مجله پیاورد سلامت
دوره ۱۶، شماره ۲، خرداد ۱۴۰۱، ۱۵۸-۱۴۵
مقاله پژوهشی

طراحی سیستم تصمیم‌یار بالینی تشخیص سرطان پستان: رویکردی مبتنی بر داده کاوی

مصطفی شنبه زاده[۱]، هادی کاظمی آرپناهی[۲]، رئوف نوپور[۳*]

دریافت مقاله: ۱۴۰۰/۱۲/۲۶
پذیرش مقاله: ۱۴۰۱/۴/۲۹

* نویسنده مسئول:
رئوف نوپور؛
دانشکده مدیریت و اطلاع‌رسانی پزشکی دانشگاه علوم پزشکی ایران

Email :
nopoor.r@iums.ac.ir

چکیده

زمینه و هدف: سرطان پستان یکی از رایج‌ترین و تهاجمی‌ترین بدخیمی‌ها در خانم‌ها می‌باشد. تشخیص به‌موقع سرطان پستان نقش مهمی در جلوگیری از پیشرفت این بیماری، اقدامات درمانی به‌موقع و در نتیجه کاهش میزان مرگ‌ومیر این بیماران دارد. یادگیری ماشین، قابلیت بالایی در تشخیص سریع و هزینه اثربخش بیماری‌ها دارد. هدف این مطالعه، طراحی سیستم تصمیم‌یار بالین Clinical Decision Support System (CDSS) بر اساس قوانین استخراج‌شده از الگوریتم منتخب درخت تصمیم با بهترین عملکرد به‌منظور تشخیص به‌موقع و مؤثر سرطان پستان است.

روش بررسی: داده‌های ۵۹۷ فرد مشکوک به سرطان پستان(۲۵۵ بیمار مبتلا و ۳۴۲ فرد سالم) به‌صورت گذشته‌نگر از پایگاه داده الکترونیکی بیمارستان آیت‌الله طالقانی شهر آبادان در قالب ۲۴ ویژگی عمدتاً سبک زندگی و سوابق پزشکی استخراج شد. پس از انتخاب مهم‌ترین متغیرها از طریق کای دو پیرسون و تحلیل واریانس یک‌طرفه($P<0/05$)، عملکرد الگوریتم‌های منتخب داده‌کاوی شامل Random Forest (RF)، J-48، Decision Stump (DS)، Rep-Tree (RT) و XG-Boost برای تشخیص سرطان پستان در بستر نرم‌افزار Weka 3.4 تحلیل شد. در نهایت سیستم تشخیصی سرطان پستان بر اساس بهترین مدل و از طریق زبان برنامه‌نویسی سی شارپ و چارچوب Dot Net Framework 3.5.4 طراحی گردید.

یافته‌ها: ۱۴ متغیر شامل سابقه‌ی فردی سرطان پستان، سابقه‌ی نمونه‌برداری از سینه، سابقه‌ی رادیوگرافی از قفسه‌ی سینه، سابقه‌ی فشارخون، افزایش کلسترول خون LDL (low-density lipoprotein)، وجود توده در ربع فوقانی داخلی سینه، هورمون‌درمانی با استروژن، هورمون‌درمانی با استروژن-پروژسترون، سابقه‌ی خانوادگی سرطان پستان، سن، سابقه‌ی سرطان‌های دیگر، نسبت اندازه‌ی دور کمر به دور باسن و مصرف میوه و سبزی ارتباط معناداری را باکلاس خروجی در سطح $P<0/05$ نشان دادند. بر اساس نتایج حاصل از ارزیابی عملکرد الگوریتم‌های منتخب، مدل RF با میزان حساسیت، ویژگی، صحت و اندازه F به‌ترتیب برابر با ۰/۹۷، ۰/۹۹، ۰/۹۸ و ۰/۹۷۴ و ۰/۹۳۶ =Area Under the Receiver Operator Characteristics (ROC) Curve (AUC) عملکرد بالاتری نسبت به سایر الگوریتم‌های منتخب داشته است و به‌عنوان مدل برتر برای تشخیص سرطان پستان پیشنهاد شد.

نتیجه‌گیری: به‌نظر می‌رسد که استفاده از متغیرهای تعدیل‌پذیر مانند سبک زندگی و ویژگی‌های هورمونی-تولیدمثلی به‌عنوان ورودی الگوریتم RF برای طراحی CDSS بتواند با صحت بهینه موارد سرطان پستان را تشخیص دهد. به‌علاوه سیستم پیشنهادی به‌طور مؤثر در محیط‌های واقعی بالینی برای تشخیص سریع و مؤثر بیماری قابل اقتباس باشد.

واژه‌های کلیدی: سرطان پستان، داده‌کاوی، مدل تشخیصی، سیستم تصمیم‌یار بالینی

۱ استادیار گروه فناوری اطلاعات سلامت، دانشکده پیراپزشکی، دانشگاه علوم پزشکی ایلام، ایلام، ایران

۲ استادیار گروه فناوری اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی آبادان، آبادان، ایران

۳ دانشجوی دکتری مدیریت اطلاعات سلامت، دانشکده مدیریت و اطلاع رسانی پزشکی، دانشگاه علوم پزشکی ایران، تهران، ایران

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

مقدمه

سرطان پستان یکی از بدخیمی‌های رایج و کشنده در میان خانم‌هاست که دارای قابلیت بالای تهاجمی و گسترشی(متاستاتیک) به بافت‌ها و اندام‌های اطراف است. این سرطان بر اساس آمار GLOBOCAN در سال ۲۰۲۱، با میزان بروز حدود ۱۱/۷ درصد از کل موارد جدید سرطان(۲۲۶۱۴۱۹ نفر)، به‌عنوان یکی از رایج‌ترین بدخیمی‌های تشخیص داده‌شده گزارش شد. همچنین این بدخیمی با ۶/۹ درصد از کل مرگ‌ومیرهای در ارتباط با سرطان(۶۸۴۹۹۶ نفر)، به‌عنوان پنجمین عامل مرگ‌ومیر ناشی از این بیماری‌ها در این سال شناخته شد(۱). بر اساس اعلام سازمان جهانی بهداشت (WHO: World Health Organization) تا پایان سال ۲۰۳۰ شیوع سرطان پستان به‌دلیل تغییر عمده در سبک زندگی مردم به شکل قابل‌توجهی افزایش می‌یابد(۳و۲). شناسایی دقیق و به‌موقع این بیماری در قالب برنامه‌های غربالگری نظام‌مند و جامع می‌تواند موجب کاهش عوارض وخیم و مرگ‌ومیرهای ناشی از آن شود. روش‌های معمول غربالگری این بیماری شامل ماموگرافی، ترموگرافی، سونوگرافی و نمونه‌برداری از پستان است(۶-۴). متاسفانه استفاده از این روش‌ها غالباً دشوار و گران‌قیمت است و در بسیاری از جوامع و به‌خصوص جوامع درحال‌توسعه به‌دلیل مشکلات برنامه‌ریزی و نبود زیرساخت‌های لازم در دسترس عموم نیست. این باعث شده تا شیوع و مرگ‌ومیرهای ناشی از سرطان پستان در جوامع درحال‌توسعه بیشتر باشد. به‌علاوه به دلیل عدم استفاده این فنون از داده‌های زمان‌محور و تجمیعی، بروز خطا در تشخیص و شناسایی موارد اجتناب‌ناپذیر است(۱۰-۷).

