Journal of Soft Computing and Information Technology (JSCIT)
Babol Noshirvani University of Technology, Babol, Iran
Journal Homepage: jscit.nit.ac.ir
Volume 9, Number 2, Summer 2020, pp. 115-123
Received: 08/28/2018, Revised: 06/28/2019, Accepted: 01/25/2020

Intelligent detection of breast cancer with feature selection based on logistic regression and support vector machine Classification

Ziba Khandezamin[1], Marjan Naderan Tahan[2*] and Mohammad Javad Rashti[3]
1- Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
2*- Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
3- Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.
[1] z.khandezamin@gmail.com, [2*]m.naderan@scu.ac.ir, and [3]mohammad.rashti@scu.ac.ir

Corresponding author's address: Marjan Naderan Tahan, Department of Computer Engineering, Faculty of Engineering, Shahid Chamran University of Ahvaz, Ahvaz, Iran.

Abstract- Breast cancer is the most common cancer among women and the existence of a precise and reliable system for the diagnosis of benign or malignant of this cancer is essential. Nowadays, using the results of needle aspiration cytology, data mining and machine learning techniques, early diagnosis of breast cancer can be done with greater accuracy. In this study, we propose a method consisting of two steps: in the first step, to eliminate the less important features, logistic regression has been used to select more important features. In the second step, the Support Vector Machine (SVM) classification algorithm has been used with three different kernel functions for the diagnosis of benign and malignant samples. To evaluate the performance of the proposed method, two data sets, WBCD and WDBC have been used with investigation of several metrics such as precision, the Area Under the ROC (AUC), true positive rate, false positive rate, accuracy and the F-measure. The results show that using the logistic regression method, it is possible to select the more efficient features, such that the proposed method reaches 98.69% in terms of classification accuracy.

Keywords- breast cancer, machine learning, feature selection, logistic regression, support vector machine.

مجله علمی پژوهشی رایانش نرم و فناوری اطلاعات
دانشگاه صنعتی نوشیروانی بابل
صفحه مجله: jscit.nit.ac.ir
شناسگر اشیاء دیجیتالی (DOI):
جلد ۹، شماره ۲، تابستان ۱۳۹۹ ، صفحه ۱۱۵-۱۲۳
دریافت: ۱۳۹۷/۰۶/۰۶، بازنگری: ۱۳۹۸/۰۴/۰۷، پذیرش: ۱۳۹۸/۱۱/۰۵

تشخیص هوشمند سرطان پستان با انتخاب ویژگی مبتنی بر رگرسیون لجستیک و دسته‌بندی ماشین بردار پشتیبان

زیبا خنده‌زمین[۱]، مرجان نادران طحان*[۲]، محمدجواد رشتی[۳]
۱- گروه کامپیوتر، دانشکده مهندسی، دانشگاه شهید چمران اهواز، اهواز، ایران.
*۲- گروه کامپیوتر، دانشکده مهندسی، دانشگاه شهید چمران اهواز ،اهواز، ایران.
۳- گروه کامپیوتر، دانشکده مهندسی، دانشگاه شهید چمران اهواز، اهواز، ایران.
[1] z.khandezamin@gmail.com, [2*]m.naderan@scu.ac.ir, [3]mohammad.rashti@scu.ac.ir

* نشانی نویسنده مسئول: مرجان نادران طحان، استادیار، اهواز، بلوار گلستان، دانشگاه شهید چمران اهواز، دانشکده مهندسی، گروه کامپیوتر.

چکیده- سرطان پستان شایع‌ترین سرطان در میان زنان است و وجود یک سیستم دقیق و مطمئن برای تشخیص خوش‌خیم و یا بدخیم بودن توده سرطان ضروری است. امروزه با استفاده از نتایج سیتولوژی آسپیراسیون سوزنی، تکنیک‌های داده‌کاوی و یادگیری ماشین می‌توان شناسایی و تشخیص زود هنگام سرطان پستان را با دقت بالاتری انجام داد. در این مقاله روشی پیشنهاد شده است که شامل دو مرحله است: در مرحله اول برای حذف ویژگی‌های کم اهمیت‌تر، از رگرسیون لجستیک استفاده شده است تا ویژگی‌های مهم‌تر انتخاب شوند. در مرحله دوم، از الگوریتم طبقه‌بندی ماشین بردار پشتیبان (SVM) با سه هسته‌ی متفاوت برای تشخیص خوش خیم و بدخیم بودن نمونه‌ها استفاده شده است. برای ارزیابی کارایی روش پیشنهادی از دو مجموعه داده WBCD و WDBC و معیارهای دقت، ناحیه زیر نمودار ROC (AUC)، نرخ مثبت حقیقی، نرخ مثبت کاذب، صحت و معیار F بررسی شده‌اند. نتایج نشان می‌دهد که با استفاده از روش رگرسیون لجستیک می‌توان انتخاب ویژگی موثرتری انجام داد، به‌گونه‌ای که روش پیشنهادی از نظر دقت طبقه‌بندی به دقت ۹۸/۶۹٪ می‌رسد.

واژه‌های کلیدی: سرطان پستان، یادگیری ماشین، انتخاب ویژگی، رگرسیون لجستیک، ماشین بردار پشتیبان.

۱- مقدمه

سرطان پستان شایع ترین سرطان در میان زنان است. این بیماری به شدت ناهمگن است و در اثر تاثیر متقابل عامل های خطر وراثتی و محیطی ایجاد می شود [۱]. این سرطان از رشد نامنظم سلولها در بافت پستان ایجاد می‌شود و تومورهایی را تشکیل می‌دهد که می‌توانند خوش‌خیم (غیرسرطانی) یا بدخیم (سرطانی) باشند [۲]. تقریباً ۲۵ درصد از مرگ و میرهای ناشی از سرطان پستان بین سنین ۴۹-۴۰ سال مشاهده می‌شود. اگرچه شیوع این بیماری در سنین قبل از ۲۵ تا ۳۰ سالگی نادر است اما بروز این سرطان در سنین کمتر حتی در سن جوانی نیز گزارش شده است. براساس آمار سازمان بهداشت جهانی، از هر ۸ تا ۱۰ نفر یک زن به سرطان پستان دچار می‌شود. بر اساس آمارهای موجود در ایران از هر ۱۰ تا ۱۵ زن، احتمال ابتلای یک زن به سرطان پستان وجود دارد. پژوهشگران میزان بالای مرگ و میر زنان بر اثر سرطان پستان را ناشی از تشخیص دیر هنگام این بیماری می‌دانند [۱].

