مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال هشتم، شماره دوم، تابستان ۱۳۹۴

انتخاب ویژگی‌های موثر در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک یادگیری ماشین

راضیه شیخ‌پور*: دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران

مهدی‌آقا صرام: دانشکده مهندسی برق و کامپیوتر، دانشگاه یزد، یزد، ایران

چکیده

مقدمه: آزمایش آسپیراسیون سوزنی روشی کم هزینه، آسان و سریع برای تشخیص دقیق و زود هنگام سرطان پستان است. با استفاده از خصوصیات استخراج شده از آزمایش آسپیراسیون سوزنی و با کمک تکنیک‌های یادگیری ماشین می‌توان سیستمی کارآمد را برای تشخیص سرطان پستان طراحی نمود که با دقت بالایی خوش‌خیم یا بدخیم بودن تومورهای پستان را تشخیص دهند. هدف از انجام این مطالعه، انتخاب ویژگی‌های موثر در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک یادگیری ماشین است.

روش بررسی: در این مطالعه از داده‌های پایگاه داده WBCD موجود در UCI که شامل ۶۸۳ نمونه خوش‌خیم و بدخیم تومور پستان که هر نمونه دارای ۹ ویژگی است استفاده شد. سپس انتخاب ویژگی با روش پیش‌رو و دسته‌بندی نوع تومور با انواع روش‌های پارامتریک مانند دسته‌بندی درجه دو، دسته‌بندی خطی و دسته‌بندی نزدیک‌ترین میانگین انجام گرفت.

یافته‌ها: روش پارامتریک دسته‌بندی درجه دو با استفاده از انتخاب ویژگی پیش‌رو، بالاترین کارایی را در تشخیص سرطان پستان دارد. این روش با انتخاب چهار ویژگی Uniformity of cell size, Bare nuclei, Bland chromatin, Mitosesدارای دقت ۹۸/۹۰٪ و حساسیت ۹۷/۸۹٪ است. همچنین در همه روش‌ها ویژگی‌های Uniformity of cell size و Bare nuclei بالاترین کارایی را دارند.

نتیجه‌گیری: نتایج این مطالعه نشان داد که با روش انتخاب ویژگی پیش‌رو و تکنیک‌های پارامتریک یادگیری ماشین، علاوه بر دست‌یابی به عملکرد بالا در تشخیص سرطان پستان، عوامل و ویژگی‌های اصلی در تشخیص سرطان پستان نیز شناسایی می‌شوند. به نظر می‌رسد این ویژگی‌ها یکی از مهم‌ترین عوامل برای کمک به تشخیص سرطان پستان هستند.

واژه‌های کلیدی: سرطان پستان، یادگیری ماشین، انتخاب ویژگی، روش‌های پارامتریک.

* نشانی نویسنده پاسخگو: یزد، خیابان چمران، خیابان رهبر، کوچه ۱۷رهبر، پلاک ۹۱، راضیه شیخ‌پور.

نشانی الکترونیک: r_sheikhkhpour@stu.yazd.ac.ir

مقدمه

سرطان پستان شایع‌ترین سرطان در زنان است (۱،۲). در هر سال یک میلیون مورد جدید از این بیماری تشخیص داده می‌شود. علی‌رغم پیشرفت‌های چشمگیر در درمان، حدود ۲۵٪ بیماران مبتلا به سرطان پستان سالانه جان خود را به علت این بیماری از دست می‌دهند (۳). شیوع و میزان مرگ و میر سرطان پستان در نژادها و موقعیت‌های جغرافیایی مختلف متفاوت است و مکانیسم‌های مختلفی در شروع و پیشبرد آن نقش دارند. فاکتورهای محیطی متعدد، تغییرات سوماتیک مانند موتاسیون در آنکوژن‌ها، ژن‌های سرکوب کننده تومور و پلی‌مورفیسم‌های ژنتیکی از عوامل به وجود آورنده آن می‌باشند (۷-۳). در حال حاضر شانس بروز سرطان پستان در زنان آمریکایی یک نفر از هر ۸ تا ۹ نفر است و سالانه باعث مرگ حدود ۴۴۰۰۰ زن مبتلا می‌شود. این بیماری شایع‌ترین علت مرگ و میر ناشی از سرطان در زنان ایرانی است. اگرچه شیوع این بیماری در سنین قبل از ۲۵ تا ۳۰ سالگی نادر است اما بروز این سرطان در سنین کمتر نیز گزارش شده است (۸).

