مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال هفتم، شماره چهارم، زمستان ۱۳۹۳

تشخیص سرطان پستان با استفاده از کاهش دو مرحله‌ای ویژگی‌های استخراج شده آسپیراسیون سوزنی و الگوریتم‌های داده‌کاوی

راضیه شیخ‌پور: دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، یزد، ایران
مهدی آقا صرام*: دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، یزد، ایران
محمدرضا زارع میرک آباد: دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، یزد، ایران
رباب شیخ‌پور: دانشکده پزشکی، واحد یزد، دانشگاه آزاد اسلامی، یزد، ایران

چکیده

مقدمه: تشخیص زودهنگام سرطان پستان نقش بسیار کلیدی در درمان و حیات بیمار ایفا می‌کند. امروزه با استفاده از خصوصیات استخراج شده از آزمایش آسپیراسیون سوزنی و الگوریتم‌های داده‌کاوی می‌توان روش‌های نوین و هوشمندی در نظام سلامت و درمان ارایه داد که با دقت بالایی قادر به تشخیص سرطان پستان باشند، هدف از انجام این مطالعه تشخیص سرطان پستان با استفاده از کاهش دو مرحله‌ای ویژگی‌های استخراج شده آسپیراسیون سوزنی و الگوریتم‌های داده‌کاوی است.

روش بررسی: در این مطالعه از داده‌های پایگاه داده WDBC موجود در UCI استفاده شد. این پایگاه شامل ۵۶۹ نمونه خوش‌خیم و بدخیم توده پستان با ۳۱ ویژگی است که ویژگی اول شماره شناسه پرونده بیمار و بقیه‌ی ویژگی‌ها نتایج کمی آزمایش آسپیراسیون سوزنی برای هر نمونه است. در این تحقیق، برای افزایش کارایی سیستم‌های تشخیص سرطان پستان روش کاهش ویژگی دو مرحله‌ای پیشنهاد شد و عملکرد روش‌های درخت تصمیم J48، بیزین ساده، طبقه‌بندی‌کننده درجه دوم، ماشین بردار پشتیبان و روش k نزدیکترین همسایه بر روی ویژگی‌های کاهش یافته مورد بررسی قرار گرفت.

یافته‌ها: بررسی‌های صورت گرفته نشان دادند که کاهش ویژگی دو مرحله‌ای موجب افزایش دقت الگوریتم‌های داده‌کاوی در تشخیص سرطان پستان می‌شود. دقت مدل ایجاد شده با استفاده از کاهش ویژگی دو مرحله‌ای مبتنی بر ضریب همبستگی و الگوریتم PCA در روش نزدیکترین همسایه براساس فاصله اقلیدسی۹۷/۵۴٪ می‌باشد که نسبت به سایر روش‌ها دارای بالاترین دقت است.

نتیجه‌گیری: نتایج این مطالعه نشان داد که با استفاده از الگوریتم‌های داده‌کاوی و کاهش ویژگی دو مرحله‌ای مبتنی بر انتخاب ویژگی و استخراج ویژگی می‌توان با دقت بالایی سرطان پستان را تشخیص داد. در واقع با استفاده از این روش‌ها می‌توان سیستم‌های نوینی برای کمک به پزشکان طراحی نمود که موجب تسهیل در فرآیندهای تشخیصی و درمانی شوند.

واژه‌های کلیدی: سرطان پستان، داده‌کاوی، کاهش ویژگی دو مرحله‌ای، دسته‌بندی.

* نشانی نویسنده پاسخگو: یزد، دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، مهدی‌آقا صرام.
نشانی الکترونیک: mehdi.sarram@yazd.ac.ir

مقدمه

امروزه به‌دلیل گسترش دانش در حوزه پزشکی و پیچیدگی تصمیمات مرتبط با تشخیص و درمان، توجه متخصصین به استفاده از ابزارهای هوشمند و سیستم‌های پشتیبان تصمیم‌گیری در امور پزشکی جلب شده است و استفاده از انواع مختلف سیستم‌های هوشمند در پزشکی رو به افزایش است (۲،۱). استفاده از این ابزارها و سیستم‌ها، می‌تواند خطاهای احتمالی ناشی از خستگی یا بی‌تجربگی متخصصین بالینی را در امر تشخیص و درمان بیماری‌ها کاهش دهد. همچنین با استفاده از این سیستم‌ها، می‌توان پایگاه داده‌های پزشکی را در زمان بسیار کمتر و با جزییات بیشتر تحلیل نمود (۴-۱).

سرطان پستان شایع‌ترین سرطان در زنان است. در این بیماری، سلول‌های بدخیم (سرطانی) در بافت پستان تشکیل می‌شوند (۵ و ۶). در حال حاضر شانس بروز سرطان پستان در زنان آمریکایی یک نفر از هر ۹ تا ۸ نفر است و سالانه باعث مرگ حدود ۴۴۰۰۰ زن مبتلا می‌شود (۷). همچنین این بیماری شایع‌ترین علت مرگ و میر ناشی از سرطان در زنان ایرانی است. اگرچه شیوع این بیماری در سنین قبل از ۳۰ تا ۲۵ سالگی نادر است اما بروز این سرطان در سنین کمتر حتی در سن جوانی نیز گزارش شده است (۸-۶).