یادگیری ماشین به‌عنوان یکی از زیرمجموعه‌های هوش‌مصنوعی از طریق استخراج الگوهای قابل‌فهم و دانش کاربردی از مجموعه داده‌های خام، موجب حمایت از تشخیص‌های پزشکی خواهد شد(۱۲و۱۱). طراحی سیستم‌های تصمیم‌یار بالینی (CDSS: Clinical Decision Support System) مبتنی بر فنون یادگیری ماشین به‌منظور شناسایی، اولویت‌بندی و ارایه راهکارهای درمانی سفارشی برای بیماران مبتلا به سرطان پستان در مرحله اول و پیش‌آگهی ابتلا و پیش‌بینی خطرات پیشرفت بیماری و مرگ‌ومیرهای ناشی از آن در مراحل بعدی ضروری است(۱۳و۱۱). پژوهش‌های متعددی در خصوص کاربرد فناوری‌های یادگیری ماشین در پیش‌آگهی، غربالگری، تشخیص و درمان بدخیمی‌های سینه انجام‌شده است. تحلیل نتایج مطالعات پیشین حاکی از عملکرد مناسب و صحت بهینه‌ی این مدل‌ها در تشخیص و غربالگری موثر سرطان پستان بوده‌است(۱۵و۱۴و۱۲). باوجود این بیشتر مطالعات از متغیرهای بالینی و ارزیابانه(آزمایشگاهی و تصویربرداری) برای طراحی سیستم استفاده کرده‌اند(۱۷و۱۶). باتوجه به در دسترس نبودن این داده‌ها برای بسیاری از افراد جامعه، در پژوهش حاضر مدل پیش‌گویانه بر اساس متغیرهای آسان‌یاب سبک زندگی، سابقه‌ای و ویژگی‌های هورمونی-تولیدمثلی طراحی گردید(۱۳). در نتیجه تعمیم‌پذیری، دسترس‌پذیری و کاربردپذیری سیستم از طریق پیاده‌سازی آن بر بستر اینترنت برای استفادهی عموم افزایش خواهد یافت. به‌علاوه اثبات شده که طراحی فناوری‌های CDSS بر اساس عوامل سبک زندگی و سوابق پزشکی افراد، قابلیت‌های تحلیلی اثربخشی را برای تشخیص به‌موقع و موثر خطرات بدخیمی سینه ارایه خواهد داد(۱۸). این الزام از این جهت مهم است که بسیاری از این عوامل خطر از نوع قابل‌تعدیل و قابل‌اصلاح هستند. طبقه‌بندی موثر این عوامل خطر از طریق فناوری‌های CDSS برای شناسایی فعالانه بیماران در معرض خطر، تشخیص زودهنگام بیماری و اتخاذ تصمیم برای پیشگیری از پیشرفت بیماری و مدیریت درمان، بسیار ضروری است(۲۱-۱۹). ازاین‌رو هدف مطالعه‌ی حاضر، ایجاد یک مدل تشخیصی هوشمند برای شناسایی سریع موارد سرطان پستان از طریق مقایسه فنون منتخب یادگیری ماشین بر روی متغیرهای عمدتاً غیربالینی است. گام‌های پژوهش حاضر عبارتند از: ۱- شناسایی و اولویت‌بندی متغیرهای پراهمیت در تشخیص بیماری، ۲- آموزش و مقایسه‌ی عملکردی مدل‌های منتخب یادگیری ماشین و ۳- طراحی رابط کاربری CDSS بر اساس بهترین الگوریتم یادگیری ماشین.

روش بررسی

هدف این مطالعه که از نوع توسعه‌ای بود، طراحی سیستم تشخیصی سرطان پستان با استفاده از الگوریتم‌های داده‌کاوی مبتنی بر قوانین بود که در سه مرحله و به شرح زیر انجام شد:

- **جمع‌آوری و توصیف پایگاه داده**

جامعه‌ی پژوهش شامل افرادی بودند که به بیمارستان آیت‌الله طالقانی شهرستان آبادان وابسته به دانشگاه علوم پزشکی آبادان مراجعه کرده بودند و نتیجه‌ی بررسی‌های تشخیصی توسط پزشک در پرونده پزشکی الکترونیکی (Electronic Medical Record, EMR) برای این دسته از افراد مثبت(دارای سرطان پستان) و منفی(فاقد سرطان پستان) گزارش شده بود. به‌منظور تحلیل داده‌ها و تولید مدل مبتنی بر قوانین، از مجموعه داده‌های موجود در پایگاه

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

داده‌ی آن بیمارستان استفاده گردید. به‌ترتیب، تعداد ۲۵۵ و ۳۴۲ نمونه زنان مرتبط با تشخیص مثبت و منفی سرطان پستان طی سال‌های ۱۳۹۹-۱۳۹۷ در پایگاه داده‌ی آن مرکز موجود بود و مورد استفاده قرار گرفت. تشخیص منفی با کد صفر و تشخیص مثبت با کد یک به‌عنوان متغیر خروجی(نتایج تشخیصی) در پایگاه داده آن مرکز موجود بودند. تمامی متغیرهای تشخیصی در جدول ۱ نشان داده شده اند.

جدول ۱: عوامل تشخیصی موثر در شناسایی سرطان پستان

نوع متغیر	دسته	نام متغیر(حالات و واحدها)	مفاهیم کدهای موجود در پایگاه داده
ورودی	**عوامل جمعیت‌شناسی**	سن(برحسب سال)، نسبت قد به وزن، نسبت اندازه دور کمر به باسن	–
	عوامل همه‌گیرشناسی	مصرف الکل(برحسب گرم در روز)، سابقه پیاده‌روی زیاد(دارد، ندارد)، شغل سخت و زیان‌آور(دارد، ندارد)، میزان فعالیت فیزیکی(برحسب ساعت در روز)، چاقی(دارد، ندارد)	دارد: کد ۱، ندارد: کد ۲ مصرف الکل: کمتر از ۵۰ گرم، بین ۵۰ تا ۱۵۰ گرم، بیش از ۱۵۰ گرم میزان فعالیت فیزیکی در روز: کمتر از نیم ساعت، نیم ساعت تا ۱ ساعت، بیشتر از ۱ ساعت
	عوامل تغذیه‌ای	مصرف سبزیجات(برحسب گرم در روز)، مصرف میوه(برحسب گرم در روز)	مصرف میوه: کمتر از ۱۰۰ گرم: کد ۱، بین ۱۰۰ تا ۲۰۰ گرم: کد ۲، بیشتر از ۲۰۰ گرم: کد ۳. مصرف سبزیجات: کمتر از ۱۵۰ گرم: کد ۱، بین ۱۵۰ تا ۳۰۰ گرم: کد ۲، بیشتر از ۳۰۰ گرم: کد ۳.
	سابقه‌ی بیماری‌های فردی	شخصی سرطان پستان(دارد، ندارد)، نمونه‌برداری از پستان(دارد، ندارد)، رادیوگرافی از قفسه سینه(دارد، ندارد)، فشارخون(دارد، ندارد)، افزایش کلسترول (LDL low-density lipoprotein:)، تری گلیسیرید و کلسترول خون (VLDL: Very low-density lipoprotein) خون(دارد، ندارد)، دیابت(دارد، ندارد)، وجود توده در ربع فوقانی پستان(دارد، ندارد)، هورمون درمانی با استروژن(دارد، ندارد)، وجود کیست در پستان(دارد، ندارد)، هورمون درمانی یا استروژن- پروژسترون(دارد، ندارد)	دارد: کد ۱، ندارد: کد ۲
	سابقه‌ی بیماری‌های خانوادگی	سابقه خانوادگی سرطان پستان(دارد، ندارد)، سابقه خانوادگی سرطان‌های دیگر(دارد، ندارد)	دارد: کد ۰، ندارد: کد ۱
خروجی	**نتیجه تشخیصی**	سرطان پستان(دارد، ندارد)	دارد: کد ۱، ندارد: کد ۰

عوامل دخیل در تشخیص سرطان پستان شامل ۲۴ متغیر بودند که در طبقات عوامل جمعیت‌شناسی، سوابق خانوادگی، سوابق شخصی، عوامل تغذیه‌ای و همه‌گیرشناسی قرار می‌گرفتند.

- **تحلیل پایگاه داده**

پس از کسب مجموعه داده و شناسایی متغیرهای تشخیص سرطان پستان و بررسی آن‌ها، داده‌ها نرمال‌سازی شدند. ابتدا نویسندگان پژوهش از طریق دریافت مشورت از سه متخصص در حوزه‌ی زنان و زایمان و یک متخصص سرطان‌شناسی آن مرکز که به تعداد ۳ نفر بودند، کل پایگاه داده را از لحاظ وجود مقادیر پرت مورد بررسی قراردادند و ارزش متغیرهایی که پرت بود با مقادیر خالی جایگزین شدند. در مرحله‌ی بعد نمونه‌هایی که بیش از ۷۰ درصد مقادیر مفقود(از ابتدا خالی و یا با حذف مقادیر مفقود) داشتند از پایگاه داده‌ی پژوهش حذف گردیدند و برای نمونه‌های با کمتر از ۷۰ درصد مقادیر مفقود از روش series mean برای جایگزینی داده‌های کمی و تعیین نسبت فراوانی برای متغیرهای کیفی استفاده گردید. در روش series mean تمامی متغیرهای مفقود با میانگین متغیرهای کمی جایگزین شدند. در روش تعیین نسبت فراوانی، ابتدا نسبت فراوانی تمامی متغیرهای کیفی اندازه گرفته شد و مقادیر مفقود به همان نسبت جایگزین شدند.

- **انتخاب ویژگی**

به‌منظور کاهش ابعاد داده و انتخاب بهترین مجموعه از متغیرهای تشخیصی سرطان پستان از روش انتخاب ویژگی استفاده گردید. این روش، علاوه بر کاهش ابعاد داده و فهم بهتر مجموعه داده موجب افزایش عملکرد الگوریتم، جلوگیری از بیش برازش، افزایش سرعت آموزش الگوریتم و قدرت محاسبات ماشین می‌شود. در این پژوهش از روش کای دو پیرسون و تحلیل واریانس یک‌طرفه در نرم‌افزار IBM SPSS Statistics V25 برای تعیین مهم‌ترین عوامل تشخیصی سرطان پستان استفاده گردید. مقدار P<۰/۰۵ به‌عنوان سطح آماره‌ی معنی‌دار برای تعیین عوامل تشخیصی استفاده شد و متغیرهایی که ارتباط آن‌ها با کلاس خروجی که همان نتیجه‌ی تشخیصی بود، در این سطح آماری معنی‌دار بود، به‌عنوان عوامل

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

مهم تشخیصی سرطان پستان در داده‌کاوی استفاده شدند.