تومورهای پستان به دو دسته خوش‌خیم و بدخیم تقسیم می‌شوند. تومورهای خوش‌خیم به ندرت مرگ‌آور هستند ولی تعدادی از تومورهای خوش‌خیم پستان نیز می‌توانند خطر ابتلا به سرطان پستان را افزایش دهند. تومورهای بدخیم جدی‌تر بوده و سرطان محسوب می‌شوند. محققان میزان بالای مرگ و میر زنان بر اثر سرطان پستان را ناشی از تشخیص دیرهنگام این بیماری می‌دانند. با تشخیص زودهنگام سرطان پستان و پیشرفت‌های به‌دست آمده در درمان، میزان بقای بیماران مبتلا به این بیماری در حال افزایش است. تشخیص به موقع سرطان پستان (حداکثر ۵ سال پس از اولین تقسیم سلول سرطانی) شانس زنده بودن بیمار سرطانی را از ۵۶٪ به ۸۶٪ افزایش می‌دهد (۹). بنابراین وجود یک سیستم دقیق و مطمئن برای تشخیص به موقع و خوش‌خیم یا بدخیم بودن توده سرطان ضروری به نظر می‌رسد. اگرچه شایع‌ترین و قطعی‌ترین روش تشخیص سرطان پستان، بیوپسی سینه و تشخیص ضایعه با روش‌های معمول آسیب شناسی بافتی است، اما از آن جا که ۸۰-۷۰ درصد بیوپسی‌های جراحی مربوط به توده‌های خوش‌خیم پستان است، انجام این گونه روش جراحی علاوه بر اتلاف هزینه، سبب ایجاد اضطراب، استرس و تشویش در بیمار می‌شود (۱۰).

آزمایش آسپیراسیون سوزنی [1](FNA) روشی کم هزینه، آسان، سریع، دارای دقت بالا و تقریباً بدون عوارض جانبی است و به صورت سرپایی قابل انجام است (۱۱). در روش FNA، مایع استخراج شده از بافت پستان برای بررسی خصوصیات سیتولوژی مورد آزمایش قرار می‌گیرد. بعد از استخراج خصوصیات سیتولوژی بیمار باید بتوان خوش‌خیم یا بدخیم بودن توده را تشخیص داد (۱۲،۱۳). در مواردی‌که با قاطعیت نتوان خوش‌خیم یا بدخیم بودن بیماری را تشخیص داد، استفاده از الگوریتم‌های کامپیوتری و تکنیک‌های یادگیری ماشین راهنمای خوبی برای پزشک هستند (۱۴،۱۵). با استفاده از خصوصیات استخراج شده از آزمایش آسپیراسیون سوزنی و با کمک تکنیک‌های یادگیری ماشین می‌توان سیستمی کارآمد را برای تشخیص سرطان پستان طراحی نمود که با دقت بالایی سرطان پستان را تشخیص دهد.

تاکنون تحقیقات زیادی در رابطه با تشخیص سرطان پستان با کمک تکنیک‌های متفاوت یادگیری ماشین انجام شده است (۳۱-۱۶). در این مطالعه، با استفاده از روش‌های پارامتریک یادگیری ماشین نظیر روش دسته‌بندی درجه دو[2]، دسته‌بندی خطی[3]، دسته‌بندی نزدیک‌ترین میانگین[4] و روش انتخاب ویژگی پیش‌رو[5] بر روی داده‌های پایگاه داده WBCD[6] به شناسایی و انتخاب ویژگی‌های موثر در سرطان پستان پرداخته می‌شود.

مواد و روش‌ها

الف) انتخاب نمونه

این پژوهش توصیفی گذشته‌نگر است که مبتنی بر اطلاعات آسپیراسیون سوزنی پرونده‌های بیماران مبتلا به سرطان پستان در بیمارستان Wisconsin می‌باشد (۲۳). مجموعه داده‌های پایگاه WBCD شامل۶۹۹ نمونه

[1] Fine needle aspiration
[2] Quadratic discriminant
[3] Linear discriminant
[4] Nearest mean classifier
[5] Forward feature selection
[6] Wisconsin Breast Cancer Dataset

با ۱۰ ویژگی است. اولین مرحله در ایجاد هر مدلی بر اساس تکنیک‌های داده‌کاوی و یادگیری ماشین، مرحله پیش پردازش[7] می‌باشد. در پایگاه داده WBCD، ۱۶ نمونه با مقادیر گمشده[8] وجود دارد که در مرحله پیش پردازش، ابتدا شماره شناسه بیمار و نمونه‌های دارای مقادیر گمشده را حذف نموده و آزمایشات را با ۶۸۳ نمونه و ۹ ویژگی ادامه می‌دهیم. هر نمونه دارای یک برچسب خوش‌خیم یا بدخیم است. از ۶۸۳ نمونه مذکور، ۴۴۴ نمونه دارای برچسب خوش‌خیم (۶۵/۰۰۷٪) و ۲۳۹ نمونه دارای برچسب بدخیم (۳۴/۹۹۳٪) هستند. مقادیر ویژگی‌ها، عددی صحیح بین ۱ تا ۱۰ است. ویژگی‌های پایگاه داده WBCD در جدول ۱ نشان داده شده است.