با تشخیص زودهنگام سرطان پستان و ارایه درمان مؤثر و راه‌کارهای ویژه، می‌توان به افزایش بقا، کاهش مرگ و ارتقا کیفیت زندگی بیماران کمک شایانی نمود. آمار نشان می‌دهد که میزان بقای بیماران مبتلا به سرطان پستان تا پنج سال پس از تشخیص، ۸۸٪ و ده سال پس از تشخیص ۸۰٪ بوده است (۹).

مرسوم‌ترین و قطعی‌ترین روش تشخیص سرطان پستان بیوپسی ضایعه می‌باشد، ولی ازآنجایی‌که ۷۰٪ موارد بیوپسی توده‌های پستان مربوط به ضایعات خوش‌خیم است، لذا استفاده از روش‌های کمتر تهاجمی مانند آسپیراسیون سوزنی ظریف[1] (FNA) از اتلاف هزینه بیمار و نیز تغییرات بافتی در ضایعه جلوگیری می‌کند. آزمایش آسپیراسیون سوزنی روشی ساده، ارزان و غیرتهاجمی برای تشخیص دقیق و زودهنگام این سرطان است (۱۰،۷). پس از استخراج خصوصیات سیتولوژی بیمار از مایع استخراج شده از بافت پستان توسط روش آسپیراسیون سوزنی ظریف، نیاز است تا خوش‌خیم یا بدخیم بودن تومور تشخیص داده شود. در مواردی‌که با قاطعیت نتوان خوش‌خیم یا بدخیم بودن بیماری را تشخیص داد، استفاده از سیستم‌های هوشمند و الگوریتم‌های داده‌کاوی راهنمای خوبی برای پزشک و متخصصین بالینی هستند (۱۰). پژوهش‌ها نشان داده‌اند که فرآیندهای تشخیص و دسته‌بندی سرطان‌ها با استفاده از فناوری‌های نوین کامپیوتری موفق عمل کرده‌اند (۱۳-۱۱). بنابراین با استفاده از ویژگی‌های استخراج شده از آزمایش آسپیراسیون سوزنی و با کمک الگوریتم‌های داده-کاوی می‌توان سرطان پستان را با دقت بالایی تشخیص داد.

تاکنون تحقیقات زیادی در رابطه با تشخیص سرطان پستان با کمک الگوریتم‌های داده‌کاوی انجام شده است که از این میان به چند مورد از مطالعات انجام شده بر روی پایگاه داده‌ی WDBC[2] می‌پردازیم.

Tan و همکاران یک تکنیک دسته‌بندی دو مرحله‌ای ترکیبی برای استخراج قوانین طبقه‌بندی ارایه دادند که توانست با دقت ۹۳/۰۴٪ سرطان پستان را تشخیص دهد (۱۴). Ster و Dobnikar نشان دادند که روش تجزیه و تحلیل گسسته خطی در تشخیص سرطان پستان با استفاده از پایگاه WDBC دارای دقت ۹۶/۸٪ است (۱۵). Aruna و همکاران از شبکه‌های عصبی مصنوعی، درخت تصمیم‌گیری و روش بیزین ساده برای مقایسه و توسعه مدل‌های پیش‌بینی سرطان پستان استفاده نمودند. نتایج آزمایشات آنها بر روی مجموعه داده‌های پایگاه WDBC نشان داد که شبکه عصبی دارای دقت ۹۳/۶۷٪، روش بیزین ساده دارای دقت ۹۲/۶۱٪ و درخت تصمیم J48 و CART دارای دقت ۹۲/۹۷٪ است (۱۶).

در این مطالعه، با استفاده از الگوریتم‌های داده‌کاوی و کاهش دو مرحله‌ای ویژگی‌های استخراج شده از آزمایش آسپیراسیون سوزنی موجود در پایگاه داده WDBC، سیستمی کارآمد برای تشخیص سرطان پستان طراحی می‌گردد که با دقت بالایی خوش‌خیم یا بدخیم بودن تومورهای پستان را تشخیص دهد.

[1] Fine Needle Aspirate

[2] Wisconsin Diagnostic Breast Cancer

مواد و روش‌ها

به منظور استفاده موفقیت آمیز از داده‌کاوی، معمولاً یک فرآیند کلی دنبال می‌شود. یکی از متدولوژی‌های انجام پروژه‌های داده‌کاوی متدولوژی CRISP است که بیشترین استفاده را در بین بقیه روش‌ها دارد. این متدولوژی شامل فازهای شناخت سیستم، شناخت داده‌ها، آماده‌سازی داده‌ها، مدل‌سازی، ارزیابی و توسعه است (۱۷). این مطالعه از نوع مطالعات گذشته‌نگر است و مدل پیشنهادی آن براساس فازهای متدولوژیCRISP ارایه شده است. مراحل مدل پیشنهادی در ادامه شرح داده می‌شوند.

شناخت سیستم:

مرحله شناخت سیستم شامل تعیین اهداف مورد نظر و عوامل موفقیت کلیدی سیستم است. رشد گسترده سرطان پستان در میان زنان و آمار بالای مرگ ومیر ناشی از آن، نیاز به سیستمی برای تشخیص زودهنگام آن را ضروری می‌سازد. در واقع هدف سیستم مورد نظر تشخیص هوشمند سرطان پستان بدون انجام عمل جراحی، با استفاده از روش‌های کمتر تهاجمی مانند آسپیراسیون سوزنی ظریف بود که از اتلاف هزینه بیمار و نیز تغییرات بافتی در پستان جلوگیری نماید و اطلاعات ضروری و دانش مورد نیاز را برای تصمیم‌گیری بهتر پزشکان ارایه دهد.