- **پیاده‌سازی و ارزیابی مدل**

در روش داده‌کاوی به‌منظور پیاده‌سازی الگوریتم‌های منتخب تولید قوانین از چهار الگوریتم DS، J-48، RF، RT و XG-Boost در نرم‌افزار Weka 3.4 و زبان برنامه‌نویسی Java استفاده گردید. در پیاده‌سازی الگوریتم‌های منتخب از ۷۰ درصد مجموعه داده به‌منظور آموزش الگوریتم استفاده گردید. همچنین ۲۰ درصد و ۱۰ درصد از مجموعه داده به‌ترتیب برای آزمایش و اعتبارسنجی مدل استفاده شد(۲۳و۲۲).

به‌منظور مقایسه‌ی عملکرد الگوریتم‌های داده‌کاوی، مدل‌های تشخیصی ایجادشده بر اساس شاخص‌های عملکردی صحت، ویژگی و حساسیت و اندازه‌ی F حاصل از ماتریس آشفتگی (confusion matrix) مقایسه شدند. همچنین به‌منظور قابلیت دسته‌بندی الگوریتم‌های تولید قوانین در حالات آموزش، آزمایش و اعتبارسنجی، سطح زیرمنحنی خصوصیت گیرنده عامل مورد (AUC: Area under the ROC Curve) تحلیل و مقایسه گردیدند و نهایتاً مناسب‌ترین مدل تشخیصی سرطان پستان به‌دست آمده و قوانین تشخیص سرطان پستان از درخت منتخب استخراج گردید. در مواردی که خطای طبقه‌بندی در گره برگ بیشتر از ۲۰ درصد کل نمونه‌های طبقه‌بندی شده بود، قوانین با تایید متخصصان به‌منظور طراحی سیستم تصمیم‌یار استفاده گردید. بدین‌صورت که از ۱۰ متخصص زنان و زایمان در شهر آبادان با استفاده از طیف لاوشی با سه طیف«ضروری است، ضروری نیست اما مفید است و ضروری نیست» نظرسنجی شد و مواردی که ضروری بودند یا به‌عبارتی میزان روایی محتوایی آن‌ها بیشتر از ۰/۶۲ به‌دست آمد، استفاده شدند(۲٤).

- **پیاده‌سازی سیستم تصمیم‌یار تشخیص سرطان پستان**

پس از تعیین مهم‌ترین قوانین، واسط کاربری سیستم تصمیم‌یار تشخیص سرطان پستان در محیط نرم‌افزاری Visual Studio V2015، در قالب زبان برنامه‌نویسی C#و چارچوب Dot Net Framework V3.5.4 طراحی گردید.

یافته‌ها

پس از شناسایی و حذف داده‌های مرتبط با نمونه‌هایی که بیش از ۷۰ درصد مقادیر مفقود داشتند، ۳٤ رکورد مرتبط با نمونه زنان با تشخیص مثبت سرطان پستان و ۳٦ رکورد مرتبط با نمونه زنان با تشخیص منفی سرطان پستان از پژوهش حذف گردیدند. در نهایت به‌ترتیب ۲۲۱ و ۳۰٦ نمونه از زنان با تشخیص مثبت و منفی سرطان پستان در پژوهش باقی‌مانده و تحلیل شدند. نتایج حاصل از تحلیل عوامل تشخیصی سرطان پستان در جدول ۲ نشان داده شده‌اند.

جدول ۲: نتایج تحلیل اهمیت اثر عوامل تشخیصی سرطان پستان

شماره	نام متغیر	نوع متغیر	سطح آماره
۱	سابقه‌ی فردی سرطان پستان	کیفی دوحالتی	۰/۰۱۱
۲	سابقه‌ی رادیوگرافی از قفسه سینه	کیفی دوحالتی	۰/۰۳
۳	سابقه‌ی فشارخون	کیفی دوحالتی	۰/۰۱۵
٤	افزایش کلسترول خون LDL	کیفی دوحالتی	۰/۰۲۳
۵	وجود توده در ربع فوقانی داخلی سینه	کیفی دوحالتی	۰/۰۲۷
٦	سابقه‌ی هورمون درمانی با استروژن	کیفی دوحالتی	۰/۰٤۳
۷	سابقه‌ی هورمون درمانی با استروژن-پروژسترون	کیفی دوحالتی	۰/۰۱۲
۸	سابقه‌ی خانوادگی سرطان پستان	کیفی دوحالتی	۰/۰۰۱
۹	سابقه‌ی سرطان‌های دیگر	کیفی دوحالتی	۰/۰۱
۱۰	نسبت اندازه‌ی دور کمر به دور باسن	کمی پیوسته	۰/۰٤۵
۱۱	مصرف میوه	کمی پیوسته	۰/۰۱۲
۱۲	مصرف سبزی	کمی پیوسته	۰/۰۱۵
۱۳	سن	کمی پیوسته	۰/۰۱۲
۱٤	BMI	کمی پیوسته	۰/۱۸
۱۵	مصرف الکل	کمی پیوسته	۰/۳۵
۱٦	چاقی	کیفی دوحالتی	۰/۲۲

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

۱۷	سابقه‌ی پیاده‌روی	کیفی دوحالتی	۰/۶۵
۱۸	شغل سخت و زیان آور	کیفی دوحالتی	۰/۵۳
۱۹	میزان فعالیت فیزیکی در روز	کیفی چندحالتی	۰/۱۳
۲۰	افزایش تری‌گلیسیرید خون	کیفی دوحالتی	۰/۲۵
۲۱	افزایش کلسترول خون VLDL	کیفی دوحالتی	۰/۳۷
۲۲	سابقه‌ی دیابت	کیفی دوحالتی	۰/۱۸
۲۳	وجود کیست در پستان	کیفی دوحالتی	۰/۱۱

نتایج حاصل از جدول ۲ نشان داد که متغیرهای سابقه‌ی فردی سرطان پستان، سابقه‌ی نمونه‌برداری از سینه، سابقه‌ی رادیوگرافی از قفسه‌ی سینه، سابقه‌ی فشارخون، افزایش کلسترول خون LDL، وجود توده در ربع فوقانی داخلی سینه، هورمون درمانی با استروژن، هورمون درمانی با استروژن–پروژسترون، سابقه‌ی خانوادگی سرطان پستان، سن، سابقه‌ی سرطان‌های دیگر، نسبت اندازه‌ی کمر به لگن، مصرف میوه و سبزی ارتباط معناداری را با متغیر تشخیص سرطان پستان در سطح آماره $P<۰/۰۵$ به‌دست آوردند و به‌عنوان متغیرهای موثر در ایجاد مدل تشخیصی سرطان پستان استفاده شدند. همچنین BMI، مصرف الکل، چاقی، سابقه‌ی پیاده‌روی زیاد، شغل سخت، فعالیت فیزیکی، سابقه‌ی افزایش تری‌گلیسرید و کلسترول VLDL در خون، سابقه‌ی دیابت و وجود کیست در پستان با میزان $P<۰/۰۵$ از پژوهش حذف گردیدند. نتایج حاصل از مقایسه و ارزیابی عملکرد هر یک از مدل‌های تشخیصی با استفاده از ماتریس آشفتگی در جدول ۲ ارایه گردیده است.

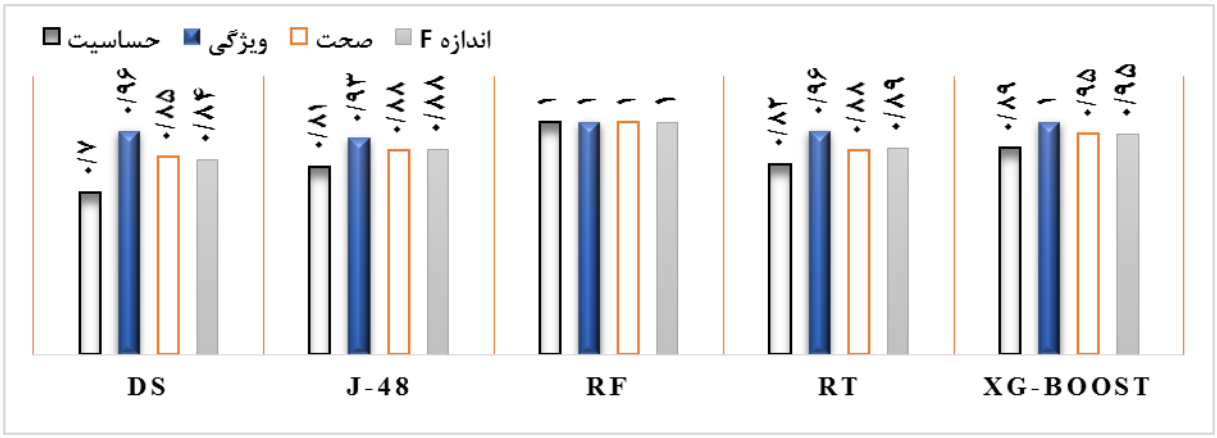
جدول ۳: نتایج حاصل از مقایسه‌ی نمونه‌های تشخیصی مثبت و منفی سرطان پستان با استفاده از مدل آشفتگی