ب) به‌کارگیری روش‌های پارامتریک در انتخاب ویژگی‌های موثر در تشخیص سرطان پستان

برای بررسی اثر هریک از ویژگی‌ها در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک دسته‌بندی درجه دو، دسته‌بندی خطی و دسته‌بندی نزدیکترین میانگین، از روش انتخاب ویژگی پیش‌رو استفاده نموده و بهترین ویژگی‌هایی که در هریک از مدل‌ها در تشخیص سرطان پستان نقش دارند را انتخاب می‌نماییم. برای این منظور، ابتدا در هریک از مدل‌ها، تنها با یک ویژگی نتایج را بررسی می‌کنیم و ویژگی با بالاترین دقت را انتخاب می‌نماییم، سپس بقیه ویژگی‌ها را به ویژگی انتخاب شده در مرحله قبل اضافه نموده و دقت مدل را با استفاده

جدول ۱: ویژگی‌های پایگاه داده WBCD

شماره ویژگی	نام ویژگی	شرح ویژگی	دامنه مقادیر ویژگی‌ها	میانگین	انحراف معیار
۱	ضخامت توده	سلول‌های خوش‌خیم تمایل به قرار گرفتن در یک لایه را دارند ولی سلول‌های سرطانی تمایل به قرار گرفتن در چند لایه را دارند	۱-۱۰	۴/۴۴	۲/۸۳
۲	یک‌نواختی اندازه سلول[1]	سلول‌های سرطانی اندازه متفاوتی دارند، در حالی‌که سلول‌های خوش‌خیم این گونه نیستند	۱-۱۰	۳/۱۵	۳/۰۷
۳	یک‌نواختی شکل سلول[1]	سلول‌های سرطانی شکل متفاوتی دارند، در حالی‌که سلول‌های خوش‌خیم این گونه نیستند	۱-۱۰	۳/۲۲	۲/۹۹
۴	چسبندگی حاشیه‌ای[1]	سلول‌های خوش‌خیم تمایل دارند که به یکدیگر متصل باشند ولی سلول‌های سرطانی تمایل دارند که این اتصال را از دست بدهند	۱-۱۰	۲/۸۳	۲/۸۶
۵	اندازه سلول‌های اپیتلیال[1]	سلول‌های اپیتلیال که به طور قابل توجهی بزرگ باشند، ممکن است یک سلول بدخیم باشند	۱-۱۰	۲/۲۳	۲/۲۲
۶	Bare nuclei	Bare nuclei واژه‌ای است که برای هسته‌هایی که توسط سیتوپلاسم احاطه نشده‌اند به کار می‌رود. آنها معمولا در سلول های خوش‌خیم دیده می‌شوند	۱-۱۰	۳/۵۴	۳/۶۴
۷	Bland chromatin	در سلول‌های خوش‌خیم، هسته (کروماتین) به شکل بافت یک‌دست و یک‌نواخت است ولی در سلول سرطانی چنین نیست	۱-۱۰	۳/۴۵	۲/۴۵
۸	Normal nucleoli	Normal nucleoli ساختارهای کوچکی هستند که در هسته سلول‌ها دیده می‌شوند. در سلول‌های خوش‌خیم این هسته‌ها خیلی کوچک هستند (اگر قابل دیدن باشند) ولی در سلول‌های سرطانی، این هسته‌ها بزرگ و مشخص هستند	۱-۱۰	۲/۸۷	۳/۰۵
۹	تقسیم میتوز	میتوز، تقسیم هسته سلول و ایجاد دو سلول دختر است. پاتولوژیست‌ها گرید سرطان را به وسیله شمارش تعداد تقسیم میتوز محاسبه می‌کنند	۱-۱۰	۱/۶۰	۱/۷۳

[7] Preprocessing
[8] Missing values

از ویژگی‌های دوتایی می‌سنجیم. در میان مجموعه ویژگی‌های دوتایی، مجموعه با بالاترین دقت را انتخاب نموده و بقیه ویژگی‌ها را به آن اضافه می‌کنیم و دقت مدل را می‌سنجیم. مراحل را به همین صورت ادامه می‌دهیم تا زمانی‌که با افزودن ویژگی به مجموعه قبلی، دقت تغییر نکند.