سه روش کلاسیک برای تشخیص سرطان سینه موجود است: معاینه فیزیکی، ماموگرافی و بیوپسی؛ روش آخر خود ممکن است یکی از چهار روش آسپیراسیون سوزن ظریف[۱](FNA)، بیوپسی سوزن

هسته، بیوپسی جراحی ویا بیوپسی گره لنفاوی باشد. ماموگرافی یکی از روش‌هایی است که برای تشخیص سرطان سینه بسیار استفاده می‌شود. تفسیر ماموگرافی مستلزم رادیولوژیست‌های بسیار ماهر است زیرا رادیولوژیست ها تفسیرهای مختلفی برای همان ماموگرافی گزارش می‌دهند [۶]. دقت ماموگرافی بین ۶۸٪ تا ۷۹٪ متغیر است. وقتی ماموگرافی توده را تشخیص داد، بیوپسی (بافت برداری) برای تعیین بدخیمی آن استفاده می‌شود. دقت بیوپسی جراحی نزدیک به صدرصد است اما پرهزینه، تهاجمی، زمان بر و دردناک است.

در مقابل از FNA به طور گسترده در تشخیص سرطان سینه استفاده می‌شود. دقت FNA با تفسیر بصری بین ۳۵٪ تا ۹۵٪ بسته به تجربه پزشک متغیر است [۳]. آزمایش آسپیراسیون سوزنی (FNA) روش سرپایی، کم هزینه، آسان و سریع است [۴]. در این روش مایع استخراج شده از بافت پستان برای بررسی خصوصیات سیتولوژی در زیر میکروسکوپ قرار می‌گیرد. بعد از استخراج خصوصیات سیتولوژی بیمار، باید بتوان خوش‌خیم یا بدخیم بودن توده را تشخیص داد [۲]. در مواردی که ممکن است پزشک در تشخیص خوش‌خیم و بدخیم بودن بیماری دچار تردید شود تکنیک‌های داده کاوی و یادگیری ماشین می‌توانند تاثیر بسزایی در این نوع تصمیم-گیری‌ها داشته باشند.

روش‌های هوشمند مختلفی در حوزه طبقه‌بندی کردن داده‌های سرطانی انجام گرفته شده است. به عنوان مثال، در [۵]، یک دسته-بند بیزین وزن‌دار برای دسته‌بندی ارائه شده است که به دقت ٪۹۸/۵۴ بر روی پایگاه داده WBCD دست یافته است.

در [۶]، نویسندگان به بررسی روش‌های SVM، C4.5، K-NN و NB بر روی پایگاه داده WBCD پرداختند که نتایج آزمایشات آنها نشان داد که روش SVM دارای دقت ٪۹۷/۱۳ ، روش C4.5 دارای دقت ٪۹۵/۱۳، روش K-NN دارای دقت ٪۹۵/۲۷ و روش NB دارای دقت ٪۹۵/۹۹ است.

در [۷]، نویسندگان با استفاده از روش EM[۲] داده‌ها را به گروه‌های خوشه بندی می‌کند و سپس از روش CART[۳] برای تولید قوانین فازی برای طبقه بندی استفاده کرده‌اند. همچنین برای غلبه بر مسئله‌ی چندخطی[۴] از روش تحلیل مولفه اساسی (PCA) در سیستم پیشنهادی استفاده کردند که در نهایت به دقت ٪۹۳/۲ دست یافتند.

در [۸]، محققان از سیستم استنتاج فازی ممدانی استفاده کردند و آزمایشات خود را بر روی پایگاه داده WBCD انجام دادند که به دقت ٪۹۳/۶ دست یافتند.

در [۹]، نویسندگان به بررسی روش Gaussian Naïve Bayes بر روی پایگاه داده WBCD پرداختند. آنها در کار خود از انتخاب ویژگی هم استفاده کردند به همین منظور از روش تحلیل مؤلفه‌ی خطی (LDA) برای انتخاب ویژگی‌های موثرتر استفاده کردند و به دقت ٪۹۶/۶ دست یافتند.

در [۱۰]، محققان از روش Pearson Correlation Coefficient برای انتخاب ویژگی و از سه دسته بند SVM، Naïve Bayes و Ensemble برای طبقه‌بندی استفاده کردند. همچنین در روشی جدا از PCA برای استخراج ویژگی استفاده کردند و بعد از آن سه دسته-بند فوق را برای کلاس‌بندی بکار بردند. در نهایت به دقت ٪۹۷/۳۹ برای دسته‌بند Naïve Bayes دست یافتند.

در [۱۱]، نویسندگان از ترکیب شبکه عصبی و الگوریتم ژنتیک استفاده و نتایج آزمایشات خود را بر روی پایگاه داده WBCD بررسی کردند. آنها به دقت ٪۹۹/۶۶ رسیدند.

در [۱۲]، پژوهشگران از سه روش Information Gain، PCA و Relief برای انتخاب ویژگی استفاده کردند و در مرحله طبقه‌بندی از Naïve Bayes، Random Forest و درخت تصمیم J48 استفاده کردند. آنها نشان دادند که با ترکیب NB و J48 به دقت ٪۹۷ می‌توان رسید.

در [۱۳]، محققان از شبکه عصبی پیشخورد چندلایه بر روی پایگاه داده WBCD استفاده کردند و به دقت ٪۹۸ دست یافتند.

در این مقاله، برای کاهش ویژگی‌های غیرموثر از الگوریتم رگرسیون لجستیک استفاده می‌شود که تاکنون در کارهای پیشین انجام نشده است. با استفاده از این الگوریتم که خطی و کم هزینه است می‌توان انتخاب ویژگی مؤثر انجام داد. درواقع، در این الگوریتم، ویژگی‌ها وزن‌دهی می‌شوند و سپس با استفاده از وزن‌های داده شده، ویژگی‌های غیر موثر حذف و ویژگی‌های موثرتر حفظ می‌شوند. سپس برای طبقه‌بندی نمونه‌ها، از روش ماشین بردار پشتیبان با سه کرنل متفاوت خطی، چندجمله‌ای و شعاعی استفاده می‌شود. همچنین در نهایت از دو مجموعه داده برای ارزیابی بهره گرفته می‌شود که در کارهای پیشین این چنین نبوده است و فقط از یک مجموعه داده استفاده می‌شده است.