شماره	نام الگوریتم	فرا پارامترهای منتخب	مثبت واقعی	منفی کاذب	مثبت کاذب	منفی واقعی
۱	DS	حداقل تعداد نمونه در هر برگ= ۱ حداقل اندازه والدین= ۲ معیار شکست= gdi قابلیت هرس= دارد	۱۴۸	۷۳	۱۴	۲۹۲
۲	J-48	ضریب اطمینان= ۰/۲۵ تعداد داده برای کاهش خطای هرس= ۳ حداقل تعداد نمونه در هر برگ= ۲ تعداد دانه= ۵	۱۷۹	۴۲	۲۲	۲۸۴
۳	RF	حداکثر عمق درخت= ۸ تعداد ویژگی‌های به‌صورت تصادفی انتخاب‌شده= ۱۰ حداکثر تعداد تکرارهای الگوریتم= ۲۰ تعداد مدل انتخاب‌شده برای ترکیب الگوریتم= ۶ درخت C5	۲۱۶	۵	۲	۳۰۴
۴	RT	حداقل وزن کلی نمونه‌ها در هر برگ= ۱ حداکثر عمق درخت= ۶ تعداد ویژگی انتخاب‌شده تصادفی= ۲ حداقل نسبت واریانس متغیرها برای تقسیم‌بندی داده= ۰/۰۰۱	۱۷۶	۴۵	۱۴	۲۹۲
۵	XG-Boost	کمینه وزن فرزندان= ۱ گاما= ۰/۵ حداکثر عمق= ۸ نرخ یادگیری= ۰/۱ آلفا= ۰ لامبدا= ۱ اتا= ۰/۱	۲۱۰	۱۱	۰	۳۰۶

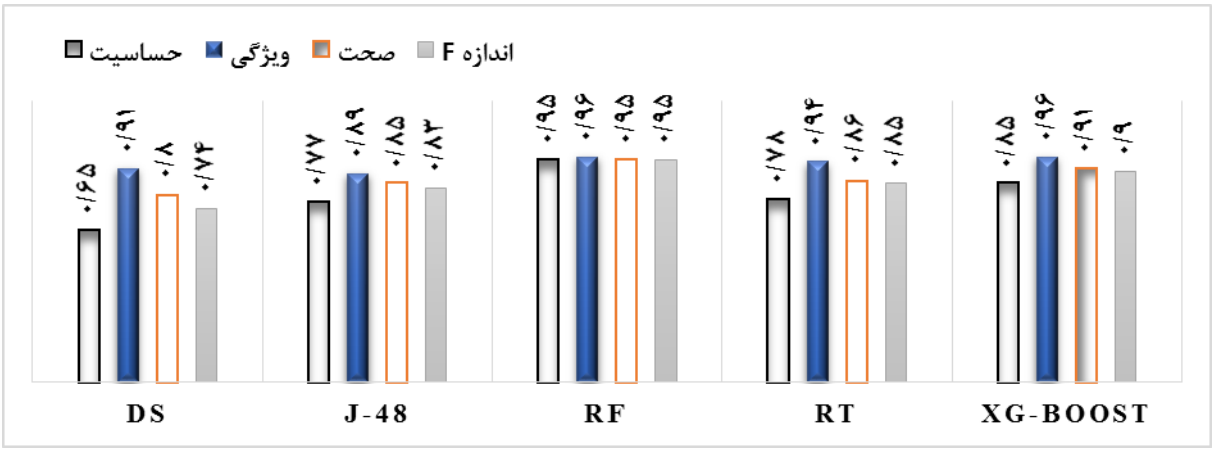
در جدول ۳، مثبت واقعی و منفی واقعی به‌ترتیب، بیانگر تعداد افراد با تشخیص مثبت و منفی سرطان پستان می‌باشند که توسط مدل به‌درستی با تشخیص مثبت طبقه‌بندی گردیدند. منفی کاذب و مثبت کاذب بیانگر تعداد افراد با تشخیص مثبت و منفی سرطان پستان می‌باشند که به نادرستی توسط مدل به‌عنوان منفی و مثبت تشخیص داده شده‌اند. نتایج حاصل از طبقه‌بندی نمونه‌های تشخیصی مثبت و منفی سرطان پستان با استفاده از جدول ۳ نشان داد که الگوریتم RF با میزان مثبت واقعی، منفی کاذب، منفی واقعی و مثبت کاذب به‌ترتیب برابر با ۲۱۶، ۵، ۲ و ۳۰۴ قابلیت دسته‌بندی بهتری را نسبت به

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

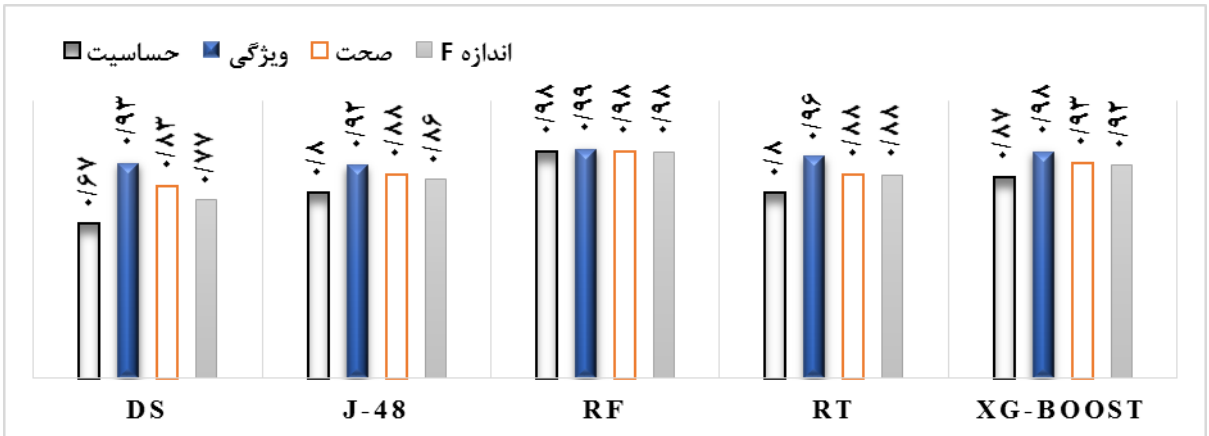
سایر الگوریتم‌های تولید قوانین در تشخیص سرطان پستان به‌دست آوردند. در ارتباط با طبقه‌بندی نمونه‌های با تشخیص منفی، الگوریتم XG-Boost با تعداد منفی واقعی ۳۰۶ عملکرد بالاتری نسبت به سایر الگوریتم‌ها داشته است. الگوریتم DS با میزان مثبت حقیقی برابر با ۱۴۸ و منفی کاذب برابر با ۷۳ پایین‌ترین عملکرد را نسبت به سایر الگوریتم در طبقه‌بندی نمونه‌های با تشخیص مثبت سرطان پستان داشته است. در رابطه با طبقه‌بندی نمونه‌های با تشخیص منفی سرطان پستان نیز الگوریتم J-48 با میزان منفی حقیقی برابر با ۲۸۴ و مثبت کاذب برابر با ۲۲ پایین‌ترین عملکرد را داشته است. نتایج حاصل از اندازه‌گیری و مقایسه‌ی شاخص‌های عملکردی اندازه F، صحت، حساسیت و ویژگی هر یک از الگوریتم‌ها در هریک از حالات آموزش، آزمایش و اعتبارسنجی مدل در نمودارهای ۳-۱ نشان داده شده است.



نمودار ۱: شاخص‌های عملکردی الگوریتم‌های منتخب در حالت آموزش



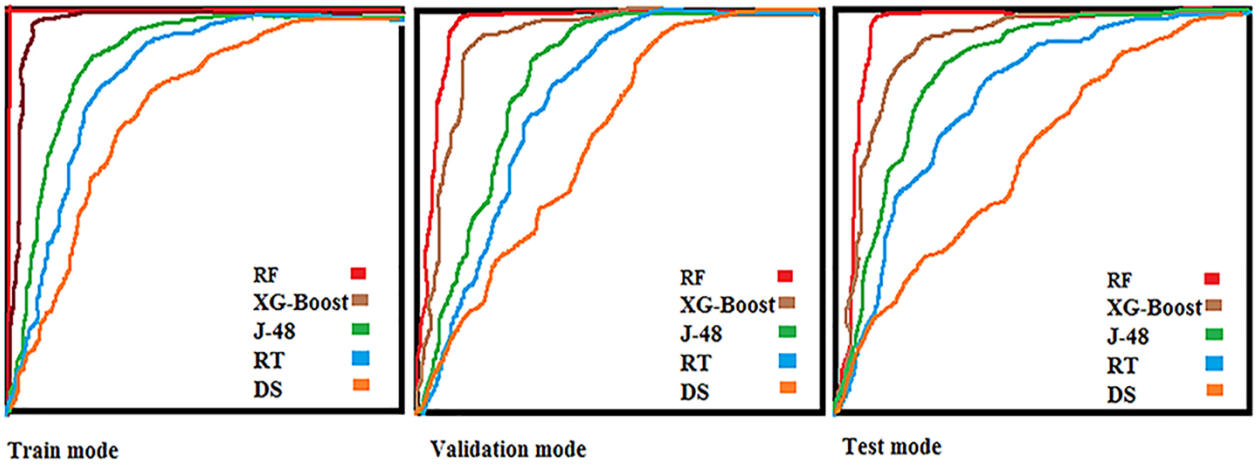
نمودار ۲: شاخص‌های عملکردی الگوریتم‌های منتخب در حالت آزمایش