با توجه به ماهیت سیستم‌های مبتنی بر یادگیری، نمونه‌های این پژوهش در دو گروه آموزش (Train) و آزمون (Test) قرار می‌گیرند و روی داده‌های گروه اول فرآیند یادگیری انجام می‌شود. بعد از فرآیند یادگیری، روش طراحی شده با استفاده از داده‌های گروه آزمون سنجیده خواهد شد و با اعمال این داده‌های جدید به روش طراحی شده کارایی آن مشخص می‌شود. برای انتخاب ویژگی‌های موثر در تشخیص نوع تومور، تمام روش‌های پارامتریک دسته‌بندی درجه دو، دسته‌بندی خطی و دسته‌بندی نزدیک‌ترین میانگین با مجموعه آموزشی یکسانی آموزش داده شده و با مجموعه آزمون یکسانی مورد آزمایش قرار می‌گیرند. مطابق (۱۷) ۶۰/۰۳٪ داده‌ها (۲۶۶ نمونه خوش‌خیم و ۱۴۴ نمونه بدخیم) به‌عنوان مجموعه داده‌های آموزش و ۳۹/۹۷٪ داده‌ها (۱۷۸ نمونه خوش‌خیم و ۹۵ نمونه بدخیم) به‌عنوان مجموعه داده‌های آزمون در نظر گرفته می‌شوند.

یافته‌ها

در این مرحله، ابتدا روش‌های پارامتریک دسته‌بندی درجه دو، دسته‌بندی خطی و دسته‌بندی نزدیکترین میانگین بر روی تمام ویژگی‌های پایگاه داده WBCD مورد بررسی قرار گرفت و دقت، حساسیت و ویژگی هرکدام از روش‌ها به دست آمد. برای پیاده‌سازی روش‌های پارامتریک از نرم‌افزار متلب استفاده شد. جدول ۲ نتایج بررسی روش‌های پارامتریک بر روی داده‌های پایگاه WBCD را نشان می‌دهد. سپس به شناسایی مهم‌ترین ویژگی‌های تاثیرگذار در تشخیص سرطان پستان پرداختیم. برای این منظور، ابتدا به بررسی اثر هر یک از ویژگی‌ها در تشخیص سرطان پستان در هر یک از مدل‌ها پرداختیم. جدول۳، مهم‌ترین ویژگی دارای بالاترین دقت در هریک از مدل‌ها را نشان می‌دهد.

از آنجایی‌که ویژگی Uniformity of cell size در تمام مدل‌ها بالاترین دقت را در تشخیص سرطان دارد، بقیه ویژگی‌ها را حذف نموده، ویژگی Uniformity of cell size را انتخاب کرده و ویژگی‌های باقی‌مانده را به آن اضافه نموده و دقت مدل‌ها را می‌سنجیم. جدول ۴، دو ویژگی دارای بالاترین دقت در هریک از مدل‌ها را نشان می‌دهد. در مرحله بعد، در هر یک از مدل‌ها مجموعه ویژگی‌های دوتایی با بالاترین دقت را انتخاب نموده و ویژگی‌های باقیمانده را به آن اضافه کرده و دقت مدل را می‌سنجیم. جدول ۵، سه ویژگی دارای بالاترین دقت در هریک از مدل‌ها را نشان می‌دهد.

جدول ۲ : نتایج بررسی روش‌های پارامتریک بر روی پایگاه WBCD

نام روش	دقت	حساسیت	ویژگی
دسته‌بندی درجه دو	۹۷/۴۴	۹۷/۸۹	۹۷/۱۹
دسته‌بندی خطی	۹۸/۱۷	۹۴/۷۴	۱۰۰
دسته‌بندی نزدیکترین میانگین	۹۸/۵۳	۹۵/۷۹	۱۰۰

سپس مجموعه ویژگی‌های سه تایی با بالاترین دقت را انتخاب نموده و ویژگی‌های باقی‌مانده را به آن اضافه کرده و دقت مدل را می‌سنجیم. جدول ۶، چهار ویژگی دارای بالاترین دقت را در هریک از مدل‌ها نشان می‌دهد.

در مرحله بعد، مجموعه ویژگی‌های چهار تایی با بالاترین دقت را انتخاب نموده و ویژگی‌های باقی‌مانده را به آن اضافه کرده و دقت مدل را می‌سنجیم. جدول ۷، پنج ویژگی دارای بالاترین دقت را در هریک از مدل‌ها نشان می‌دهد. همان‌گونه که در جدول ۷ مشخص شده است، از آنجایی‌که در همه‌ی مدل‌ها با افزودن ویژگی‌های جدید به مجموعه چهار ویژگی به دست آمده در مرحله قبل، دقت افزایش نیافت، دیگر آزمایش‌ها را ادامه نداده و پی می‌بریم که در دسته‌بندی درجه دو با چهار ویژگی {Uniformity of cell size, Bare nuclei, Bland chromatin, Mitoses} و در دسته‌بندی خطی و دسته‌بندی نزدیکترین میانگین با چهار ویژگی {Uniformity of cell size, Bare nuclei, normal nucleoli, Marginal adhesion}

می‌توانیم بهترین پیش‌بینی را در مورد سرطان پستان بر روی داده‌های پایگاه WBCD انجام دهیم.