شناخت داده‌ها:

در مرحله شناخت داده‌ها جمع‌آوری داده‌های اولیه، توصیف داده‌ها، بازرسی و بررسی داده‌ها و اعتبارسنجی کیفیت داده‌ها انجام می‌شود. داده‌های مورد استفاده در این مطالعه مبتنی بر ویژگی‌های استخراج شده آسپیراسیون سوزنی پرونده‌های بیماران مبتلا به سرطان پستان در بیمارستان Wisconsin بود و از پایگاه داده WDBC به‌دست آمده است (۱۸).

ویژگی‌های پایگاه داده WDBC به صورت زیر به‌دست آمده است. ابتدا توسط آزمایش آسپیراسیون سوزنی مایع بافت پستان استخراج می‌شود. معاینه بصری این نمونه در زیر میکروسکوپ انجام شده و تصویر به یک تصویر خاکستری تبدیل می‌شود. سپس با یک برنامه کامپیوتری و تکنیک‌های پردازش تصویر، مرز هسته سلول‌ها در این تصویر میکروسکوپیک مشخص شده و ویژگی‌هایی مانند شعاع، بافت، محیط، مساحت، همواری، فشردگی، تقعر، نقاط مقعر، تقارن و بعد فراکتالی برای هر هسته محاسبه می‌شود. سرانجام 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 به‌دست می‌آید.

مجموعه داده‌های پایگاه WDBC شامل اطلاعات ۵۶۹ بیمار با ۳۱ ویژگی است که ویژگی اول شماره شناسه پرونده بیمار و بقیه‌ی ویژگی‌ها نتایج کمی آزمایش آسپیراسیون سوزنی برای هر بیمار است. هر بیمار با یک برچسب *خوش‌خیم* یا *بدخیم* مشخص می‌گردد که نشان دهنده وضعیت تومور در وی می‌باشد. از ۵۶۹ نمونه موجود در این پایگاه، ۳۵۷ نمونه دارای برچسب خوش‌خیم و ۲۱۲ نمونه دارای برچسب بدخیم هستند.

آماده‌سازی داده‌ها:

مرحله آماده‌سازی داده‌ها با عنوان پیش‌پردازش داده‌ها[3] شناخته می‌شود و جهت بهبود کیفیت داده‌های واقعی برای داده‌کاوی لازم است. (۱۹).

در مرحله آماده‌سازی داده‌ها در مدل پیشنهادی، ابتدا شماره شناسه بیمار را حذف نموده و سپس داده‌ها را نرمال‌سازی نمودیم. نرمال‌سازی داده‌ها روشی است که هنگامی‌که داده‌ها در محدوده یا دامنه متفاوتی قرار داشته باشند آنها را در دامنه مشابه قرار می‌دهد و معمولاً منجر به کسب نتایج بهتر می‌شود (۱۹). از آنجایی‌که داده‌های پایگاه داده WDBC دارای دامنه‌های متفاوتی هستند، ممکن است ویژگی‌های دارای مقادیر بزرگ اثر زیادتری در تابع هزینه نسبت به ویژگی‌های با مقادیر کم داشته باشند که این مشکل با نرمالیزه نمودن ویژگی‌ها در دامنه‌های مشابه برطرف خواهد شد. برای نرمال‌سازی داده‌های پایگاه WDBC از روش نرمال‌سازی مینیمم- ماکزیمم[4] استفاده نمودیم.

یکی دیگر از روش‌های پیش‌پردازش داده‌ها، کاهش ویژگی‌ها است. روش‌های کاهش ویژگی، نمایش کوتاه‌تری از مجموعه داده‌های اولیه را محاسبه و ارایه می‌کنند. این روش‌ها به دو دسته انتخاب ویژگی و استخراج ویژگی (ترکیب ویژگی) تقسیم می‌شوند (۲۰). مسئله انتخاب ویژگی در واقع شناسایی و انتخاب یک زیرمجموعه مفید

[3] Preprocessing
[4] min-max normalization

از ویژگی‌ها از میان مجموعه داده‌های اولیه است که حداکثر توان را در پیشگویی خروجی دارا باشند (۲۱، ۲۰). استخراج ویژگی فرآیند نگاشت و انتقال ویژگی‌ها از فضای با ابعاد بیشتر به فضای با ابعاد و متغیرهای کمتر است (۲۲،۲۰).

در مدل پیشنهادی این مطالعه برای دسته‌بندی داده‌های سرطان پستان، از کاهش دو مرحله‌ای ویژگی‌ها استفاده شد. بدین صورت که در مرحله اول براساس یک روش انتخاب ویژگی، زیرمجموعه‌ای مفید از ویژگی‌ها انتخاب شد. برای این منظور با استفاده از روش‌های انتخاب ویژگی مبتنی بر درخت تصمیم، انتخاب ویژگی مبتنی بر ضریب همبستگی و انتخاب ویژگی CfsSubsetEval (روش انتخاب ویژگی موجود در ابزار داده‌کاوی Weka) زیرمجوعه‌ای مهم از ویژگی‌ها را انتخاب کردیم. سپس در مرحله بعد با اعمال روش استخراج ویژگی PCA بر روی ویژگی‌های باقیمانده ترکیبی از این ویژگی‌ها با ابعاد کمتر به دست آمد. بدین صورت با کاهش دو مرحله‌ای ویژگی‌ها، ترکیبی از ویژگی‌های مفید با ابعاد کمتر حاصل شد. تکنیک‌های کاهش ویژگی دو مرحله‌ای پیشنهادی، به صورت زیر نام‌گذاری شدند:

۱-کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر درخت تصمیم J48 (موجود در ابزار داده‌کاوی Weka) و اعمال روش استخراج ویژگی PCA بر روی آن (J48+ PCA).