نمودار ۳: شاخص‌های عملکردی الگوریتم‌های منتخب در حالت اعتبارسنجی

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

بر اساس نمودارهای ۳-۱، مدل RF با میزان حساسیت ۱ در حالت آموزش، حساسیت ۰/۹۵ در حالت آزمایش و ۰/۹۸ در حالت اعتبارسنجی، نتیجه بهتری را نسبت به سایر الگوریتم‌های منتخب در طبقه‌بندی نمونه‌های با تشخیص مثبت سرطان پستان در هریک از حالات آموزش، آزمایش و اعتبارسنجی داشته است. همچنین تمامی الگوریتم‌های منتخب قابلیت طبقه‌بندی خوبی در ارتباط با نمونه‌های با تشخیص منفی سرطان پستان داشتند؛ به‌عبارتی‌دیگر، ویژگی تمامی الگوریتم‌های منتخب به میزان بالاتر از ۰/۹ در هریک از حالات به‌دست آمده بود. در مجموع، نتایج حاصل از طبقه‌بندی نمونه‌های پژوهش با استفاده از الگوریتم‌های منتخب نشان داد که درخت تصمیم RF قابلیت عملکردی بالاتری در طبقه‌بندی نمونه‌های با تشخیص مثبت و منفی سرطان پستان در هریک حالات آموزش، آزمایش و اعتبارسنجی داشته است. سایر الگوریتم‌ها نیز از عملکرد نسبی خوبی برخوردار بودند. نتایج حاصل از AUC الگوریتم‌های طبقه‌بندی در هریک از حالات آموزش، آزمایش و اعتبارسنجی مدل تشخیصی در شکل ۱ نشان داده شده است(محور عمودی معرف حساسیت و محور افقی معرف مکمل ویژگی می‌باشند).



شکل ۱: سطح زیرمنحنی خصوصیت گیرنده عامل مدل‌های منتخب

بر اساس نتایج شکل ۱، منحنی خصوصیت گیرنده عامل الگوریتم RF نسبت به سایر الگوریتم‌ها به محور عمودی نزدیک‌تر بود و با میزان AUC_{-train}= ۰/۹۸۸، $AUC_{-validation}$= ۰/۹۵۶ و AUC_{-test}= ۰/۸۷۵ و AUC_{-total}= ۰/۹۳۶ عملکرد بهتری را در مقایسه با سایر الگوریتم‌های منتخب در تمامی مراحل آموزش، اعتبارسنجی و آزمایش الگوریتم در طبقه‌بندی از خود نشان داد؛ بنابراین، به‌عنوان مناسب‌ترین مدل تشخیصی سرطان پستان در نظر گرفته شد. درخت DS نیز با میزان AUC_{-train}= ۰/۷۹٤، $AUC_{-validation}$= ۰/۷٦٦ و AUC_{-test}= ۰/۷۱۲ و AUC_{-total}= ۰/۷۳۱ کارایی پایین‌تری را نسبت به سایر مدل‌های تشخیصی از خود نشان داد. در شکل ۲ ساختار درخت تصمیم RF نشان داده شده است.

```
of Breast Cancer = 0
 Fruit = 1 : 1 (8/0)
 Fruit = 2
 |   H of Breast Specimen = 1 : 1 (11/1)
 |   H of Breast Specimen = 2 : 0 (14/7)
 Fruit = 3 : 1 (11/1)
of Breast Cancer = 1
 Physical Activity = 1
 |   H of Breast Specimen = 1 : 1 (7/1)
 |   H of Breast Specimen = 2 : 0 (16/7)
 Physical Activity = 2
 |   H of Breast Specimen = 1
 |   |   WaistRatioPelvic < 95.83
 |   |   |   H of Chest Radiotherapy = 1 : 1 (9/2)
 |   |   |   H of Chest Radiotherapy = 2 : 0 (13/4)
 |   |   WaistRatioPelvic >= 95.83 : 1 (8/0)
```

شکل ۲: قسمتی از درخت RF تشخیص سرطان پستان

بر اساس درخت RF، درمجموع ٦٥ قانون تشخیص سرطان پستان استخراج شدند و در طراحی سیستم تصمیم‌یار استفاده شدند. در زیر به هفت نمونه از قوانین مستخرج تشخیص سرطان پستان از درخت RF اشاره شده است:

١- اگر سابقه خانوادگی سرطان پستان در فرد وجود نداشته باشد و مصرف میوه در فرد کمتر از ١٠٠ گرم در روز باشد، تشخیص مثبت است.

٢- اگر سابقه خانوادگی سرطان پستان در فرد وجود نداشته باشد و مصرف میوه در فرد بین ١٠٠ تا ٢٠٠ گرم در روز باشد و سابقه‌ی نمونه‌برداری از سینه نداشته باشد، تشخیص منفی است.

٣- اگر سابقه خانوادگی سرطان پستان در فرد وجود داشته باشد و مصرف میوه در فرد بین ١٠٠ تا ٢٠٠ گرم در روز باشد و سابقه‌ی نمونه‌برداری از سینه نداشته باشد، تشخیص منفی است.

٤- اگر سابقه خانوادگی سرطان پستان در فرد وجود داشته باشد و میزان فعالیت فیزیکی در فرد کمتر از نیم ساعت در روز باشد و سابقه‌ی نمونه‌برداری سینه در فرد وجود داشته باشد، تشخیص مثبت است.

٥- اگر سابقه خانوادگی سرطان پستان در فرد وجود داشته باشد و میزان فعالیت فیزیکی در فرد کمتر از نیم ساعت در روز باشد و سابقه‌ی نمونه‌برداری سینه در فرد وجود نداشته باشد، تشخیص منفی است.

٦- اگر سابقه خانوادگی سرطان پستان در فرد وجود داشته باشد و میزان فعالیت فیزیکی در فرد نیم تا یک ساعت در روز باشد و سابقه‌ی نمونه‌برداری سینه در فرد وجود داشته باشد و نسبت دور کمر به لگن در فرد کمتر از ٩٥/٨ باشد و سابقه‌ی رادیوتراپی از قفسه سینه در فرد وجود داشته باشد، تشخیص مثبت است.

٧- اگر سابقه خانوادگی سرطان پستان در فرد وجود داشته باشد و میزان فعالیت فیزیکی در فرد نیم تا یک ساعت در روز باشد و سابقه‌ی نمونه‌برداری سینه در فرد وجود داشته باشد و نسبت دور کمر به لگن در فرد کمتر از ٩٥/٨ باشد و سابقه‌ی رادیوتراپی از قفسه سینه در فرد وجود داشته باشد، تشخیص منفی است.

در قانون اول هشت نمونه در برگ طبقه‌بندی شده بود و تمامی این نمونه‌ها به‌درستی طبقه‌بندی شده بودند و معرف تشخیص مثبت سرطان پستان بودند، در قانون دوم ١١ نمونه به‌صورت کل طبقه‌بندی شده بودند که تنها یک نمونه به اشتباه طبقه‌بندی شده بود. در قانون سوم از ١٤ نمونه طبقه‌بندی‌شده، ٧ نمونه به‌درستی طبقه‌بندی شده بودند. اهمیت نسبی هریک از متغیرهای تشخیصی با استفاده از مدل RF در جدول ٤ نشان داده شده است.

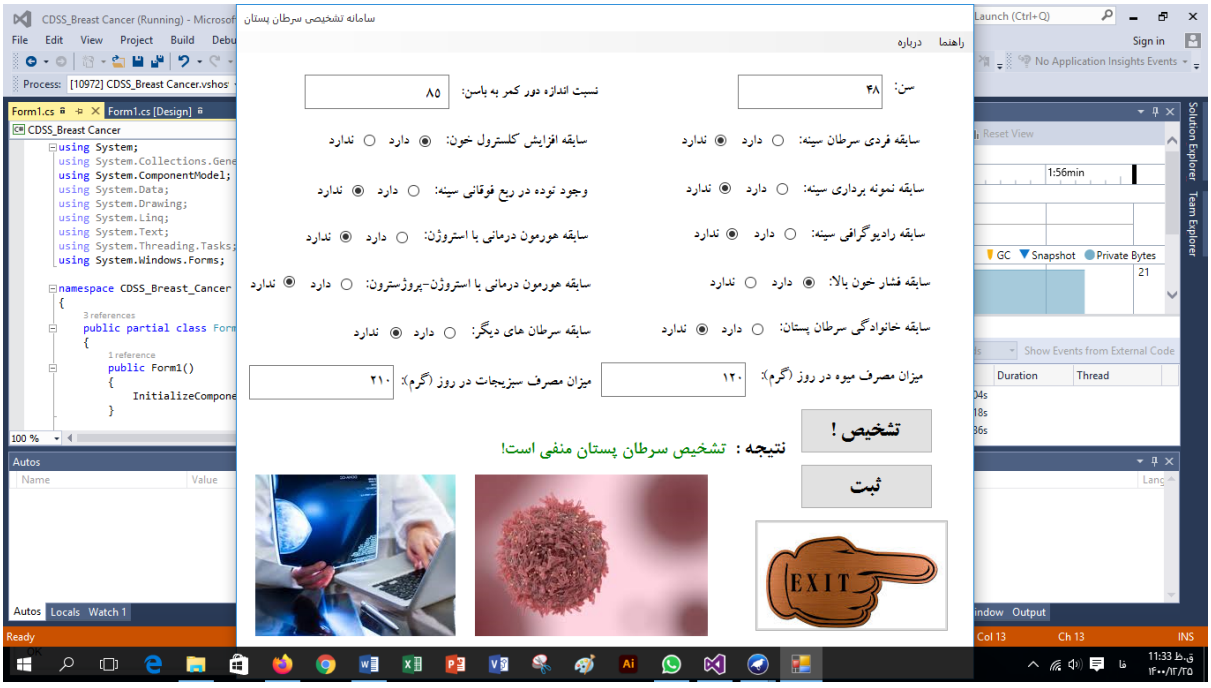
جدول ۴: اهمیت نسبی متغیرهای تشخیصی سرطان پستان با استفاده از مدل RF