همان‌گونه که از نتایج جداول ۲ و ۶ مشخص می‌شود، علاوه بر شناسایی مهم‌ترین ویژگی‌ها، با استفاده از ویژگی‌های کمتر به عملکردی مشابه یا بهتر در مقایسه با تمام ویژگی‌ها دست یافتیم. در روش‌های یادگیری ماشین دست‌یابی به نتایج مشابه یا بهتر با استفاده از ویژگی‌های کمتر حائز اهمیت است.

جدول ۳: مهم‌ترین ویژگی دارای بالاترین دقت در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک

نام روش	ویژگی	دقت	حساسیت	ویژگی
دسته‌بندی درجه دو	{Uniformity of cell size}	۹۵/۹۷	۹۰/۵۳	۹۸/۸۸
دسته‌بندی خطی	{Uniformity of cell size}	۹۵/۹۷	۹۰/۵۳	۹۸/۸۸
دسته‌بندی نزدیکترین میانگین	{Uniformity of cell size}	۹۵/۹۷	۹۰/۵۳	۹۸/۸۸

جدول ۴: دو ویژگی دارای بالاترین دقت در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک

نام روش	ویژگی	دقت	حساسیت	ویژگی
دسته‌بندی درجه دو	{Uniformity of cell size, Bare nuclei}	۹۷/۸۰	۹۷/۸۹	۹۷/۷۵
دسته‌بندی خطی	{Uniformity of cell size, Bare nuclei}	۹۷/۴۴	۹۳/۶۸	۹۹/۴۴
دسته‌بندی نزدیکترین میانگین	{Uniformity of cell size, Bare nuclei}	۹۷/۴۴	۹۳/ ۶۸	۹۹/۴۴

جدول ۵: سه ویژگی دارای بالاترین دقت در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک

نام روش	ویژگی	دقت	حساسیت	ویژگی
دسته‌بندی درجه دو	{Uniformity of cell size, Bare nuclei, Bland chromatin}	۹۸/۵۳	۹۶/۸۴	۹۹/۴۴
دسته‌بندی خطی	{Uniformity of cell size, Bare nuclei, Normal nucleoli}	۹۸/۱۷	۹۴/۷۴	۱۰۰
دسته‌بندی نزدیکترین میانگین	{Uniformity of cell size, Bare nuclei, Normal nucleoli}	۹۷/۸۰	۹۳/۶۸	۱۰۰

جدول ۶: چهار ویژگی دارای بالاترین دقت در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک

نام روش	ویژگی	دقت	حساسیت	ویژگی
دسته‌بندی درجه دو	{Uniformity of cell size, Bare nuclei, Bland chromatin, Mitoses}	۹۸/۹۰	۹۷/۸۹	۹۹/۴۴
دسته‌بندی خطی	{Uniformity of cell size, Bare nuclei, Normal nucleoli, Marginal adhesion}	۹۸/۵۳	۹۵/۷۹	۱۰۰
دسته‌بندی نزدیکترین میانگین	{Uniformity of cell size, Bare nuclei, Normal nucleoli, Marginal adhesion}	۹۸/۵۳	۹۵/۷۹	۱۰۰

جدول ۷: پنج ویژگی دارای بالاترین دقت در تشخیص سرطان پستان با استفاده از مدل‌های پارامتریک

نام روش	ویژگی	دقت	حساسیت	ویژگی
دسته‌بندی درجه دو	{Uniformity of cell size, Bare nucleoli, Bland chromatin, Mitoses, Clump thickness}	۹۸/۹۰	۹۷/۸۹	۹۹/۴۴
دسته‌بندی خطی	{Uniformity of cell size, Bare nucleoli, Normal nucleoli, Marginal adhesion, Bland chromatin}	۹۸/۵۳	۹۵/۷۹	۱۰۰
دسته‌بندی نزدیکترین میانگین	{Uniformity of cell size, Bare nucleoli, Normal nucleoli, Marginal adhesion, Bland chromatin}	۹۸/۵۳	۹۵/۷۹	۱۰۰

بحث

هدف اصلی مطالعه حاضر شناسایی مهم‌ترین ویژگی‌ها در تشخیص سرطان پستان با استفاده از روش‌های پارامتریک یادگیری ماشین است. سیستم پیشنهادی این مطالعه شامل انتخاب ویژگی پیش‌رو و دسته‌بندی نوع تومور با انواع روش‌های پارامتریک دسته‌بندی درجه دو، دسته‌بندی خطی و دسته‌بندی نزدیک‌ترین میانگین می‌باشد.

روش پارامتریک دسته‌بندی درجه دو با شناسایی چهار ویژگی مهم و دقت۹۸/۹۰٪ و حساسیت ۹۷/۸۹٪ دارای عملکرد بالاتری نسبت به روش‌های دسته‌بندی خطی و نزدیک‌ترین میانگین با دقت ۹۸/۵۳٪ و حساسیت ۹۵/۷۹ در تشخیص سرطان پستان است.