ادامه‌ی ساختار مقاله بدین صورت است که در بخش دوم روش پیشنهادی ارائه می‌شود. در بخش سوم نتایج شبیه‌سازی و مقایسه با روش‌های پیشین ارائه شده است. در نهایت در بخش چهارم، نتیجه‌گیری و پیشنهادهایی برای کارهای آتی بیان خواهد شد.

۲- روش پیشنهادی

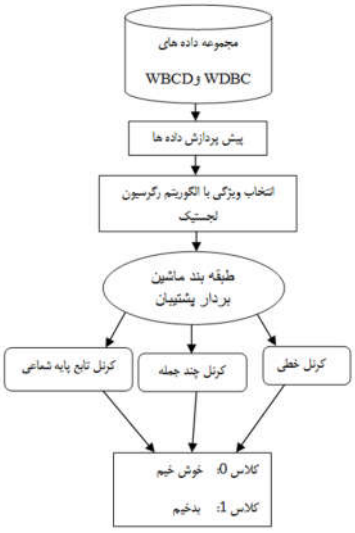
در این مطالعه، مدلی برای طبقه‌بندی بین دو نوع تومور خوش‌خیم و بدخیم برای تشخیص سرطان سینه با استفاده از مجموعه داده‌های WBCD و WDBC ارائه شده است. در این مدل از رگرسیون لجستیک برای انتخاب زیرمجموعه‌ای از ویژگی‌های که برای تشخیص مناسب‌تر هستند استفاده می‌شود. راه حل مورد استفاده برای آموزش، طبقه بند ماشین بردار پشتیبان با سه تابع کرنل متفاوت، تابع پایه شعاعی و خطی و چند جمله‌ای، است. در نهایت عملکرد روش پیشنهادی را با استفاده از معیارهای دقت، ناحیه زیر نمودار ROC (AUC)، نرخ مثبت حقیقی، نرخ مثبت کاذب، صحت و معیار F بررسی می‌کنیم. شکل ۱، فلوچارت روش پیشنهادی را نشان می‌دهد.

۲-۱- مجموعه داده

در این مقاله، از دو پایگاه داده مختلف با تعداد ویژگی‌های متفاوت استفاده شده است تا قدرت روش انتخاب ویژگی بیشتر قابل بررسی باشد. پایگاه داده‌های WBCD و WDBC که در اکثر کارهای پیشین نیز استفاده شده‌اند برای این منظور انتخاب شده‌اند. این دو مجموعه داده از مخزن یادگیری ماشین UCI گرفته شده‌اند [۱۴]. جداول ۱ و ۲ ویژگی‌های نمونه‌های این دو مجموعه داده را نشان می‌دهد. مجموعه داده‌های پایگاه WBCD شامل اطلاعات ۶۹۹ بیمار با ۱۰ ویژگی است که ویژگی اول شماره شناسه پرونده بیمار و بقیه‌ی ویژگی‌ها نتایج کمی آزمایش آسپیراسیون سوزنی برای هر بیمار است. هر نمونه با برچسب خوش‌خیم یا بدخیم مشخص می‌گردد. ۲۴۱ نمونه بدخیم و ۴۵۸ نمونه خوش‌خیم هستند.

مجموعه داده‌های پایگاه WDBC شامل اطلاعات ۵۶۹ بیمار با ۳۱ ویژگی است که ویژگی اول شماره شناسه پرونده بیمار و ۳۰ویژگی باقیمانده از تصویر دیجیتالی آزمایش آسپیراسیون سوزنی توده پستان به دست آمده که این ویژگی‌ها خصوصیات هسته سلول در تصویر را بیان می‌کنند. در واقع برای هر ویژگی، میانگین،خطای استاندارد و بزرگترین مقدار (میانگین سه تا از بزرگترین مقادیر) این ویژگی‌ها محاسبه شده و به این ترتیب ۳۰ ویژگی با مقدار عددی حقیقی برای هر نمونه به دست می آید. هر یک از نمونه ها با یک برچسب خوش خیم یا بدخیم مشخص می‌گردند. از ۵۶۹ نمونه مذکور، ۳۵۶ نمونه دارای برچسب خوش خیم و ۲۱۳ نمونه دارای برچسب بدخیم هستند [۳].

همان‌طور که از جداول مشاهده می‌شود، مجموعه داده WBCD دارای ۹ ویژگی و مجموعه داده WDBC دارای ۳۱ ویژگی است.این موضوع سبب می‌شود که در مرحله‌ی انتخاب ویژگی در بخش بعدی، ورودی تابع دو نوع متفاوت باشد.



شکل۱: فلوچارت روش پیشنهادی

جدول۱: ویژگی‌های پایگاه داده WBCD

شماره ویژگی	نام ویژگی	دامنه مقادیر ویژگی
۱	ضخامت توده	۱-۱۰
۲	یکنواختی اندازه سلول	۱-۱۰
۳	یکنواختی شکل سلول	۱-۱۰
۴	چسبندگی حاشیه ای	۱-۱۰
۵	اندازه سلول های اپتلیال	۱-۱۰
۶	Bare nuclei	۱-۱۰
۷	Bland chromatin	۱-۱۰
۸	Normal nucleoli	۱-۱۰

جدول۲: ویژگی‌های پایگاه داده WDBC

شماره ویژگی	توصیف ویژگی
۱	شعاع (میانگین فاصله از مرکز تا نقاط روی محیط)
۲	بافت (انحراف معیار مقادیر خاکستری پیکسلها)
۳	محیط
۴	مساحت
۵	همواری (تغییرات محلی در طول شعاع‌ها)
۶	فشردگی (محیط به توان ۲ تقسیم بر مساحت منهای یک)
۷	تحدب(میزان شدت قسمتهای محدب کانتور)
۸	نقاط محدب (تعداد قسمتهای محدب کانتور)
۹	تقارن
۱۰	ابعاد فراکتالی (“coastline approximation” – 1)

۲-۲- رگرسیون لجستیک

هدف اصلی انتخاب ویژگی کاهش تعداد ویژگی‌های استفاده شده در طبقه‌بندی است، بطوری که دقت طبقه‌بندی در حد قابل قبولی

حفظ شود. انتخاب ویژگی، اندازه ورودی داده‌ها را به مدل پیش‌بینی کاهش می‌دهد و همچنین زمان اجرا را کاهش می‌دهد. روش‌های انتخاب ویژگی به دو گروه با ناظر و بدون ناظر تقسیم می‌شوند. یک روش انتخاب ویژگی با ناظر، طیف وسیعی از زیرمجموعه ویژگی‌ها را با استفاده از یک تابع ارزیابی یا معیار، ارزیابی می‌کند تا فقط آن ویژگی‌هایی که مربوط به کلاس‌های تصمیم‌گیری داده‌ها هستند را انتخاب کند [۱۵].