نام متغیر	تعداد گره‌های استفاده‌شده برای متغیر(متوسط کاهش ناخالصی)	میزان اهمیت نسبی
سن	٥٣٦	٠/٦١
نسبت اندازه دور کمر به دور باسن	٢٠٨	٠/٣١
سابقه‌ی سرطان‌های دیگر	١٤٣	٠/٣
سابقه‌ی خانوادگی سرطان پستان	١٦٦	٠/٢٩
مصرف میوه	١٢٨	٠/٢٧
سابقه‌ی شخصی سرطان پستان	١٠٢	٠/٢٦
وجود توده در ربع فوقانی داخلی سینه	١٣	٠/٢٤
سابقه‌ی نمونه‌گیری از سینه	١٢٧	٠/٢٤
مصرف سبزی	١١٥	٠/٢١
سابقه‌ی رادیوتراپی از قفسه سینه	١١٢	٠/١٩
فعالیت فیزیکی	١٠٨	٠/١٩
سابقه‌ی فشارخون	٥٣	٠/١٨
افزایش تری‌گلیسیرید خون	١٦	٠/١٦

بر اساس جدول ٤، متغیر سن با میزان اهمیت ٠/٦١ و استفاده بیشتر از گره‌های مدل RF با n=٥٣٦ به‌عنوان مهم‌ترین متغیر تشخیصی سرطان پستان و افزایش تری‌گلیسیرید خون با میزان ٠/١٦ اهمیت نسبی و ١٦ گره استفاده شده در مدل، به‌عنوان کم‌اهمیت‌ترین متغیر توسط مدل RF در تشخیص سرطان پستان در نظر گرفته شد.

شکل ٣، سیستم تصمیم‌یار تشخیص سرطان پستان را در قالب زبان C# در محیط نرم‌افزاری Visual Studio V2015 نشان می‌دهد.

[DOR: 20.1001.1.17358132.1401.16.2.5.4]



شکل ۳: رابط کاربری سیستم تصمیم‌یار تشخیص سرطان پستان

بر اساس شکل ۳، پزشکان با واردکردن اطلاعات مرتبط با افرادی که مشکوک به سرطان پستان باشند، می‌توانند نتیجه تشخیصی سرطان پستان حاصل از سامانه را دریافت کنند و در نتیجه این ابزار می‌تواند به‌عنوان یک ابزار کمک تشخیصی برای تشخیص زودرس و پیش‌آگهی سرطان پستان در زنان استفاده شود.

بحث

امروزه کاربرد روش‌های یادگیری ماشین در حمایت از تشخیص‌های پزشکی برای غربالگری سرطان پستان به‌دلیل افزایش کارآمدی این روش‌ها به‌منظور شناسایی و طبقه‌بندی بیماری رو به افزایش است. اثبات شده که استفاده از سیستم‌های تصمیم‌یار بالینی مبتنی بر یادگیری ماشین موجب تشخیص از ابتلا به بیماری سرطان پستان و شناسایی سریع موارد برای بهبود پیامدهای درمانی می‌شود(۱۹و۱۴). در پژوهش حاضر ابتدا به شناسایی عوامل تشخیصی سرطان پستان با استفاده از روش کای دو پیرسون و تحلیل واریانس یک‌طرفه پرداخته شد. سپس به‌منظور ساخت مدل پیش‌بینی، مدل‌های داده‌کاوی مبتنی بر قوانین شامل RT، RF، DS و J-48 برای تشخیص سرطان پستان بر روی‌داده‌های تعداد ۵۹۷ فرد مشکوک به سرطان پستان(۲۵۵ بیمار مبتلا و ۳۴۲ فرد سالم) پیاده‌سازی و مقایسه گردید. ۲۴ متغیر اولیه برای توصیف مجموعه داده و استخراج داده مورد استفاده قرار گرفت. با توجه به حجیم‌بودن پایگاه داده‌ی حاضر و وجود متغیرهای کم‌اهمیت، تحلیل انتخاب ویژگی به‌عنوان یکی از پیش‌نیازهای پیشنهادی در بهینه‌سازی ورودی‌ها و ارتقای عملکرد دسته‌بندی استفاده شد. در مرحله‌ی نخست با استفاده از فن انتخاب ویژگی تاثیرگذارترین متغیرها شناسایی گردید. بنابراین به‌منظور مقایسه‌ی عملکرد روش‌های انتخاب ویژگی، مدل‌های منتخب داده‌کاوی بر روی متغیرهای منتخب حاصل از تحلیل انتخاب ویژگی پیاده‌سازی شدند. پس از تحلیل داده‌ها ۱۴ متغیر که عمدتاً از کلاس‌های سبک زندگی، سابقه‌ای و هورمونی–تولیدمثلی بودند، به‌عنوان مهم‌ترین عوامل پیش‌بینی‌کننده شناسایی شدند.

در زمینه‌ی کاربرد یادگیری ماشین کارهای تحقیقاتی متعددی برای تشخیص و غربالگری سرطان پستان انجام‌شده که در برخی از آن‌ها مقایسه‌ی الگوریتم‌های مختلف به‌منظور پیشنهاد کارآمدترین آن‌ها در دستور کار بوده است. این الزام با توجه به ماهیت تهاجمی و کشنده‌ی بدخیمی‌های سینه باهدف حمایت از تصمیمات بالینی و انتخاب بهترین مداخله‌ی درمانی حیاتی است. در مطالعه‌ی Lakshmi و همکاران از ترکیب قابلیت ۱۱ روش پایه داده‌کاوی برای تشخیص بیماری سرطان پستان استفاده شد. پس از پیاده‌سازی مدل‌ها بهترین عملکرد مربوط به الگوریتم RF با صحت طبقه‌بندی ۹۸/۸ درصد بود(۱۷). نتایج مطالعه‌ی Karatza و همکاران(۲۰۲۱) حاکی از عملکرد بهینه مدل RF با صحت تشخیصی ۹۶/۴۶ درصد برای طبقه‌بندی سرطان پستان بود(۲۵). به‌طور مشابه، Nanglia و همکاران(۲۰۲۲) نشان دادند از بین روش‌های منتخب داده‌کاوی، الگوریتم RF با صحت ۹۳/۹ درصد بهترین قابلیت را برای پیش‌بینی سرطان پستان داشت(۲۶).

مطالعه‌ی Yifan و همکاران نشان داد که الگوریتم RF یکپارچه‌شده با الگوریتم ada-boost با میزان صحت ۰/۹۸۵ عملکرد بهتری نسبت به سایر الگوریتم‌های داده‌کاوی در تشخیص سرطان پستان دارد(۲۷). Kober و همکاران پس از پیاده‌سازی مدل‌های داده‌کاوی مختلف به این نتیجه رسیدند که پیاده‌سازی الگوریتم RF در تشخیص سرطان پستان بالاترین امتیاز را کسب کرده است(۲۸). صحت طبقه‌بندی تشخیصی مدل RF به‌عنوان بهترین مدل تشخیصی برای طبقه‌بندی سرطان پستان(مثبت یا منفی) در مطالعه‌ی Jain و همکاران ۹۶/۵ درصد بود(۲۹). در مطالعه‌ی دیگر توسط Omondiagbe و همکاران پس از مقایسه الگوریتم‌های یادگیری ماشین متعدد برای طبقه‌بندی تشخیصی سرطان پستان نشان دادند که الگوریتم RF با میزان صحت، حساسیت، اختصاصیت و سطح زیرنمودار خصوصیت گیرنده‌ی عامل به‌ترتیب ۸۲/۵، ۹۸، ۱۰۰ و ۹۸/۹ درصد بالاترین عملکرد را داشت(۱۵). نتایج مطالعات مروری انجام‌شده توسط Nindrea و همکاران، Li و همکاران و Yassin و همکاران نشان داد که الگوریتم RF یکی از کارآمدترین و پرکاربردترین روش‌ها برای طبقه‌بندی تشخیصی بدخیمی‌های سینه معرفی‌شده است(۱۶و۱۴و۱۲).