۲-کاهش ویژگی با استفاده از روش انتخاب ویژگی مبتنی بر ضریب همبستگی (Correlation Coefficient) داده‌ها با پارامتر تصمیم‌گیری (برچسب کلاس) و اعمال روش استخراج ویژگی PCA بر روی آن (CC+ PCA).

۳-کاهش ویژگی با استفاده از روش انتخاب ویژگی CfsSubsetEval (موجود در ابزار داده‌کاوی Weka) و اعمال روش استخراج ویژگی PCA بر روی آن (Cfs+PCA).

مدل‌سازی:

در این مرحله با استفاده از الگوریتم‌های مختلف داده‌کاوی به مدل‌سازی داده‌های پردازش شده پرداختیم. مدل‌سازی با استفاده از نرم‌فزار Matlab R2013a و ابزار داده‌کاوی Weka انجام گرفت. برای مدل‌سازی از روش‌های درخت تصمیم J48، بیزین ساده[5]، طبقه‌بندی‌کننده درجه دوم، ماشین بردار پشتیبان (SVM-RBF) و روش k نزدیک‌ترین همسایه استفاده نمودیم که در ادامه شرح داده شده‌اند.

- طبقه‌بندی کننده درجه دوم: این روش براساس قانون بیزین است و فرض می‌کند که نمونه‌ها از یک توزیع آماری پیروی می‌کنند. در این روش با تخمین کوواریانس و میانگین داده‌ها و جای‌گذاری آنها در مدل فرضی می‌توان به تصمیم‌گیری در مورد برچسب داده‌ها پرداخت (۲۰).
- روش بیزین ساده: این روش مبتنی بر قانون بیزین است و فرض می‌کند ویژگی‌ها از هم مستقل هستند. در روش بیزین ساده تنها نیاز است تا واریانس ویژگی‌ها به ازای هر کلاس محاسبه شود و نیازی به محاسبه ماتریس کوواریانس نیست (۲۰،۱۹).
- روش k نزدیک‌ترین همسایه[6] (KNN): این روش یک تکنیک دسته‌بندی است که تصمیم‌گیری در مورد اینکه یک نمونه جدید در کدام کلاس قرار گیرد با بررسی تعدادی (k) از شبیه‌ترین نمونه‌ها یا همسایه‌ها انجام می‌شود. این روش برای یافتن شباهت بین نمونه‌ها نیاز به یک معیار فاصله نظیر فاصله اقلیدسی یا فاصله منهتن دارد (۱۹).
- درخت تصمیم: درخت تصمیم ساختاری شبیه به فلوچارت دارد و با مرتب کردن نمونه‌ها در درخت از گره ریشه به سمت گره‌های برگ آنها را دسته‌بندی می‌کند (۱۹).
- روش ماشین بردار پشتیبان (SVM): این روش با ساخت یک ابرسطح (که عبارت است از یک معادله خطی)، سعی دارد بهترین ابرسطحی را پیدا کند که با حداکثر فاصله، داده‌های مربوط به دو طبقه را از هم تفکیک کند (۲۰).

در مورد روش ایجاد مدل پیش‌بینی، ابتدا باید با استفاده از تکنیکی مجموعه داده‌ها را به زیرمجموعه‌هایی جداگانه برای آموزش و آزمایش مدل تفکیک کرد. تکنیک انتخاب شده در این مطالعه، روش اعتبارسنجی متقاطع با ده تکرار[7] بود. اعتبارسنجی متقاطع به این دلیل انتخاب شد

[5] NaïveBayes
[6] Nearest Neighbors
[7] Fold Cross Validation

که آزمایش‌ها نشان داده‌اند بهترین انتخاب برای به‌دست آوردن دقیق‌ترین تخمین است.

ارزیابی:

در این مرحله به ارزیابی نتایج حاصل از مدل‌سازی پرداختیم. برای ارزیابی مدل‌ها از شاخص‌های دقت، حساسیت و ویژگی استفاده شد.

توسعه:

در این مرحله، با توجه به نتایج به‌دست آمده در مرحله قبل، مدلی که دارای عملکرد مناسبی است برای دسته‌بندی داده‌ها انتخاب شد.

یافته‌ها

بعد از اعمال تکنیک‌های کاهش ویژگی پیشنهاد شده بر روی داده‌هایWDBC، تعداد ویژگی‌ها به صورت نشان داده شده در جدول ۱ کاهش یافت.

