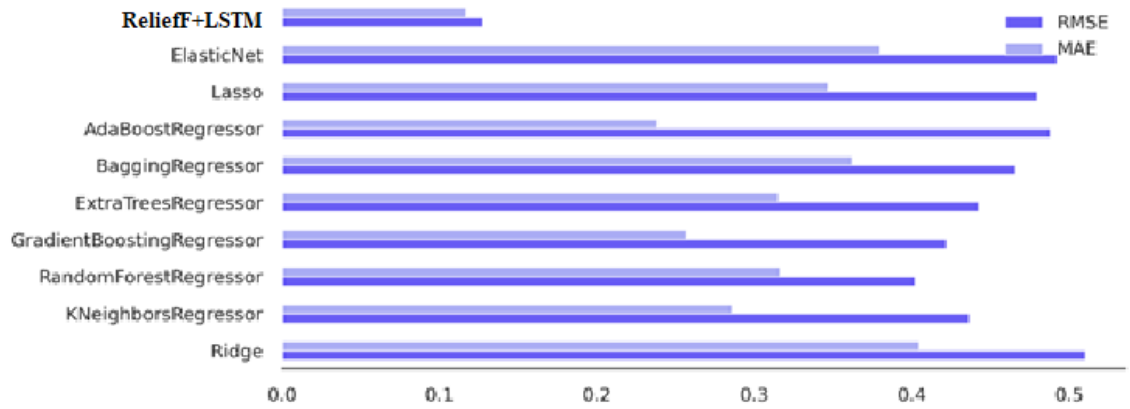
تمام آزمایش‌های انجام گرفته برای مجموعه داده لوسمی، برای مجموعه داده سرطان کلورکتال نیز با همان نرم‌افزار شبیه‌سازی شده در پایتون و با همان پارامترها انجام گرفته و نتایج حاصل در شکل‌های ۷ و ۸ و جدول ۳ گردآوری و نمایش داده شده‌اند.



شکل ٦- مقایسه میزان خطای مدل پیشنهادی با ۹ الگوریتم دیگر در مجموعه داده لوسمی خون



شکل ۷- مقایسه روش پیشنهادی با ۹ الگوریتم دیگر در مجموعه داده کلورکتال



شکل ۸- مقایسه میزان خطای روش پیشنهادی با ۹ الگوریتم دیگر در مجموعه داده کلورکتال

جدول ۳- مقایسه روش پیشنهادی با ۹ الگوریتم دیگر در مجموعه داده کلورکتال

Method	Accuracy	Precision	Recall	F1	RMSE	MAE
Ridge	71.4286	53.5714	47.22	50	0.510247	**0.404620**
Random-Forest	80.95	82.1429	79.8077	78.8077	0.416305	**0.336667**
Elastic	80.95	59.5238	52.778	54.8778	0.492103	**0.379719**
Gradiantboosting	71.42	74.3750	68.0556	67.8571	0.443705	**0.278430**
Extratree	66.67	69.8529	62.5	61.080	0.431983	**0.313333**
Bagging	76.1905	75.9615	75	75.2941	0.435343	**0.304762**
Adaboost	80.9524	80.556	80.556	80.556	0.436436	**0.190476**
Lasso	85.71	86.0577	84.722	85.1765	0.479494	**0.346538**
Kneighbor	71.42	83.33	66.67	65	0.436436	**0.285714**
+LSTM Relieff	**92.31**	**91.6650**	**90**	**90.824**	**0.12760**	**0.11670**

یکی از معیارهای مهم در سنجش میزان کارآمدی مدل را می‌توان معیار صحت در نظر گرفت. این معیار تمامی مقادیر صحیح را در جای درست و غلط را در جای غلط در نظر می‌گیرد، همان‌طور که در جدول ۳ مشاهده می‌شود، معیار صحت برای روش پیشنهادی در مجموعه داده سرطان کلورکتال ٦/٦٪ بهتر از نزدیک‌ترین الگوریتم (Lasso) است و همچنین مقدار معیارهای دقت، ٥/٦١٪، معیار بازخوانی، ٥/٢٨٪ و معیار f1 نیز ٥/٦٥٪ بهبود داده شده‌اند.

هرچقدر نرخ یادگیری در مدل کوچک‌تر انتخاب شود، امکان بهبود تابع خطا وجود دارد. برازش کنترل نشده باعث افزایش و کاهش ناگهانی خطا می‌شود که در مدل پیشنهادی کاهش خطا به‌صورت ثابت ادامه‌دار بوده است. از نظر میزان خطا، مقدار خطای RMSE برای این مجموعه داده ۰/۱۲ است که به نسبت دیگر الگوریتم‌ها ۰/۲۳ خطای کمتری داشته است. نزدیکی دیگر معیارهای صحت، دقت و بازخوانی نیز گواه عملکرد متناسب مدل پیشنهادی در این مجموعه داده را دارد.