ویژگی‌های {Uniformity of cell size, Bare nuclei, Bland chromatin, Mitoses} بالاترین کارایی را در روش دسته‌بندی درجه دو و ویژگی‌های {Uniformity of cell size, Bare nuclei, Normal nucleoli, Marginal adhesion} بالاترین کارایی را در روش‌های دسته‌بندی خطی و نزدیک‌ترین میانگین دارند. در تمامی روش‌های انجام شده، ویژگی‌های Uniformity of cell size و Bare nuclei نسبت به سایر ویژگی‌ها در تشخیص سرطان پستان از دقت بالاتری برخوردار بودند، بنابراین به نظر می‌رسد این ویژگی‌ها اثر مهمی در تشخیص سرطان پستان داشته باشند.

در این مطالعه علاوه بر شناسایی مهم‌ترین ویژگی‌ها، با استفاده از ویژگی‌های کمتر به عملکردی یکسان یا بهتر بر حسب شاخص‌های دقت، حساسیت و ویژگی در مقایسه با تمام ویژگی‌ها دست یافتیم. اگر اطلاعات پایگاه WBCD شامل ویژگی‌های دیگری مانند سن افراد، سن اولین بارداری و تعداد بارداری و... بود، اطلاعات جامع‌تر و کامل‌تر برای انتخاب ویژگی‌های موثر در تشخیص سرطان پستان در اختیار داشتیم.

فلاحتی و جعفری درسال ۲۰۱۱ سیستم خبره‌ای برای تشخیص سرطان پستان با استفاده از داده‌های پیش پردازش و شبکه بیزین طراحی نمودند که توانست با دقت ۹۸/۱٪ داده‌های پایگاه WBCD را به درستی تشخیص دهد (۲۹). Aruna و همکاران به بررسی و مقایسه دسته‌بندی‌کننده‌های یادگیری با نظارت نظیر روش بیزین ساده، SVM-RBF، شبکه عصبیRBF، درخت تصمیم J48 وCART برروی مجموعه داده‌های پایگاه WBCD پرداختند تا بهترین دسته‌بندی‌کننده را بر روی این داده‌ها پیدا نمایند. نتایج آزمایشات آنها نشان داد که SVM-RBF با دقت ۹۶/۸۴٪ نسبت به دسته‌بندی‌کننده‌های دیگر دقت بیشتری دارد (۳۳). Medjahed و همکاران با استفاده از روش دسته‌بندی نزدیک‌ترین همسایه و معیارهای مختلف فاصله به دسته-بندی داده‌های پایگاه WBCD پرداختند. آنها با استفاده از فاصله اقلیدسی و k=1 به دقت ۹۸/۷٪ دست یافتند (۳۴). Kiyan و همکاران در سال ۲۰۰۳ به بررسی تکنیک‌های RBF، GRNN و PNN بر روی مجموعه داده‌های WBCD پرداختند. نتایج آزمایشات آنها نشان داد که روش RBF دارای دقت ۹۶/۱۸٪، روش PNN دارای دقت ۹۸/۸٪ و روش MLP دارای دقت ۹۵/۷۴٪ است (۳۵). Chen و همکاران در سال ۲۰۱۱ با استفاده از

روش SVM-RS بر روی پایگاه WBCD نشان دادند که ویژگی‌های {Uniformity of cell shape, Marginal adhesion, Bare nuclei, Mitoses, Clump thickness} از اهمیت بیشتری برخوردارند (۳۶). Marcano-Cedeño و همکاران در سال ۲۰۱۱ با بهبود در آموزش شبکه عصبی برای دسته‌بندی الگوها به دقت ۹۹/۲۶٪ دست یافتند. الگوریتم پیشنهادی آنها (MLP{Uniformity of cell size}AM) از تئوری اطلاعات شانون الهام گرفته بود (۱۷). Chaurasia و Chakrabarti در سال۲۰۱۳ روشی برای تشخیص سرطان پستان با استفاده از ماشین بردار پشتیبان ارایه دادند. روش پیشنهادی آنها بر روی داده‌های پایگاه WBCD به دقت ۹۶/۴٪ دست یافت. در مطالعه آنها دقت روش‌های مختلف بر روی داده‌های این پایگاه بین ۹۴/۷۷٪ و ۹۸/۸٪ گزارش شد (۳۷). Palaniappan و Pushparaj در سال ۲۰۱۳ روشی برای تشخیص سرطان پستان با استفاده از شبکه عصبی و قوانین همبستگی پیشنهاد دادند که با نرخ ۹۸/۴٪ داده‌های پایگاه WBCD را به درستی پیش بینی نمود (۳۸). Tan و همکاران یک تکنیک دسته‌بندی دو مرحله‌ای ترکیبی برای استخراج قوانین دسته‌بندی ارایه دادند. روش پیشنهادی آنها به دقت ۹۷/۵۷٪ بر روی پایگاه WBCD دست یافت (۳۹). در مطالعه Lavanya و همکاران در سال ۲۰۱۱، کارایی دسته‌بندی‌کننده درخت تصمیم CART بدون انتخاب ویژگی بر روی پایگاه‌های WBCD بررسی شد. این محققان با تکنیک دسته‌بندی دو مرحله‌ای ترکیبی به دقت ۹۴/۸۴٪ رسیدند (۴۰)، در حالی که مطالعه ما با روش پارامتریک دسته‌بندی درجه دو و انتخاب ویژگی پیش‌رو به دقت ۹۸/۹۰٪ رسید. بدین ترتیب با روش دسته‌بندی درجه دو، دقت شناسایی سیستم‌های تشخیص سرطان افزایش یافت، این درحالی است که نسبت به سیستم‌های مشابه از تعداد کمتری ویژگی استفاده شده است. کاهش تعداد ویژگی‌ها در سیستم‌های تشخیص هوشمند باعث افزایش سرعت، ساده‌سازی و کاهش پیچیدگی مدل می‌شود.