روش‌های سنتی انتخاب ویژگی به دو گروه فیلتر و رَپِر تقسیم می‌شوند. حذف یا انتخاب ویژگی در روش فیلتر بر اساس معیارهایی همچون اندازه‌گیری آماری، وزن‌دهی ویژگی و اطلاعات متقابل تعیین می‌شود. از طرف دیگر، ارزیابی عملکرد ویژگی مستقل از هر مدل یادگیری است از این رو برای پیاده سازی آسان و سریع‌تر است [۱۶].

روشی که برای انتخاب ویژگی در این مقاله استفاده شده است، روش رگرسیون لجستیک است که جزء روش‌های یادگیری با ناظر است. برای حذف یا انتخاب ویژگی نیز از وزن‌دهی ویژگی‌ها استفاده می‌شود. در رگرسیون لجستیک [۱۷] از تابع سیگموئید، در (۱)، برای دسته‌بندی استفاده می‌شود که تضمین می‌کند خروجی در بازه [0-1] است.

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1+ e^{-\theta^T x}} \qquad (۱)$$

ورودی این تابع با توجه به اینکه از دو پایگاه داده استفاده شده است، که برای یک مجموعه داده ۹ ویژگی و برای مجموعه داده دوم ۳۰ ویژگی وجود دارد، به صورت زیر است:

$$\theta^T x = \theta_0 + \theta_1 x_1 + \cdots + \theta_9 x_9 \qquad (۲)$$

$$\theta^T x = \theta_0 + \theta_1 x_1 + \cdots + \theta_{30} x_{30} \qquad (۳)$$

که در آن θ_0پارامتر بایاس است،θ_i که برای فرمول (۲) $1 \le i \le 9$ است و برای فرمول (۳) $1 \le i \le 30$است، پارامترهای مجهول هستند که باید محاسبه شوند و x_i، مقادیر ویژگی‌ها است. تابع هدف بصورت فرمول (۴) تعریف می‌شود:

$$Cost(h_\theta(x), y) = \begin{cases} -\log(h_\theta(x)) if y = 1 \\ -\log(1 - h_\theta(x)) \quad 0 therwise \end{cases} \qquad (۴)$$

که در آن، y برچسب کلاس است و $y \in \{0,1\}$. می‌توان تابع هزینه در فرمول (۴) را به شکل ساده شده‌ی زیر نوشت:

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost\left(h_\theta(x^i), y^i\right) = \qquad (۵)$$

$$-\frac{1}{m}\left[\sum_{i=1}^{m} y^i \log h_\theta(x^i) + (1 - y^i)\log(1 - h_\theta(x^i))\right]$$

ورودی تابع، θها (پارامترهای مجهول) هستند، متغیر m در این تابع تعداد نمونه‌ها است.برای بدست آوردن پارامترهای مجهول هر کلاس، برچسب آن کلاس، یک و بر چسب بقیه کلاس‌ها صفر در نظر گرفته می‌شود، به همین دلیل y دارای مقادیر صفر و یک است. می‌بایست مقدارپارامتر θ بگونه‌ای محاسبه شود که تابع هزینه در فرمول (۵) کمینه شود. تابع هزینه محدب و مشتق‌پذیر است با توجه به محدب بودن تابع هدف می‌توان از روش گرادیان نزولی استفاده کرد. بنابراین:

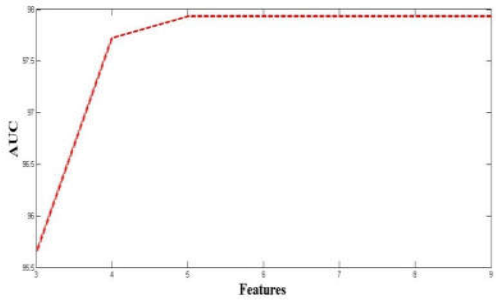
$$\theta_J = \theta_j + \alpha\left(y^i - h_\theta(x^i)\right)x_\theta(i) \qquad (۶)$$

پس از کلاس‌بندی مجموعه داده‌ها توسط الگوریتم رگرسیون لجستیک، ۹ مقدار تتا یا ضریب برای ۹ ویژگی مجموعه داده WBCD در بازه‌ی [۰/۱۰,۴/۸۲] و ۳۰ مقدار تتا برای مجموعه داده WDBC در محدوده‌ی [-۴/۲۰,۱۰/۳۷] بدست می‌آید. سپس نمودار تعداد ویژگی‌ها بر حسب مقدار AUC[۵] رسم می‌شود، ویژگی‌های که ضریب آنها دارای مقادیر بزرگتر هستند زودتر به نمودار اضافه می‌شوند. که این نمودار در شکل‌های ۲ و ۳ مشاهده می‌شوند.

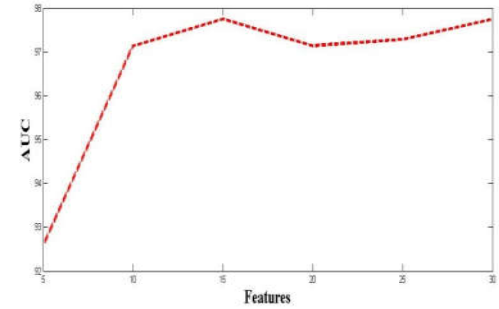
در این نمودارها محور افقی نشان دهنده تعدادویژگی‌هاست که از ابتدای محور افقی ویژگی‌هایی که ضرایب بزرگتری دارند تعداد آنها بر روی محور نشان داده شده است و محور عمودی مقدار به دست آمده برای AUC بر حسب تعداد ویژگی‌هاست. طبق این نمودارها کمترین تعداد ویژگی که مقدار AUC قابل قبولی دارند انتخاب می‌شوند. مثلا در شکل ۳ مقدار AUC بدست آمده برای ۱۵ویژگی و ۳۰ ویژگی به یک میزان بوده است، به همین دلیل کمترین تعداد ویژگی که برای این کلاس مفید است ۱۵ ویژگی است. با توجه به نمودارها ویژگی‌های که برای هر مجموعه داده بدست می‌آید در جدول ۳ نشان داده شده است.