در مطالعه‌ی حاضر پس از پیاده‌سازی الگوریتم‌های مختلف، مدل RF با میزان ویژگی، صحت، دقت و حساسیت به‌ترتیب برابر با ۰/۹۷، ۰/۹۹، ۰/۹۸ و ۰/۹۷۴ و سطح زیرنمودار ۰/۹۳۶ عملکرد بالاتری نسبت به سایر الگوریتم‌های منتخب داشته است و به‌عنوان مدل تشخیصی سرطان پستان در نظر گرفته شد.

تاکنون مطالعاتی در زمینه‌ی شناسایی متغیرهای تاثیرگذار غیربالینی در تشخیص سرطان پستان به‌عنوان یکی از پیش‌نیازهای مهم در پیاده‌سازی فناوری‌های CDSS انجام پذیرفته است. در این مطالعات از فنون مبتنی بر مدل‌های یادگیری ماشین برای تشخیص سرطان پستان و شناسایی موارد سرطانی از افراد سالم استفاده‌شده است. متغیرهای مورداستفاده در این پژوهش‌ها(۲۵و۱۹و۱۷) به‌عنوان ورودی مدل‌های داده‌کاوی معمولاً از رده‌های سبک زندگی، سابقه‌ای، اجتماعی-اقتصادی، دموگرافیکی و هورمونی-تولیدمثلی بودند. این متغیرها عبارتند از: ۱- متغیرهای جمعیت شناختی و اجتماعی-اقتصادی: سن، سطح سواد، میزان درآمد، شغل، وضعیت تأهل، شاخص توده‌بدنی و نسبت دورکمر، ۲- داده‌های سبک زندگی از جمله رژیم غذایی، فعالیت بدنی، مصرف نمک، مصرف غذاهای پر فیبر، دخانیات و الکل، ۳- متغیرهای سابقه‌ای/ عوامل خطر از جمله بیماری زمینه‌ای فشارخون، دیابت، عوامل ژنتیکی، سابقه خانوادگی سرطان پستان، سابقه فردی کیست پستان، عوامل شغلی و محیطی، در معرض سموم و آلاینده‌های شیمیایی و ٤- عوامل خطر تولیدمثلی-هورمونی شامل: سن ازدواج، سن قاعدگی، فرزندآوری، شیردهی و مدت‌زمان آن، وضعیت حاملگی، سن اولین حاملگی، وضعیت قاعدگی، سن قاعدگی، نظم در چرخه قاعدگی، طول چرخه قاعدگی، هورمون درمانی، مصرف قرص خوراکی ضدحاملگی و مدت‌زمان مصرف آن. در پژوهش حاضر پس از انجام تحلیل انتخاب ویژگی متغیرهای عمدتاً سبک زندگی، سابقه‌ای و ویژگی‌های هورمونی-تولیدمثلی مانند متغیرهای سابقه فردی سرطان پستان، سابقه‌ی نمونه‌برداری از سینه، سابقه‌ی رادیوگرافی از قفسه سینه، سابقه فشارخون، افزایش کلسترول خون، وجود توده در ربع فوقانی سینه، هورمون درمانی با استروژن، هورمون درمانی با استروژن-پروژسترون، سابقه خانوادگی سرطان پستان، سابقه‌ی سرطان‌های دیگر، نسبت اندازه کمر به لگن و مصرف میوه و سبزی به‌عنوان موثرترین متغیرها در تشخیص سرطان پستان شناسایی شد. این درحالی است که در بیشتر پژوهش‌ها(۳۲-۳۰) از متغیرهای بالینی(مانند اندازه، موقعیت و میزان تمایز و مرفولوژی تومور)، پاراکلینیکی و تصویربرداری(مانند نتایج آزمایشگاهی و پردازش تصویر) برای تغذیه مدل‌های داده‌کاوی تشخیص سرطان پستان استفاده شده است.

مدل پیشنهادی حاضر قادر خواهد بود تا بدخیمی‌های سینه را در مراحل اولیه و از طریق متغیرهای غالباً سبک زندگی و سوابق پزشکی با میزان صحت بهینه تشخیص دهد. احتمالاً راه‌اندازی CDSS مبتنی بر این مدل به‌صورت موثر و کاربردی برای محیط‌های بالینی قابل‌استفاده باشد. باوجود این دارای محدودیت‌ها و چالش‌هایی است که نیازمند توجه است. در این پژوهش با توجه به ماهیت گذشته‌نگر بودن پایگاه داده، وجود فیلدهای اطلاعاتی نویزی(مانند موارد غیریکپارچه، ناقص، غیرطبیعی، بی‌معنی و دارای خطا) و خالی اجتناب‌ناپذیر بود. برای این منظور فیلدهای نویزی از طریق مراجعه به پرونده بالینی افراد و پرسش از پزشک معالج برطرف شد. برای برطرف‌شدن فیلدهای خالی از روش‌های series mean برای جایگزینی داده‌های کمی و تعیین نسبت فراوانی برای داده‌های کیفی استفاده شد. تک‌مرکزی بودن، کوچک‌بودن پایگاه داده، عدم‌استفاده از متغیرهای بالینی و رادیولوژیکی، نپرداختن به اعتبارسنجی بیرونی برای ارزیابی مدل‌ها و استفاده از تنها چند روش داده‌کاوی از دیگر محدودیت‌های پژوهش حاضر است؛ بنابراین پیشنهاد می‌گردد که به‌منظور بهبود کیفیت مدل‌سازی و کاهش پیش‌داوری در تشخیص، پژوهش‌های بیشتری پس از انجام اعتبارسنجی‌های

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

بیرونی دقیق‌تر بر روی پایگاه‌های داده بزرگ‌تر، چندمرکزی و آینده‌نگر انجام پذیرد. در پایان به‌نظر می‌رسد که استفاده از متغیرهای بالینی و پاراکلینیکی در کنار موارد سبک زندگی در بهبود صحت مدل‌سازی تاثیرگذار باشند.

نتیجه‌گیری

شناسایی سریع بدخیمی‌های سینه از طریق روش‌های فناورانه و غیرتهاجمی موجب بهبود کیفیت خدمات درمانی و ارایه درمان‌های سفارشی می‌شود. طراحی CDSS مبتنی بر کارآمدترین روش‌های یادگیری ماشین باهدف تشخیص بدخیمی‌های سینه با استفاده از عوامل سبک زندگی، سابقه‌ای و ویژگی‌های هورمونی-تولیدمثلی زمینه‌ساز کاهش عوارض وخیم بیماری و افزایش شانس بقای بیماران خواهد شد. ازاین‌رو پژوهش حاضر به‌صورت گذشته‌نگر و از طریق شناسایی متغیرهای مهم در تشخیص سرطان پستان سعی در بهبود عملکرد مدل‌های داده‌کاوی داشت. سپس عملکرد الگوریتم‌های منتخب داده‌کاوی بر روی متغیرهای منتخب افراد مبتلا به سرطان پستان و افراد سالم مدل‌سازی و مقایسه شدند؛ بنابراین در پژوهش حاضر پس از پیاده‌سازی مدل‌های داده‌کاوی، الگوریتم RF به‌عنوان بهترین مدل معرفی شد. استفاده از این الگوریتم در شناسایی موثر بیماران موجب حفظ هزینه اثربخشی درمان، اولویت‌بندی صحیح منابع و بهبود ایمنی و کیفیت مراقبتی خواهد شد. باوجود این نیاز به مطالعات بیشتر برای پیاده‌سازی مدل پیشنهادی در قالب فناوری CDSS در محیط واقعی بالینی و به‌صورت یکپارچه با سایر سیستم‌های تحلیل خطر می‌باشد.

تشکر و قدردانی

مقاله‌ی حاضر برگرفته از یک طرح پژوهشی مصوب در دانشگاه علوم پزشکی ایلام با کد IR.MEDILAM.REC.1399.0220 می‌باشد. از مسئولان معاونت تحقیقات این دانشگاه و همچنین بیمارستان آیت‌الله طالقانی آبادان که در انجام این پژوهش گروه تحقیق را یاری رساندند، تشکر و قدردانی می‌گردد.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians 2021; 71(3): 209-49.

2. Lei S, Zheng R, Zhang S, Chen R, Wang S, Sun K, et al. Breast cancer incidence and mortality in women in China: Temporal trends and projections to 2030. Cancer Biology and Medicine 2021; 18(3): 900-9.

3. Yildirim NK, Ozkan M, Ilgun AS, Sarsenov D, Alco G, Aktepe F, et al. Possible role of stress, coping strategies, and life style in the development of Breast cancer. The International Journal of Psychiatry in Medicine 2018; 53(3): 207-20.

4. Salod Z & Singh Y. Comparison of the performance of machine learning algorithms in Breast cancer screening and detection: A protocol. Journal of Public Health Research 2019; 8(3): 1677.