نتیجه‌گیری

در این مطالعه، با استفاده از روش انتخاب ویژگی پیش‌رو و تکنیک‌های پارامتریک یادگیری ماشین، علاوه بر دست‌یابی به دقت بالا در تشخیص بیماری‌ها، عوامل و ویژگی‌های اصلی در تشخیص سرطان پستان نیز شناسایی شدند. به نظر می‌رسد این ویژگی‌ها یکی از مهم‌ترین عوامل برای کمک به تشخیص زودهنگام سرطان پستان هستند.

References

1. Sheikhpour R, Ghasemi N, Yaghmaei P, Mohiti J. Immunohistochemical assessment of p53 protein and its correlation with clinicopathological parameters in breast cancer patients. Indian Journal Science and Technology 2014; 7(4): 472-9.
2. 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. International Journal of Cancer 2008; 123(7):1536–44.
3. Sheikhpour R, Taghipour Zahir S. Evaluation of Tp53 codon72 polymorphism and resulted protein in breast cancer patients in Yazd city. Iranian Journal of Breast Disease 2014; 7 (3): 20-9.
4. Mousavi SM, Montazeri A, Mohagheghi MA, Jarrahi AM, Harirchi I, Najafi M, Ebrahimi M, Furuwatari C, Yagi A, Yamagami O. Breast cancer in Iran: an epidemiological review. Breast J 2007; 13(4):383-91.
5. Lakhani SR, Vijver MJ, Jacquemier J, Anderson TJ, Osin PP, McGuffog L, Douglas F. The Pathology of Familial Breast Cancer: Predictive Value of Immunohistochemical Markers Estrogen Receptor, Progesterone Receptor, HER-2, and p53 in Patients with Mutations in BRCA1 and BRCA2. J Clin Oncol 2002; 20(9): 2310-8.

6. Sheikhpour R, Mohiti Ardekani J. The effect of progesterone on p53 protein in T47D cell line. J Urmia Uni Med Sci 2014; 25(10): 954-60.

7. Vojtesek B, Lane DP. Regulation of p53 protein expression in human breast cancer cell lines. J Cell Sci 1993; 105: 607-12.

8. Sheikhpour R, Agha Sarram M, Zare Mirakabad M, Sheikhpour R. Breast Cancer Detection Using Two-Step Reduction of Features Extracted From Fine Needle Aspirate and Data Mining Algorithms. Iranian Journal of Breast Disease 2015; 7(4): 43-51.

9. Alipour M, Hadadnia J. An Accurate Intelligent Breast Cancer Diagnosis System. Iranian Journal of Breast Disease 2009; 2(2):33-40.

10. Litigate J. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. Clinical Cancer Research 2004; 10(8): 2725-37.

11. Maglogiannis I, Zafiropoulos E, Anagnostopoulos I. An intelligent system for automated breast cancer diagnosis and prognosis using SVM based classifiers. Applied intelligence 2009; 30(1): 24-36.

12. Khooei AR, Mehrabi Bahar M, Ghaemi M, Mirshahi M. Sensitivity and specificity of CNB in diagnosis of breast masses. Iranian journal of Basic Medical Sciences 2005; 8(2): 6-100.

13. Zhaohui L, Xiaoming W, Shengwen G, Binggang Y. Diagnosis of breast cancer tumor based on manifold learning and support vector machine. Proceeding of International Conference on Information and Automation 2008; 703-7.