همان‌گونه‌که در جدول ۱ مشخص شده است، تعداد ویژگی‌های پایگاه WDBC با استفاده از تمام روش‌های کاهش ویژگی دو مرحله‌ای به طور قابل توجهی کاهش یافت. پس از کاهش ویژگی‌های پایگاه WDBC، روش‌های درخت تصمیم J48، بیزین ساده، طبقه‌بندی‌کننده درجه دوم، ماشین بردار پشتیبان (SVM-RBF) و روش k نزدیکترین همسایه بر روی این داده‌ها اجرا گردیدند و دقت، حساسیت و ویژگی هرکدام از روش‌ها با استفاده از روش اعتبارسنجی متقاطع 10-foldمحاسبه شدند. جدول ۲ نتایج بررسی روش‌های گوناگون دسته‌بندی را با استفاده از شاخص دقت بر روی داده‌های پایگاه WBCD نشان می‌دهد.

همان‌گونه در جدول ۲ مشاهده می‌شود، با استفاده از کاهش ویژگی دو مرحله‌ای J48+ PCA دقت روش‌های درخت تصمیم J48، SVM-RBF و روش بیزین ساده بهبود یافت . دقت تمام روش‌ها با استفاده از کاهش ویژگی دو مرحله‌ایCC+ PCA افزایش یافت. این بدان معنی است که در تمام روش‌ها، حذف ویژگی‌های با ضریب همبستگی پایین و اعمال الگوریتم PCA بر روی داده‌های باقیمانده باعث افزایش عملکرد روش دسته‌بندی با استفاده از شاخص دقت شد. با استفاده از کاهش ویژگی دو مرحله‌ای Cfs+PCA دقت تمام روش‌ها به جز روش k نزدیکترین همسایه بهبود یافت. در میان تمام روش‌ها، روش k نزدیکترین همسایه براساس فاصله اقلیدسی با استفاده از کاهش دو مرحله‌ای CC+ PCA به بالاترین عملکرد براساس شاخص دقت دست یافت.

در روش‌های درخت تصمیم J48، طبقه‌بندی‌کننده درجه دوم و روش k نزدیک‌ترین همسایه، کاهش ویژگی دو مرحله‌ای CC+PCA در مقایسه با روش‌های کاهش ویژگی J48+ PCA و Cfs+ PCA دارای عملکرد بهتری براساس شاخص دقت بود. در روش k نزدیک‌ترین

جدول ۱: تعداد ویژگی‌ها پس از اعمال هر یک از تکنیک‌های کاهش ویژگی

نام روش	تعداد ویژگی‌ها
داده‌های نرمال بدون کاهش ویژگی	۳۰
کاهش ویژگی با استفاده از درخت تصمیم J48 و االگوریتم PCA (J48+ PCA)	۴
کاهش ویژگی با استفاده از ضریب همبستگی و الگوریتم PCA (CC+ PCA)	۵
کاهش ویژگی با استفاده از روش CfsSubsetEval و الگوریتم PCA (Cfs+ PCA)	۴

جدول۲: نتایج بررسی روش‌های دسته‌بندی گوناگون با استفاده از شاخص دقت

Method	J48	SVM-RBF	NaiveBayes	Quadratic	KNN+ Euclidean	KNN+ Manhattan
Full dataset	۹۳/۳۲	۹۰/۵۱	۹۳/۱۵	۹۵/۹۶	۹۷/۳۶	۹۷/۰۱
J48+PCA	۹۴/۳۸	۹۵/۲۵	۹۶/۱۳	۹۵/۶۱	۹۶/۴۹	۹۵/۹۶
CC+PCA	۹۴/۷۳	۹۴/۹۰	۹۳/۱۵	۹۶/۴۹	۹۷/۵۴	۹۷/۰۱
Cfs+PCA	۹۴/۰۲	۹۵/۶۱	۹۴/۲۰	۹۶/۱۳	۹۶/۱۳	۹۶/۳۱

همسایه با استفاده از فاصله اقلیدسی و منهتن، کاهش ویژگی دو مرحله‌ای CC+ PCA باعث بهبود دقت شد. به طور کلی روش k نزدیک‌ترین همسایه مبتنی بر فاصله اقلیدسی در مقایسه با روش‌های دیگر دارای دقت بالاتری برای تشخیص سرطان پستان بود.

جدول ۳ نتایج بررسی روش‌های دسته‌بندی را با استفاده از شاخص حساسیت بر روی داده‌های حاصل از تکنیک-های کاهش ویژگی نشان می‌دهد.

همان‌گونه‌که در جدول ۳ نشان داده شده است، با استفاده از کاهش ویژگی دو مرحله‌ای J48+ PCA عملکرد روش‌های درخت تصمیم J48، SVM-RBF و روش بیزین ساده براساس شاخص حساسیت بهبود یافت.

جدول ۴ نتایج بررسی روش‌های گوناگون دسته‌بندی را با استفاده از شاخص ویژگی بر روی داده‌های حاصل از تکنیک‌های کاهش ویژگی نشان می‌دهد.

همان‌گونه در جدول ۴ مشاهده می‌شود، با استفاده از کاهش ویژگی دو مرحله‌ای J48+ PCA عملکرد روش‌های درخت تصمیم J48، SVM-RBF و روش بیزین ساده براساس شاخص ویژگی بهبود یافت. روش‌های کاهش ویژگی دو مرحله‌ای CC+PCA و Cfs+PCA باعث بهبود عملکرد روش‌های درخت تصمیمJ48، SVM-RBF، روش بیزین ساده و طبقه‌بندی‌کننده درجه دوم براساس شاخص ویژگی شدند.