جدول ۳: تعداد و شماره ویژگی‌های انتخاب شده برای هر مجموعه داده

مجموعه داده	تعداد ویژگی	شماره ویژگی
WBCD	۵	۱-۳-۶-۷-۹
WDBC	۱۵	-۲۲-۲۴-۲۵-۲۷-۲۸-۲۹-۳۰ ۲-۷-۸-۱۱-۱۳-۱۴-۱۵-۲۱



شکل ۲: نمودار AUC بر حسب مقادیر مرتب شده ویژگی‌ها با استفاده از ضرایب ویژگی‌ها برای مجموعه WBCD



شکل ۳: نمودار AUC بر حسب مقادیر مرتب شده ویژگی‌ها با استفاده از ضرایب ویژگی‌ها برای مجموعه WDBC

۲-۳- ماشین بردار پشتیبان

ماشین بردار پشتیبان یکی از تکنیک‌های طبقه‌بندی با ناظر است که به دلیل قابلیت‌های برتر آن بطور گسترده‌ای مورد استفاده قرار می‌گیرد. این طبقه‌بند می‌تواند مشکلات طبقه‌بندی غیرخطی را با نگاشت داده‌های آموزشی اصلی به یک فضای ویژگی با ابعاد بالا با یک تابع هسته را رفع کند. پس از آن بهترین ابرصفحه جداساز را تعیین می‌کند که به عنوان مرز جدایی داده‌های دو کلاس عمل می‌کند. وقتی یک دسته بند دودویی برای پیش بینی بین دوکلاس خوش‌خیم و بدخیم استفاده می‌شود، این ابرصفحه باید حاشیه‌هایی که بعنوان فاصله بین نزدیکترین نقاط آموزشی شناخته شده هستند، بیشینه کند. در این مقاله، سه تابع هسته استفاده شده است که عبارتند از: تابع پایه شعاعی، چند جمله‌ای و خطی.

در صورتی که داده ها بصورت خطی مجزا از هم باشند، ماشین بردار پشتیبان به ماشین های خطی برای تولید یک سطح بهینه که داده ها را بدون خطا و با حداکثر فاصله میان صفحه و نزدیک‌ترین نقاط آموزشی (بردارهای پشتیبان) تفکیک می‌کند. بنابراین در این حالت، قواعد تصمیم‌گیری که تعریف می‌شود توسط یک صفحه بهینه که طبقات تصمیم‌گیری باینری را تفکیک می کند، به صورت (۷) است:

$$Y = sign(\sum_{i=1}^{N} y_i \alpha_i (X \times X) + b)) \qquad (۷)$$

که در آن Y خروجی رابطه، y_i ارزش طبقه نمونه آموزشی X_i و پارامترهای a_i و b_i تعیین کننده ابر صفحه هستند. در صورتی که داده‌ها بصورت خطی قابل تفکیک نباشند رابطه (۷) به رابطه (۸) تغییر می‌یابد:

$$Y = sign(\sum_{i=1}^{N} y_i a_i K(X \times X_i) + b) \qquad (۸)$$

که در آن، تابع $K(X, X_i)$ تابع کرنلی است که برای ایجاد ماشین‌های با انواع مختلفی از سطوح تصمیم‌گیری غیرخطی در فضای داده‌ها، ضرب های داخلی تولید می‌کند. کرنل‌های استفاده شده در این پژوهش در جدول ۴ نشان داده شده است.

در نهایت ورودی طبقه‌بند ماشین بردار پشتیبان، دیتاست با ویژگی-های کمتر است که ویژگی‌های غیرضروری آن حذف شده‌اند.

جدول ۴: توابع کرنل

تابع	کرنل
$(1 + X_i X_j)^d$	چند جمله ای
$X_i X_j$	خطی
$e^{-\frac{\|X_i - X_j\|}{2\sigma^2}}$	پایه شعاعی

۳- شبیه‌سازی، نتایج و ارزیابی

برای شبیه سازی روش پیشنهادی از نرم افزار متلب ۲۰۱۳ استفاده شده است. اولین مرحله در ایجاد هر مدلی بر اساس تکنیک‌های داده کاوی، مرحله پیش پردازش می‌باشد که جهت بهبود کیفیت داده‌های واقعی برای داده کاوی لازم است.

۳-۱- پیش‌پردازش و معیارهای ارزیابی

ابتدا هر دو دسته مجموعه داده WBCD و WDBC با استفاده از (۹) نرمال شده‌اند [۱۳]، به طوریکه داده‌ها حداقل 0 و حداکثر ۱ هستند.نرمال سازی داده‌ها منجر به کسب نتایج بهتر می‌شود. بنابراین:

$$normvalue = \frac{\text{realvalue} - minvaludataset}{maxvaluedataset - \text{minvaluedataset}} \qquad (۹)$$

سپس نمونه‌های این پژوهش در دو گروه آموزش و آزمون قرار می‌گیرند و روی داده‌های گروه اول فرایند یادگیری انجام می‌شود.۸۰ درصد داده ها بعنوان مجموعه داده های آموزش و ۲۰ درصد آن، بعنوان مجموعه داده آزمون در نظر گرفته شده است. برای ارزیابی کارایی روش پیشنهادی معیارهای دقت، ناحیه زیر نمودار ROC (AUC)، نرخ مثبت حقیقی، نرخ مثبت کاذب، صحت و معیار F بررسی شده‌اند. این معیارها مبتنی بر معیارهای پایه‌ای زیر هستند:

- مثبت درست (TP): تعداد نمونه‌های مثبت که درست تشخیص داده شده‌اند.
- منفی درست (TN): تعداد نمونه‌های منفی که درست تشخیص داده شده‌اند.
- مثبت غلط (FP): تعداد نمونه‌های منفی که مثبت تشخیص داده شده‌اند.
- منفی غلط (FN): تعداد نمونه‌های مثبت که منفی تشخیص داده شده‌اند.