بحث

یکی از راه‌های تشخیص سرطان استفاده از ریزآرایه و تحلیل داده‌های آن است که در مقایسه با روش اسکن که اشعه‌های تولید شده از دستگاه‌های تصویربرداری ضررهای زیادی برای انسان به‌همراه دارد، روش امن‌تری است. با توجه به این که توالی کامل ژنوم انسانی در ریزآرایه‌ها در دسترس است می‌توان از این ابزار و آنالیز داده‌ها جهت تبدیل و تفسیر آن‌ها در مورد تشخیص سرطان استفاده کرد. همان‌طور که ملاحظه شد در این مقاله از یک مدل مبتنی بر شبکه عصبی LSTM به‌همراه استخراج ویژگی با روش ReliefF و خوشه‌بندی K-means برای تشخیص سرطان‌های خون و کلورکتال بهره برده شده است، روش پیشنهادی با ۹ الگوریتم دیگر از نظر معیارهای دقت، صحت، بازخوانی و F1 و خطاهای RMSE و MAE مقایسه شده است که در همه معیارها بهتر از دیگر الگوریتم‌ها عمل کرده است.

داده‌های ریزآرایه شامل ژن‌های خیلی زیادی است که برخی از آن‌ها تکراری و برخی برای رسیدن به یک هدف، نامرتبط هستند.

در این مقاله ابتدا این ژن‌ها توسط الگوریتم K-means خوشه‌بندی می‌شوند و در ادامه توسط الگوریتم ReliefF ژن‌های مفید برای تشخیص سرطان انتخاب می‌شوند. در واقع یکی از دلایل بهبود دادن دقت و صحت توسط روش پیشنهادی همین انتخاب ژن‌های مفید از میان ژن‌های خیلی زیاد است که باعث یادگیری بهتر توسط مدل استفاده شده در شبکه یادگیری عمیق می‌شود. نرخ یادگیری در مدل شبکه یادگیری عمیق نیز کوچک انتخاب شده است تا امکان بهبود تابع خطا وجود داشته باشد.

با توجه به نتایج به‌دست آمده می‌توان از روش پیشنهادی در تشخیص زودهنگام سرطان‌ها و حتی جلوگیری از مرگ و میر ناشی از این بیماری خطرناک استفاده کرد. باتوجه به این که نتایج حاصل از این مطالعه می‌تواند به‌عنوان یک رهیافت مناسب در این زمینه ارائه شود، موارد زیر به‌عنوان کارهای آتی قابل پیشنهاد می‌باشد:

- در کارهای آتی، می‌توان از ترکیب Ensamble چندین شبکه عصبی عمیق و یادگیری عمیق مانند شبکه کانولوشن، LSTM و DNN و یا دیگر ترکیب‌ها استفاده نمود.
- می‌توان از شبکه GAN برای کار بر روی ریزآرایه‌ها در کارهای آتی استفاده نمود.
- برای بررسی قابلیت تعمیم روش پیشنهادی، می‌توان آن را در کارهای آینده بر روی دیگر مجموعه داده‌های موجود در زمینه ریزآرایه‌ها بررسی کرد.

منابع

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

Begum S, Chakraborty S, Banerjee A, Das S, Sarkar R, Chakraborty D (2018) Gene selection for diagnosis of cancer in microarray data using memetic algorithm. InIntelligent Engineering Informatics (441-449). Springer, Singapore.

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

Coustan-Smith E, Song G, Shurtleff S, Yeoh AE, Chng WJ, Chen SP, Rubnitz JE, Pui CH, Downing JR, Campana D (2018) Universal monitoring of minimal residual disease in acute myeloid leukemia. JCI insight 3.

Gal O, Auslander N, Fan Y, Meerzaman D (2019) Predicting Complete Remission of Acute Myeloid Leukemia: Machine Learning Applied to Gene Expression. Cancer informatics 18:1176935119835544.

Gal Y, Ghahramani, Z (2016) Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international conference on machine learning (1050-1059).

Ghosh M, Begum S, Sarkar R, Chakraborty D, Maulik U (2019) Recursive memetic algorithm for gene selection in microarray data. Expert Systems with Applications 116:172-85.

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

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

Harikiran J, Ramakrishna D, Avinash B, Lakshmi P, Kiran Kumar R (2014) A new method of gridding for spot detection in microarray images CEIS 5:25-33.

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

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

Lee SI, Celik S, Logsdon BA, Lundberg SM, Martins TJ, Oehler VG, Estey EH, Miller CP, Chien S, Dai J, Saxena A (2018) A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. Nature communications 9:1-3.

Medigue C (2015) ImaGene 9.0 - Leading-Edge Microarray Analysis Software, BioDiscovery, http://www.biodiscovery.com, Accessed Dec 1.

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

Sreedevi A, Jangamashetti D (2009) Automatically Locating Spots in DNA Microarray Image Using Genetic Algorithm without Gridding, International Association of Computer science and Information Technology - Spring Conference p: 178-181.

Tarca A.L,Romero R, Draghici S (2006) Analysis of microarray experiments of gene expression profiling 195:373-88.

Toloie Ashlaqi A, Mohsen Taheri, S (2010) Designing an expert system for suggesting the blood cancer treatment. Journal of Health Administration 13:41-50.
Wang T, Li T, Shao G, Wu S (2015) An improved Kmeans clustering method for cDNA microarray image segmentation 14:7771-81.
Wallack D (2015) Data Analysis with ScanAlyze, Department of Biology, Davidson College, Muhlenberg College, PA, http://www.bio.davidson.edu.
Wang L, Chu F, Xie W (2007) Accurate cancer classification using expressions of very few genes. IEEE/ACM 4:40-53.
Yin W, Chen T, Zhou S.X, Chakraborty A (2005) Background correction for cDNA microarray images using the TV+ Ll model 21:2410-6.