جدول۳: نتایج بررسی روش‌های دسته‌بندی گوناگون با استفاده از شاخص حساسیت

Method	J48	SVM-RBF	NaiveBayes	Quadratic	KNN+ Euclidean	KNN+ Manhattan
Full dataset	۹۲/۹۲	۹۱/۹۸	۸۹/۶۲	۹۴/۸۱	۹۳/۴۰	۹۲/۹۲
J48+PCA	۹۱/۵۱	۹۵/۲۸	۹۳/۸۷	۹۱/۹۸	۹۲/۴۵	۹۱/۹۸
CC+PCA	۹۱/۹۸	۹۵/۷۵	۸۸/۸۶	۹۵/۲۸	۹۴/۳۴	۹۳/۸۷
Cfs+PCA	۹۰/۰۹	۹۴/۸۱	۹۰/۵۷	۹۲/۴۵	۹۲/۹۲	۹۱/۵۱

جدول۴: نتایج بررسی روش‌های دسته‌بندی گوناگون با استفاده از شاخص ویژگی

Method	J48	SVM-RBF	NaiveBayes	Quadratic	KNN+ Euclidean	KNN+ Manhattan
Full dataset	۹۳/۵۶	۸۹/۶۴	۹۵/۲۴	۹۶/۶۴	۹۹/۷۲	۹۹/۴۴
J48+PCA	۹۶/۰۸	۹۵/۲۴	۹۷/۴۸	۹۷/۷۶	۹۸/۸۸	۹۸/۳۲
CC+PCA	۹۶/۳۶	۹۴/۴۰	۹۵/۸۰	۹۷/۲۰	۹۹/۴۴	۹۸/۸۸
Cfs+PCA	۹۶/۳۶	۹۶/۰۸	۹۶/۳۶	۹۸/۳۲	۹۸/۰۴	۹۹/۱۶

حساسیت روش‌های SVM-RBF، طبقه‌بندی‌کننده درجه دوم و روش k نزدیک‌ترین همسایه براساس فاصله-های اقلیدسی و منهتن با استفاده از کاهش ویژگی دو مرحله‌ای CC+PCA بهبود یافت. با استفاده از کاهش ویژگی دو مرحله‌ای Cfs+PCA عملکرد روش‌های SVM-RBF و بیزین ساده براساس شاخص حساسیت بهبود یافت. در میان تمام روش‌ها، روش SVM-RBF با استفاده از کاهش دو مرحله‌ای CC+PCA دارای بالاترین عملکرد براساس شاخص حساسیت بود.

بحث

در این مطالعه با استفاده از الگوریتم‌های داده‌کاوی و کاهش ویژگی دو مرحله‌ای به‌دسته‌بندی سرطان پستان بر اساس ویژگی‌های استخراج شده آسپیراسیون سوزنی پرداختیم. مدل کاهش ویژگی دو مرحله‌ای پیشنهادی باعث کاهش قابل ملاحظه‌ای در ابعاد داده‌ها شد.

با انجام سه روش انتخاب ویژگی با استفاده از درخت تصمیم J48، ضریب همبستگی و CfsSubsetEval بر روی داده‌های WDBC مشخص گردید که پنج ویژگی میانگین بافت، میانگین تقعر، خطای استاندارد مساحت،

بزرگترین مساحت و بزرگترین نقاط مقعر دارای اهمیت بیشتری هستند. در واقع این پنج ویژگی توسط هر سه روش به عنوان ویژگی‌های تأثیرگذار در تشخیص سرطان پستان انتخاب شدند که در مطالعات قبلی گزارش نشده است. نتایج مدل‌سازی نشان دادند که دقت تمام الگوریتم‌های داده‌کاوی با استفاده از روش کاهش ویژگی دو مرحله‌ای مبتنی بر ضریب همبستگی و الگوریتم PCA (CC+PCA) بهبود یافته است. همچنین نتایج مدل‌سازی نشان دادند که روش k نزدیک‌ترین همسایه با استفاده از فاصله اقلیدسی و کاهش ویژگی دو مرحله‌ای CC+PCA دارای بالاترین دقت در تشخیص سرطان پستان است. مزیت مدل پیشنهادی نسبت به سیستم‌های مشابه (۱۶ و ۲۳)، استفاده از تعداد ویژگی کمتر است که این امر موجب افزایش سرعت، ساده سازی و کاهش پیچیدگی مدل می‌شود.