بنابراین، با استفاده از معیارهای پایه‌ای تعریف شده در بالا، معیارهای مورد استفاده عبارتند از:

- دقت: به درصد طبقه بندی درست توسط طبقه بند SVM آموزش دیده بر روی مجموعه تست اشاره دارد که توسط (۱۰) بیان می‌شود.
- ناحیه زیر منحنی ROC برای تمایز داده‌ها در کلاس‌های داده شده (مانند خوش‌خیم و بدخیم) استفاده می‌شود. هدف تعیین نقطه منفصل برای طبقه‌بند است که بیشترین تعداد مثبت حقیقی و تعداد کم مثبت‌های کاذب را بدست آورد.
- حساسیت (نرخ مثبت حقیقی): عبارت است از مقداری برای مشخص کردن توانایی سیستم در تشخیص و دسته‌بندی موارد واقعاً بیمار (سرطانی) که سیستم آنها را صحیح و سرطانی تشخیص می‌دهد که توسط (۱۱) بیان می‌شود.
- صحت: به معنی نسبت تعداد نمونه‌های صحیح طبقه‌بندی شده توسط طبقه‌بند به کل تعداد نمونه‌ها (که طبقه بند چه به صورت صحیح و چه به صورت غلط طبقه بندی کرده است) و توسط (۱۲) محاسبه می‌شود.
- معیار F: پارامتر مناسب برای ارزیابی کیفیت کلاس‌بندی است و همچنین توصیف کننده میانگین هارمونیک بین دو کمیت صحت و حساسیت، که توسط (۱۳) بیان می‌شود [۲۰].

$$\text{دقت} = \frac{TP + TN}{TP + FP + TN + FN} \quad (۱۰)$$

$$\text{حساسیت} = \frac{TP}{TP + FN} \quad (۱۱)$$

$$\text{صحت} = \frac{TP}{TP + FP} \quad (۱۲)$$

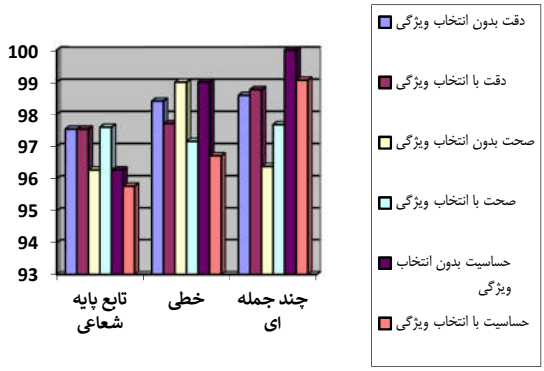
$$F_{\text{معیار}} = 2 \times \frac{\text{حساسیت} \times \text{صحت}}{\text{حساسیت} + \text{صحت}} \quad (۱۳)$$

۳-۲- ارزیابی کارایی روش انتخاب ویژگی

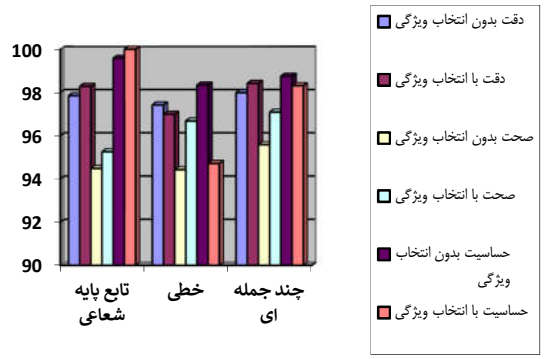
از آنجا که انتخاب ویژگی‌های مؤثر در تشخیص سرطان پستان، در هیچ کدام از کارهای پیشین دیگر انجام نشده است، برای ارزیابی کارایی روش انتخاب ویژگی، می‌بایست کارایی روش به ازای برخی از معیارها یک بار بدون استفاده از الگوریتم انتخاب ویژگی و یک بار با از استفاده از الگوریتم انتخاب ویژگی بررسی شود. برای این منظور، سه معیار دقت، صحت و حساسیت انتخاب شدند و نتایج روی دو مجموعه داده بدست آمد که در شکل‌های ۴ و ۵ نشان داده شده است. در این دو شکل، نتایج دسته‌بندی به ازای هر سه تابع هسته‌ی خطی، چندجمله‌ای و شعاعی ارائه شده است.

همان‌گونه که مشاهده می‌شود، نتایج الگوریتم انتخاب ویژگی، در مورد مجموعه داده WBCD نتایج بهتری نسبت به WDBC دارد (شکل ۵ نسبت به شکل ۴). از شکل ۴ می‌توان دریافت که معیار دقت و صحت به ازای کرنل چند جمله ای با روش انتخاب ویژگی پیشنهادی افزایش یافته است. همچنین شکل ۵ نیز نشان می‌دهد که دقت کرنل چند جمله ای و کرنل تابع پایه شعاعی با روش انتخاب ویژگی پیشنهادی افزایش یافته است. علاوه بر این معیار صحت به ازای هر سه تابع کرنل، با روش انتخاب ویژگی پیشنهادی افزایش محسوسی را نشان می‌دهد (بجز یک مورد تابع کرنل چندجمله‌ای در مجموعه داده WDBC).

علاوه بر آن، نتایج به ازای تابع هسته‌ی چندجمله‌ای و استفاده از الگوریتم انتخاب ویژگی، نتایج ملموس‌تری نسبت به دو تابع هسته‌ی دیگر دارد.



شکل ۴: مقایسه بین نتایج استفاده از انتخاب ویژگی با مجموعه داده WDBC



شکل ۵: مقایسه بین نتایج استفاده از انتخاب ویژگی با مجموعه داده WBCD

۳-۳- ارزیابی کارایی روش طبقه‌بندی

بعد از انتخاب ویژگی‌های موثرتر، عمل طبقه‌بندی با طبقه‌بند ماشین بردار پشتیبان به ازای سه تابع هسته‌ی پایه شعاعی، تابع چند جمله‌ای و تابع خطی انجام شده است. ۸۰٪ داده‌ها به عنوان مجموعه داده‌های آموزش و ۲۰٪ داده‌ها به عنوان مجموعه داده‌های آزمون در نظر گرفته می‌شوند.