14. Alipoour M, Haddadnia J. An Accurate Intelligent Breast Cancer Diagnosis System. Iranian Journal of Breast Disease 2009; 2(2): 33-40.

15. Ghayomi zadeh H, Droudgar moghadam A, Hadad nia J. Clustering of breast cancer via thermal images using a combination of MLP and SOM neural network. Iranian Journal of breast Disease 2012; 5(2,3): 70-83.

16. Jesmin Nahar,Tasadduq Imam, Kevin S. Tickle, A.B.M. Shawkat Ali, Yi-Ping Phoebe Chen. Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer. Expert System with Applications 2012; 39:12371- 7.

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

18. Salama G. I, Abdelhalim M. B, Zeid M. A. E. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. International Journal of computer and Information Technology 2012; 1(1): 36-43.

19. Georgiou VL, Malefaki SN, Alevizos PD, Vrahatis MN. Evolutionary Bayesian Probabilistic Neural Networks. Proceeding of International Conferance on Numerical Analysis and Applied Mathematics 2006; 393-6.

20. Nattkemper TW, Arnrich B, Lichte O, Timm W, Degenhard A, Pointon L, ... & Leach MO. Evaluation of radiological features for breast tumour classification in clinical screening with machine learning methods. Artificial Intelligence in Medicine 2005; 34: 129-39.

21. Xu Y, Zhu Q, Wang J. Breast cancer diagnosis based on a kernel orthogonal transform. Neural Computing and Applications 2012; 21(8): 1865-70.

22. Jaganathan P, Rajkumar N, Nagalakshmi R. A Kernel Based Feature Selection Method Used in the Diagnosis of Wisconsin Breast Cancer Dataset. In Advances in Computing and Communications 2011; 683-90.

23. Sotiriou C, Neo SY, McShane LM, Korn EL, Long PM, Jazaeri A, Martiat P, Fox SB, Harris A, Liu ET. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciences 2003; 100(18): 10393-8.

24. Rajesh K, Anand S. Analysis of SEER Dataset for Breast Cancer Diagnosis using C4. 5 Classification Algorithms. International Journal of Advanced

Research in Computer and Communication Engineering 2012; 1.

25. Mangasarian OL. Unsupervised classification via convex absolute value inequalities. Data Mining Institute Technical Report 2014; 14-01.

26. Salama GI, Abdelhalim MB, Zeid MAE. Experimental comparison of classifiers for breast cancer diagnosis. In Proceeding of Seventh International Conference on Computer Engineering & Systems (ICCES) 2012; 180-5.

27. Naghibi S, Teshnehlab M, Shoorehdeli MA. Breast cancer classification based on advanced multi dimensional fuzzy neural network. Journal of medical systems. 2012; 36(5): 2713-20.

28. Alickovic E, Subasi A. Usage of Simplified Fuzzy ARTMAP for improvement of classification performances. South East Europe Journal of Soft Computing 2013; 2(2).

29. Fallahi A, Jafari S. An expert system for detection of breast cancer using data preprocessing and Bayesian network. Int J Adv Sci Technol 2011; 34: 65-70.

30. Zhang L, Wang L, Wang X, Liu K, Abraham A. Research of neural network classifier based on FCM and PSO for breast cancer classification. In Hybrid Artificial Intelligent Systems 2012; 647-54.

31. Jacob SG, Ramani RG. Efficient classifier for classification of prognostic breast cancer data through data mining techniques. In Proceedings of the World Congress on Engineering and Computer Science 2012; 1.

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

33. Aruna S, Rajagopalan DS, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology 2011; 2: 37-45.

34. Medjahed SA, Saadi TA, Benyettou A. Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules. International Journal of Computer Application 2013; 62(1): 1-5.

35. Kiyan T, Yildirim T. Breast cancer diagnosis using statistical neural networks. IU-Journal of Electrical & Electronics Engineering 2011; 4(2), 1149-53.

36. Chen H-L, Bo Yang Jie Liu, Da-You Liu. A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis. Expert System Application 2011; 38: 9014-22.

37. Chaurasia S, Chakrabarti P. An Approach with Support Vector Machine using Variable Features Selection on Breast Cancer Prognosis. International Journal of Advanced Research in Artificial Intelligence 2013; 2(9): 38-42.

38. Palaniappan S, Pushparaj T. A Novel Prediction on Breast Cancer from the Basis of Association rules and Neural Network. International Journal of Computer Science and Mobile Computing–IJCSMC 2013; (4): 269-77.

39. Tan KC, Yu Q, Heng CM, Lee TH. Evolutionary computing for knowledge discovery in medical diagnosis. Artificial Intelligence in Medicin 2003; 27(2): 129-54.

40. Lavanya D, Rani KU. Ensemble decision tree classifier for breast cancer data. International Journal of Information Technology Convergence and Services 2012; 2(1): 17-24.