Bamakan و همکاران در سال ۲۰۱۴ روشی برای انتخاب ویژگی براساس تحلیل پوششی داده‌های مجتمع و مدل آنتروپی ارائه دادند که با استفاده از روش‌های دسته‌بندی SVM، درخت تصمیم C5.0 و رگرسیون لجستیک بر روی پایگاه WDBC به ترتیب به دقت ۸۹/۸۶٪، ۹۳/۹۲٪ و ۹۵/۹۵٪ دست یافت. آنها همچنین با استفاده از روش رگرسیون لجستیک و تکنیک انتخاب ویژگی CfsSubsetEval به دقت ۹۵/۹۵٪ و با استفاده از روش رگرسیون لجستیک و تکنیک انتخاب ویژگی فیلتر به دقت ۹۶/۶۲٪ دست یافتند. در مطالعه آنها روش SVM با تکنیک‌های انتخاب ویژگی فیلتر و CfsSubset Eval با دقت ۸۷/۸۴٪ داده‌ها را به درستی پیش‌بینی نمود (۲۳). Xue و همکاران در سال ۲۰۱۲ با استفاده از روش BPSO دو مرحله‌ای بر روی داده‌های WDBC به دقت ۹۲/۹۸٪ و در سال ۲۰۱۴ با استفاده از PSO(4-2) به دقت ۹۳/۹۸٪ دست یافتند (۲۴ و ۲۵). Yao و همکاران در سال ۲۰۱۳ با استفاده از روش‌های RF، MARS و RF&MARS بر روی داده‌های WDBC به ترتیب به دقت ۹۶/۲۶٪، ۹۶/۷٪ و ۹۶/۲۹٪ دست یافتند. آنها همچنین نشان دادندکه روش درخت تصمیم C4.5 دارای دقت ۹۳/۱۶٪ و روش SVMMaj دارای دقت ۹۵/۸۵٪ است (۲۶).

Aruna و همکاران در سال ۲۰۱۱ به بررسی و مقایسه طبقه‌کننده‌های یادگیری با نظارت بر روی مجموعه داده‌های پایگاه WDBC پرداختند تا بهترین طبقه‌بندی‌کننده را بر روی این داده‌ها پیدا نمایند. در آزمایشات آنها روش بیزین ساده به دقت ۹۲/۶۱٪، شبکه عصبی به دقت ۹۳/۶۷٪ و درخت تصمیم J48 و CART به دقت ۹۲/۹۷٪ دست یافت (۱۶). Tan و همکاران در سال ۲۰۰۳ یک تکنیک دسته‌بندی دو مرحله‌ای ترکیبی برای استخراج قوانین طبقه‌بندی ارایه دادند. روش پیشنهادی آنها به دقت ۹۳/۰۴٪ بر روی پایگاه داده WDBC دست یافت (۱۴). Lavanya و Rani در سال ۲۰۱۱ به بررسی کارایی درخت تصمیمCART با انتخاب ویژگی و بدون انتخاب ویژگی بر روی پایگاه WDBC پرداختند (۲۷). در مطالعه آنها درخت تصمیم CART توانست بدون انتخاب ویژگی با دقت ۹۲/۹۷٪ و با انتخاب ویژگی با دقت ۹۲/۰۹٪ داده‌های WDBC را به درستی پیش بینی نماید.

Salama و همکاران در سال ۲۰۱۲ با استفاده از روش بیزین ساده و درخت تصمیم J48 به ترتیب به دقت ۹۲/۹۷٪ و ۹۳/۱۵٪ بر روی پایگاه داده WDBC دست یافتند (۱۳). Maldonado و همکاران در سال ۲۰۱۱ با استفاده از تکنیک انتخاب ویژگی Fisher و طبقه‌بندی کننده SVM به دقت ۹۴/۷٪ و با تکنیک انتخاب ویژگی RFE و طبقه‌بندی‌کننده SVMبه دقت ۹۵/۲۵٪ دست یافتند (۲۸). Ster و Dobnikar نشان دادند که روش تجزیه و تحلیل گسسته خطی بر روی پایگاه WDBC دارای دقت ۹۶/۸٪ است (۱۵). Abonyi و Szeifert با تکنیک خوشه‌بندی فازی نظارت شده بر روی داده‌های WDBC به دقت ۹۵/۷۵٪ دست یافتند (۲۹). Joachims نشان داد که تکنیک فازی- عصبی بر روی داده‌های WDBC دارای دقت ۹۵/۰۶٪ است (۳۰).

نتایج مطالعات ذکر شده محققان بر روی داده‌های پایگاه WDBC در مقایسه با روش کاهش دو مرحله‌ای ویژگی پیشنهادی حاکی از برتری روش پیشنهادی این مطالعه است.

نتیجه‌گیری

در این مطالعه، با استفاده از الگوریتم‌های داده‌کاوی و روش کاهش ویژگی دو مرحله‌ای مبتنی بر روش‌های انتخاب و استخراج ویژگی، دقت شناسایی سیستم‌های

سرطان پستان افزایش یافت. در واقع با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوین و با صرفه‌تری در نظام سلامت و درمان ارائه کرد که با دقت بالایی قادر به تشخیص سرطان پستان باشند. استفاده از این سیستم‌ها می‌تواند موجب کاهش بیوپسی‌های غیر ضروری شود و دقت تشخیص سرطان پستان را بهبود بخشد. پیشنهاد می‌شود که برای کاهش ویژگی دومرحله‌ای از روش‌های دیگر انتخاب ویژگی نظیر بهره‌ی اطلاعاتی و نسبت بهره در ترکیب با الگوریتم PCA استفاده شود. همچنین می‌توان از روش‌های دیگر استخراج ویژگی نظیر روش LDA در ترکیب با روش‌های مختلف انتخاب ویژگی استفاده کرد و اثر اعمال این روش‌ها بر روی داده‌های استخراج شده از تومورهای پستان را با استفاده از الگوریتم‌های مختلف داده‌کاوی بررسی نمود.

References

1. Milovic B. Prediction and decision making in Health Care using Data Mining. Int J Publ Health Sci (IJPHS) 2012; 1(2): 69-78.