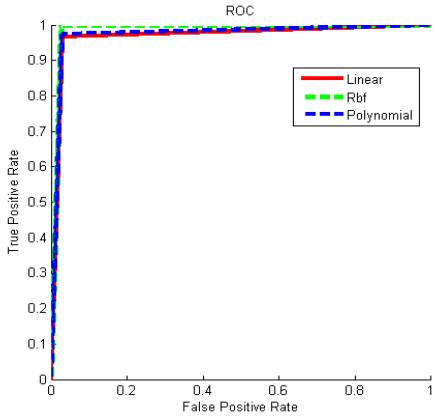
شکل ۶، خروجی واقعی (مقدار پیش بینی شده) که توسط مدل یادگیری بدست آمده با مقدار هدف (مقدار تشخیصی شناخته شده) را در مقایسه با کارهای پیشین نشان می‌دهد (معیار دقت). روش پیشنهادی SVM با توابع هسته پایه شعاعی، خطی و چند جمله‌ای نیز در سه ستون آخر نشان داده شده است. همان‌طور که از شکل مشاهده می‌شود SVM به ازای دو تابع RBF و چندجمله‌ای کارایی خوبی را نشان می‌دهد ولی به ازای تابع خطی چندان مناسب نیست. همچنین نتایج نمودار شکل ۶، نشان دهنده‌ی بالا بودن دقت روش پیشنهادی با دو تابع چندجمله‌ای و RBF نسبت به تمام کارهای پیشین، بجز دو مطالعه‌ی [۱۱] و [۱۹] دارد.

نتایج آزمایش عملکرد SVM به ازای دیگر معیارها و با استفاده از توابع کرنل مختلف بر روی مجموعه داده WBCD و WDBC در جداول ۵ و ۶ نشان داده شده است. همان‌طور که از جدول مشاهده می‌شود، در معیارهای مختلف کرنل ها متفاوت عمل کرده اند.

منحنی ROC برای سه کرنل نیز در شکل ۷ ارائه شده است. نتایج نشان می‌دهد که در معیار دقت، کرنل چند جمله ای بیشترین مقدار را بدست آورده است و در معیار AUC کرنل تابع پایه شعاعی بیشترین مقدار را بدست آورده است. طبق نتایج کرنل RBF قادر بوده است که همه‌ی نمونه‌های بدخیم را درست تشخیص دهد. طبق نتایج کرنل RBF قادر بوده است که همه‌ی نمونه‌های بدخیم را درست تشخیص دهد.

جدول ۵: نتایج طبقه‌بند SVM بر روی WBCD

معیارهای ارزیابی	توابع کرنل ماشین بردار پشتیبان		
	تابع پایه شعاعی	خطی	چند جمله ای
دقت	۰/۹۸۲۸	۰/۹۶۹۹	۰/۹۸۴۲
ناحیه زیر نمودار ROC	۰/۹۸۶۹	۰/۹۶۹۲	۰/۹۸۱۱
نرخ مثبت کاذب	۰/۰۲۶۲	۰/۰۲۸۳	۰/۰۱۵۱
حساسیت	۱	۰/۹۴۷۱	۰/۹۸۳۱
صحت	۰/۹۵۲۵	۰/۹۶۶۸	۰/۹۷۰۹
معیار F	۰/۹۷۵۶	۰/۹۵۶۷	۰/۹۷۷۸



شکل ۷: منحنی ROC ازعملکرد SVM با سه کرنل به ازای مجموعه داده WBCD

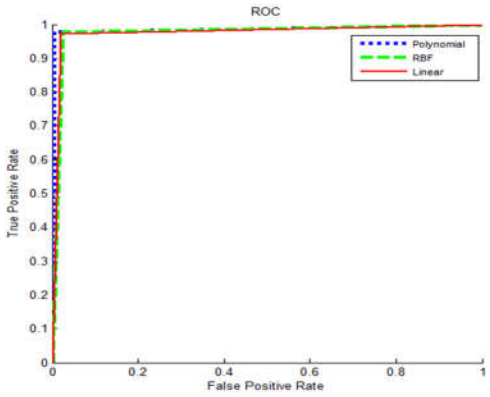
همچنین، نتایج آزمایشات عملکرد SVM با سه کرنل بر روی مجموعه داده WDBC در جدول ۵ نشان داده شده است. جدول نشان می‌دهد **که کرنل چند جمله‌ای موفق تر عمل کرده است.**

منحنی ROC با سه کرنل بر روی مجموعه داده WDBC نیز در شکل ۸ ارائه شده است. در معیارهای دقت و AUC، کرنل چندجمله‌ای بیشترین مقدار را بدست آورده است. در نهایت کرنل چندجمله‌ای قادر بوده است نمونه‌های بدخیم را با درصد بالاتری نسبت به دو کرنل دیگر تشخیص دهد.

جدول ۶: نتایج طبقه‌بند SVM بر روی WDBC

معیارهای ارزیابی	توابع کرنل ماشین بردار پشتیبان		
	تابع پایه شعاعی	خطی	چند جمله ای
دقت	۰/۹۷۵۴	۰/۹۷۷۱	۰/۹۸۷۷
سطح زیر نمودار ROC	۰/۹۷۱۹	۰/۹۷۵۱	۰/۹۸۸۳
نرخ مثبت کاذب	۰/۰۱۴	۰/۰۱۶۸	۰/۰۰۱۴
حساسیت	۰/۹۵۷۷	۰/۹۶۷۱	۰/۹۹۰۶
صحت	۰/۹۷۶۰	۰/۹۷۱۶	۰/۹۷۶۸
معیارF	۰/۹۶۶۷	۰/۹۶۹۳	۰/۹۷۳۶

شکل ٦: مقایسه روش پیشنهادی و سایر مطالعات بر روی WBCD و WDBC



شکل ۸: منحنیROC ازعملکرد SVM با سه کرنل به ازای مجموعه داده WDBC

۴- نتیجه‌گیری و پیشنهادهای آتی

در این مقاله، برای انتخاب مجموعه‌ای بهینه از ویژگی‌ها، برای پیش-بینی سرطان سینه، از روش رگرسیون لجستیک استفاده شده که تاکنون در کارهای پیشین این حوزه مشاهده نشده است. سپس برای دسته‌بندی نمونه‌ها به دو گروه خوش‌خیم و بدخیم از طبقه‌بند SVM با سه تابع کرنل چندجمله‌ای، خطی و پایه شعاعی استفاده شده است.