5. Douangnoulack P & Boonjing V. Building minimal classification rules for Breast cancer diagnosis, Chiang Mai, Thailand: 10th International Conference on Knowledge and Smart Technology (KST), 2018.

6. Khandezamin Z, Naderan M & Rashti MJ. Detection and classification of Breast cancer using logistic regression feature selection and GMDH classifier. Journal of Biomedical Informatics 2020; 111(1): 103591.

7. Trimboli RM, Codari M, Guazzi M & Sardanelli F. Screening mammography beyond Breast cancer: Breast arterial calcifications as a sex-specific biomarker of cardiovascular risk. European Journal of Radiology 2019; 119(1): 108636.

8. Weller M, Sarmento G, Waleska Barros A, De Macedo Andrade AC, Dantas Guimaraes B, Ferreira Junior CA, et al. Breast cancer risk perception and mammography screening behavior of women in northeast Brazil. Women's Health Reports (New Rochelle, NY) 2020; 1(1): 150-8.

9. Lee JS & Oh M. Breast cancer screening in asian women with dense Breast by mammography: A cross-sectional observational study. Asian Pacific Journal of Cancer Prevention: APJCP 2021; 22(4): 1165-70.

10. Wu AM, Morse AR, Seiple WH, Talwar N, Hansen SO, Lee PP, et al. Reduced mammography screening for Breast cancer among women with visual impairment. Ophthalmology 2021; 128(2): 317-23.

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

11. Casal Guisande M, Comesana Campos A, Dutra I, Cerqueiro Pequeno J & Bouza Rodriguez JB. Design and development of an intelligent clinical decision support system applied to the evaluation of Breast cancer risk. Journal of Personalized Medicine 2022; 12(2): 169.

12. Nindrea RD, Aryandono T, Lazuardi L & Dwiprahasto I. Diagnostic accuracy of different machine learning algorithms for Breast cancer risk calculation: A meta-analysis. Asian Pacific Journal of Cancer Prevention: APJCP 2018; 19(7): 1747-52.

13. Jiang X, Wells A, Brufsky A & Neapolitan R. A clinical decision support system learned from data to personalize treatment recommendations towards preventing Breast cancer metastasis. PLoS One 2019; 14(3): e0213292.

14. Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. Predicting Breast cancer 5-year survival using machine learning: A systematic review. PloS One 2021; 16(4): e0250370.

15. Omondiagbe DA, Veeramani S & Sidhu AS. Machine learning classification techniques for Breast cancer diagnosis. Available at: https://iopscience.iop.org/article/10.1088/1757-899X/495/1/012033/pdf. 2019.

16. Yassin NI, Omran S, El Houby EM & Allam H. Machine learning techniques for Breast cancer computer-aided diagnosis using different image modalities: A systematic review. Computer Methods and Programs in Biomedicine 2018; 156(1): 25-45.

17. Lakshmi D, Gurrela SR & Kuncharam M. A comparative study on Breast cancer tissues using conventional and modern machine learning models. Switzerland: Springer; 2021: 693-9.

18. Gonzalez F. Using clinical decision support systems in Breast cancer treatment: A critical review. Cancer Nursing Practice 2022; 21(4): e1804.

19. Massafra R, Latorre A, Fanizzi A, Bellotti R, Didonna V, Giotta F, et al. A clinical decision support system for predicting invasive Breast cancer recurrence: Preliminary results. Frontiers in Oncology 2021; 11(1): 576007.

20. Mazo C, Kearns C, Mooney C & Gallagher WM. Clinical decision support systems in Breast cancer: A systematic review. Cancers 2020; 12(2): 369.

21. Xu F, Sepulveda MJ, Jiang Z, Wang H, Li J, Liu Z, et al. Effect of an artificial intelligence clinical decision support system on treatment decisions for complex Breast cancer. JCO Clinical Cancer Informatics 2020; 4(1): 824-38.

22. Kotthoff L, Thornton C, Hoos HH, Hutter F & Leyton Brown K. Auto-WEKA: Automatic model selection and hyperparameter optimization in WEKA. Available at: https://library.oapen.org/bitstream/handle/20.500.12657/23012/1007149.pdf#page=89. 2019.

23. Crowther PS & Cox RJ. A method for optimal division of data sets for use in neural networks, Berlin: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems, 2005.

24. Lawshe CH. A quantitative approach to content validity. Personnel Psychology 1975; 28(4): 563-75.

25. Karatza P, Dalakleidi K, Athanasiou M & Nikita KS. Interpretability methods of machine learning algorithms with applications in Breast cancer diagnosis, Mexico: 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2021.

26. Nanglia S, Ahmad M, Ali Khan F & Jhanjhi NZ. An enhanced Predictive heterogeneous ensemble model for Breast cancer prediction. Biomedical Signal Processing and Control 2022; 72(1): 103279.

27. Yifan D, Jialin L & Boxi F. Forecast model of Breast cancer diagnosis based on RF-AdaBoost, Beijing, China: International Conference on Communications, Information System and Computer Engineering (CISCE), 2021.

28. Kober KM, Roy R, Dhruva A, Conley YP, Chan RJ, Cooper B, et al. Prediction of evening fatigue severity in outpatients receiving chemotherapy: Less may be more. Fatigue: Biomedicine, Health and Behavior 2021; 9(1): 14-32.

29. Jain P, Patel D, Verma JP & Tanwar S. Computer-aided-diagnosis system for symptom detection of Breast and cervical cancer. Singapore: Springer; 2021: 743-58.

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

30. Boeri C, Chiappa C, Galli F, De Berardinis V, Bardelli L, Carcano G, at al. Machine learning techniques in Breast cancer prognosis prediction: A primary evaluation. Cancer Medicine 2020; 9(9): 3234-43.

31. Yue W, Wang Z, Chen H, Payne AM & Liu X. Machine learning with applications in Breast cancer diagnosis and prognosis. Designs 2018; 2(2): 1-17.

32. Jebarani PE, Umadevi N, Dang H & Pomplun M. A novel hybrid k-means and gmm machine learning model for Breast cancer detection. IEEE Access 2021; 9(1): 146153-62.

[DOR: 20.1001.1.17358132.1401.16.2.5.4]

Journal of Payavard Salamat
Vol. 16, No. 2, Jun 2022

Design of Clinical Decision Support System to Diagnose Breast Cancer: An Approach Using Data Mining

Mostafa Shanbehzadeh[1] (Ph.D.), Hadi Kazemi-Arpanahi[2] (Ph.D.), Raoof Nopour[3*] (M.S.)

1 Assistant Professor, Department of Health Information Technology, School of Allied Medical Sciences, Ilam University of Medical Sciences, Ilam, Iran
2 Assistant Professor, Department of Health Information Technology, School of Health Management and Information Sciences, Abadan University of Medical Sciences, Abadan, Iran
3 Ph.D. Candidate in Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran

Abstract

Received: 17 Mar. 2022
Accepted: 20 Jul. 2022

Background and Aim: Breast cancer is one of the most common and aggressive malignancies in women. Timely diagnosis of breast cancer plays an important role in preventing the progression of this disease, timely treatment measures, and aftermath reducing the mortality rate of these patients. Machine learning has the potential ability to diagnose diseases quickly and cost-effectively. This study aims to design a CDSS based on the rules extracted from the decision tree algorithm with the best performance to diagnose breast cancer in a timely and effective manner.

Materials and Methods: The data of 597 suspected people with breast cancer (255 patients and 342 healthy people) were retrospectively extracted from the electronic database of Ayatollah Taleghani Hospital in Abadan city with 24 characteristics, mainly pertained to lifestyle and medical histories. After selecting the most important variables by using the Chi-square Pearson and one-way analysis of variance ($P<0.05$), the performance of selected data mining algorithms including RF, J-48, DS, RT and XG -Boost was evaluated for breast cancer diagnosis in Weka 3.4 software. Finally, the breast cancer diagnostic system was designed based on the best model and through C# programming language and Dot Net Framework V3.5.4.

Results: Fourteen variables including personal history of breast cancer, breast sampling, and chest X-ray, high blood pressure, increased LDL blood cholesterol, presence of mass in upper inner quadrant of the breast, hormone therapy with estrogen, hormone therapy with Estrogen-progesterone, family history of breast cancer, age, history of other cancers, waist-to-hip ratio and fruit and vegetable consumption showed a significant relationship with the output class at the $P<0.05$. Based on the results of the performance evaluation of selected algorithms, the RF model with sensitivity, specificity, accuracy, and F- measure equal to 0.97, 0.99, 0.98, 0.974, respectively, AUC=0.936 had higher performance than other selected algorithms and was suggested as the best model for breast cancer diagnosis.

Conclusion: It seems that using modifiable variables such as lifestyle and reproductive-hormonal characteristics as input to the RF algorithm to design the CDSS, can detect breast cancer cases with optimal accuracy. In addition, the proposed system can be effectively adapted in real clinical environments for quick and effective disease diagnosis.

Keywords: Breast Cancer, Data Mining, Diagnostic Model, Clinical Decision Support System

*** Corresponding Author:**
Nopour R
Email:
nopoor.r@iums.ac.ir

[DOR: 20.1001.1.17358132.1401.16.2.5.4]