۲. صدوقی فرحناز، شیخ‌طاهری عباس. کاربرد سیستم‌های هوش مصنوعی در تصمیم‌گیری‌های پزشکی: مزایا و چالش‌ها. مدیریت اطلاعات سلامت، ۱۳۹۰؛ ۸ (۳): ۴۴۵-۴۴۰.

3. Hariz M, Adnan M, Husain W, Rashid N. A. Data Mining for Medical Systems: A Review. Int Conf Adv Comput Inform Tech - ACIT 2012;17-22.

4. Rafe V, Farhoud RH. A Survey on Data Mining Approaches in Medicine. Int Res J Appl Basic Sci 2013; 4 (1): 196-202.

5. Sheikhpour R, Ghasemi N, Yaghmaei P, Mohiti J. Immunohistochemical assessment of p53 protein and its correlation with clinicopathological parameters in breast cancer patients. Indian J Sci Tech 2014; 7(4): 472-9.

6. Wang YA, Johnson SK, Brown BL, Carragher LM, Sakkaf KL, Royds JA, et al. Enhanced anticancer effect of a phosphatidylinositol-3 kinase inhibitor and doxorubicin on human breast epithelial cell lines with different p53 and oestrogen receptor status. Int J Canc 2008; 123(7):1536-44.

۷. مدنی سیدحمید، ایزدی بابک، کنانی مالک، خزاعی صدیقه، حمزه لوی مریم، مولایی توانا پرستو. بررسی ارزش تشخیصی آسپیراسیون سوزنی (FNA) توده‌های قابل لمس پستان در بیماران مراجعه کننده به بیمارستان امام رضا (ع) کرمانشاه. مجله علوم پزشکی ارومیه، ۱۳۹۱؛ ۲۳ (۴): ۴۲۶-۴۲۲

8. Richie RC, Swanson JO. Breast Cancer: A Review of the Literature. J Insur Med 2003; 35:85 -101.

۹. طلوعی اشلقی عباس، پورابراهیمی علی، ابراهیمی ماندانا، قاسم احمد لیلا. پیش بینی عود مجدد سرطان پستان به کمک سه تکنیک داده‌کاوی. فصلنامه بیماری‌های پستان ایران، ۱۳۹۱؛ ۵ (۴): ۳۴-۲۳.

۱۰.محمد علیپور، جواد حدادنیا. معرفی یک سیستم هوشمند برای تشخیص دقیق سرطان پستان. فصلنامه بیماری‌های پستان ایران، ۱۳۸۸؛ ۲ (۲): ۴۰-۳۳.

11. Nahar J, Imam T, Tickle KS, Shawkat ABM, Chen YPP. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. Expert Syst Appl 2012; 39: 12371-7.

12. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst App. 2011; 38: 9573-9.

13. Salama GI, Abdelhalim MB, Zeid MAE. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. Int J Comput Sci Inform Tech 2012; 1(1): 2277- 0764.

14. Tan KC, Yu Q, Heng CM, Lee TH. Evolutionary computing for knowledge

discovery in medical diagnosis. Artif Intell Med 2003; (27):129-54.

15. Ster B, Dobnikar A. Neural networks in medical diagnosis: Comparison with other methods. Proc Int Conf Eng Appl neural networks.427-30.

16. Aruna S, Rajagopalan DS, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput Sci Inform Tech 2011; 2: 37-45.

17. Chapman P, Clinton J, Kerber R, Khabaza T, Reinartz T, Shearer C, Wirth R. CRISP-DM 1.0 Step-by-step data mining guide. 2000.

18. Wolberg WH, Street WN, Mangasarian OL. Image analysis and machine learning applied to breast cancer diagnosis and prognosis. Anal Quant Cytol Histol 1995; 17(2): 77-87.

19. Han J, Kamber M, Pei J. Data Mining Concepts and Techniques. (3th ed.), Morgan kaufmann 2012.

20. Alpaydin E. Introduction to machine learning. (2th ed.). London: MIT press 2010.

21. Zhu M, Song J. An Embedded Backward Feature Selection Method for MCLP Classification Algorithm. Procedia Comput Sci 2013: 17; 1047-54.

22. Santosa V, Datiaa N, Patoa MPM. Ensemble feature ranking applied to medical data. Conf Electron Telecomm Comput-CETC. 2014.

23. Bamakan SMH, Gholami PA. Novel Feature Selection Method based on an Integrated Data Envelopment Analysis and Entropy Model. Procedia Comput Sci 2014; 31: 632-8.

24. Xue B, Zhang M, Browne WN. New fitness functions in binary particle swarm optimization for feature selection. Proc IEEE Congr Evol Comput (CEC). 2012

25. Xue B, Zhang M, Browne WN. Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms. Appl Soft Comput 2014; 18: 261-76.

26. Yao D, Yang J, Zhan X. A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines. J Compu 2013; 8(1): 170-7.

27. Lavanya D, Rani DKU. Analysis of feature selection with classification: Breast cancer datasets. Indian J Comput Sci Eng 2011; 2(5): 756-63.

28. Maldonado S, Weber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines. Inform Sci 2011; 181(1): 115-28.

29. Abonyi J, Szeifert F. Supervised fuzzy clustering for the identification of fuzzy classifiers. Pattern Recogn Lett 2003; 14(24): 2195-207.

30. Joachims T. Transductive inference for text classification using support vector machines. Proc Int Conf Mach Learn. Slovenia 1999.