نتایج شبیه‌سازی‌ها روی دو مجموعه داده WBCD و WDBC نشان می‌دهند که روش انتخاب ویژگی مبتنی بر رگرسیون لجستیک منجر به افزایش معیارهای دقت و صحت اندازه‌گیری شده می‌شود. علاوه بر این، کرنل چندجمله‌ای دقت بالاتری برروی مجموعه داده‌ی WBCD بدست آورد در حالی که کرنل RBF دقت بالاتری برروی مجموعه داده WDBC بدست آورد.

به عنوان کارهای آتی، پیشنهاد می‌شود روش‌های یادگیری عمیق که دقت بهتری دارند و انتخاب ویژگی را همراه با دسته‌بندی می-توانند انجام دهند، بررسی شود که در این حوزه تاکنون انجام نشده است.

سپاسگزاری

نویسندگان مراتب تشکر و قدردانی خود را از معاونت پژوهشی دانشگاه شهید چمران اهواز، جهت حمایت مالی انجام این تحقیق و مرکز پردازشهای سریع دانشگاه شهید چمران اهواز جهت در اختیار قرار دادن منابع محاسباتی ابراز می‌دارد.

مراجع

[1] R. Sheikhpour, R. Sheikhpour, “Breast cancer diagnosis using non-parametric kernel density estimation,” Razi Journal on Medical Sciences (RJMS), Iran University of Medical Sciences, Vol. 23, No. 144, pp. 30-40, 2016.

[2] R. Sheikhpour, M. Agha Sarram, R. Sheikhpour, “Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer,” Applied Soft Computing. Vol. 40, pp. 113-131, 2016.

[3] S. Aalaei, H. Shahraki, AR. Rowhanimanesh, S. Eslami, “Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets,” Iran Journal on Basic Medical Sciences, Vol. 19, No. 5, pp. 476-482, 2016.

[4] G. RMA Sizilio, C. RM Leite, A. MG Guerreiro, A. DD Neto, “Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis,” BioMedical Engineering Online 11, 83, 2012. doi.org/10.1186/1475-925X-11-83

[5] M. Karabatak. “A new classifier for breast cancer detection based on Naïve Bayesian,” Measurement, Vol. 72, pp. 32-36, 2015.

[6] H. Asri, H. Mousannif, H. Al Moatassime, T. Noel, “Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis,” Procedia Computer Science. Vol. 83, pp. 1064-1069, 2016.

[7] M. Nilshahi, O. Ibrahim, H. Ahmadi, L.A. Shahmoradi, “Knowledge-Based System for Breast Cancer Classification Using Fuzzy Logic Method,” Telematics and Informatics, Vol. 34, No. 4, pp. 133–144, 2017.

[8] B. M. Gayathri, C. P. Sumathi, “Mamdani Fuzzy Inference system for Breast cancer risk detection,” IEEE International Conference on

Computational Intelligence and Computing Research (ICCIC), Madurai, India, Dec. 2015.

[9] B. M. Gayathri, C. P. Sumathi, "An Automated Technique using Gaussian Naïve Bayes Classifier to Classify Breast Cancer," International Journal of Computer Application, Vol. 148, No. 6, pp. 16-21, 2016.

[10] A. Hazra, S. Kumar Mandal, A. Gupta, "Study and Analysis of Breast Cancer Cell Detection using Naïve Bayes, SVM and Ensemble Algorithms," International Journal of Computer Applications, Vol. 145, No. 2, pp. 39-45, 2016.

[11] A. Bhardwaj, A. Tiwari, "Breast Cancer Diagnosis Using Genetically Optimized Neural Network Model," Expert Systems with Applications, Vol. 42, No. 10, pp. 4611-4620, 2015.

[12] N. Modi, K. Ghanchi, "Comparative Analysis of Feature Selection Methods and Associated Machine Learning Algorithms on Wisconsin Breast Cancer Dataset (WBCD)," Proceedings of International Conference on ICT for Sustainable Development, Advances in Intelligent Systems and Computing, Vol. 408, pp. 215-224, 2016.

[13] L. Abdel-Ilah, L. Sahinbegoviü, "Using machine learning tool in classification of breast cancer," International Conference in Medical and Biological Engineering in Bosnia and Herzegovina, IFMBE proceedings, Vol. 62, pp. 3-8, March 2017.

[14] UCI Machine Learning Repository, Breast Cancer Wisconsin(Original)Dataset https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Original)

[15] H. Hannah Inbarani, M. Bagyamathi, A. T. Azar, "A novel hybrid feature selection method based on rough set and improved harmony search," Neural Computing and Applications, Vol. 26, No. 8, pp. 1859-1880, 2015

[16] F. Ahmad, N. A. Mat Isa, Z. Hussain, M. K. Osman, S. N. Sulaiman, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer," Pattern Analysis and Applications, Vol. 18, No. 4, pp. 861-870, 2015.

[17] E. Besharati, M. Naderan, E. Namjoo, "LR-HIDS: Logistic Regression Host-based Intrusion Detection System for Cloud Environments," Journal of Ambient Intelligence and Humanized Computing, Vol. 10, No. 9, pp. 3669-3692, 2019.

[18] A. Ahmadi, P. Afshar, "Intelligent breast cancer recognition using particle swarm optimization and support vector machines," Journal of Experimental & Theoretical Artificial Intelligence, Vol. 28, No. 6, pp. 1021-1034, 2015.

[19] L. Peng, W. Chen, W. Zhou, F. Li, J. Yang, J. Zhang, "An immune-inspired semi-supervised algorithm for breast cancer diagnosis," Computer Methods and Programs in Biomedicine, Vol. 134, pp. 259-265, 2016.

[20] A. Mert, N. Kılıç, N., Bilgili, E., Akan, A. "Breast Cancer Detection with Reduced Feature Set," Computational and Mathematical Methods in Medicine, Vol. 2015, Article ID 265138.

[21] Mert, A., Kılıc, N., Akan, A.An improved hybrid feature reduction for increased breast cancer diagnostic performance.Biomedical Engineering Letters. 2014; 4(3), 285–291.

پاورقی‌ها:

[1]Fine Needle Aspiration
[2]Expectation Maximization
[3]Classification and Regression Trees
[4]Multi-collinearity
[5]Area Under the Curve