مجله دانش و تندرستی در علوم پایه پزشکی، دوره ۱۵، شماره ۳، پاییز ۱۳۹۹
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شاهرود

DOI: 10.22100/jkh.v15i3.2454

مقاله پژوهشی

ارایه یک سیستم کمک تصمیم‌یار بالینی جهت تشخیص سرطان پستان

مصطفی برومندزاده[۱]، الهام پروین‌نیا[*۱]

۱- گروه مهندسی کامپیوتر- واحد شیراز- دانشگاه آزاد اسلامی- شیراز- ایران.

تاریخ دریافت: ۱۳۹۹/۰۷/۱۴، **تاریخ پذیرش:** ۱۳۹۹/۰۹/۱۷

چکیده

مقدمه: *سرطان پستان از مهمترین عوامل مرگ و میر زنان است. بنابراین دقت و سرعت تشخیص بیماری در تعیین روال درمان، بسیار حیاتی است؛ در این راستا برای یکسان‌سازی گزارش‌های ماموگرافی از سیستم طبقه‌بندی BI-RADS استفاده شده است. با این وجود تفاوت نظر پزشکان در مورد مقادیر BI-RADS زیاد است. هدف این مقاله تشخیص BI-RADS با پردازش زبان طبیعی گزارش‌های ماموگرافی و اطلاعات کلینیکی حاصل از سوابق پرونده الکترونیک بیمار و ترکیب آن‌ها برای تعیین زیر گروه‌های مولکولی و کمک به روند تشخیصی بیماری و پیگیری درمان می‌باشد.*

مواد و روش‌ها: *در این مطالعه از ۱۲۰۰ گزارش ماموگرافی و اطلاعات سوابق پرونده الکترونیک مرکز آموزشی درمانی نمازی بین سال‌های ۱۳۹۶-۱۳۹۴ استفاده شد. با پردازش گزارش‌ها، ۱۶۰ ویژگی متناسب با آنها ایجاد و با مراجعه به سوابق پرونده الکترونیک افراد، ۱۸ ویژگی استخراج شد. از مجموعه بردارها با ۱۷۸ ویژگی، مقادیر BI-RADS با استفاده از ماشین بردار پشتیبان (SVM) و زیرگروه‌های مولکولی توسط بیزین‌ساده پیش‌بینی گردید و مورد ارزیابی قرار گرفت.*

نتایج: *برای ارزیابی نتایج، مقادیر دقت، ارزش اخباری مثبت، ارزش اخباری منفی، حساسیت و خاصیت، برای تشخیص BI-RADS و زیرگروه‌های مولکولی محاسبه شدند. میزان دقت برای تشخیص BI-RADS، ۸۵/۴۲٪ و برای تشخیص زیرگروه‌های مولکولی ۷۲/۳۱٪ به‌دست آمد.*

نتیجه‌گیری: *سیستم تصمیم‌یار ارایه‌شده، مدلی مناسب برای کمک به پزشک جهت تشخیص سرطان پستان و دسته‌بندی بیماران بود. همچنین مشخص گردید که ترکیب اطلاعات، شامل سوابق پرونده الکترونیک بیماران و زیرگروه‌های مولکولی تعیین‌شده در کنار گزارش‌های ماموگرافی می‌تواند در تشخیص بیماری و تعیین بهینه روال درمان مفید باشد.*

واژه‌های کلیدی: کمک تصمیم‌یار، زیرگروه‌های مولکولی، اطلاعات بالینی، سرطان پستان.

[*] **نویسنده مسئول:** شیراز، شهر صدرا، پردیس دانشگاه آزاد اسلامی واحد شیراز، دانشکده مهندسی۱، گروه کامپیوتر، **تلفن:** ۰۷۱۳۶۴۱۰۰۴۱، **نمابر:** ۰۷۱۳۶۴۱۰۰۵۹، **Email:** parvinnia@iaushiraz.ac.ir

ارجاع: برومندزاده مصطفی، پروین‌نیا الهام. ارایه یک سیستم کمک تصمیم‌یار بالینی جهت تشخیص سرطان پستان. مجله دانش و تندرستی در علوم پایه پزشکی ۱۳۹۹؛۱۵(۳):۵۴-۶۶.

مقدمه

سرطان یکی از عوامل اصلی بروز مرگ و میر در جهان امروز است. سرطان بعد از بیماری‌های قلبی عروقی دومین عامل شایع مرگ و میر در کشورهای توسعه‌یافته و سومین عامل مرگ در کشورهای کمتر توسعه‌یافته است و به تنهایی بیش از بیماری‌های سل، ایدز و مالاریا افراد را به کام مرگ می‌کشاند (۱). به‌طوری که اگر اقدامات پیشگیرانه انجام نشود در ۱۰ سال آینده شاهد مرگ و میر بیش از ۸۵ میلیون نفر در جهان خواهیم بود (۲). در حال حاضر سرطان عامل ۱۲٪ مرگ و میرها در سراسر جهان است (۳). در ایران سرطان سومین عامل مرگ و میر است و سالانه بیش از ۳۰۰۰۰ نفر از مردم در اثر سرطان جان خود را از دست می‌دهند. یکی از موارد شایع سرطان در زنان سرطان پستان است. براساس آمارهای ارایه‌شده، ۱۹/۹٪ مرگ‌ومیر ناشی از بیماری سرطان در زنان، مربوطبه سرطان پستان است (۴). طبق آمارهای منتشر شده توسط سازمان جهانی بهداشت، از هر ۸ تا ۱۰ زن، یک نفر به سرطان پستان مبتلا می‌شود. این آمار برای کشور ایران، احتمال ابتلای یک نفر به ازای هر ۱۰ تا ۱۵ زن را نشان می‌دهد (۵). اما نکته مهم در بحث تشخیص سرطان پستان تفاوت میانگین سن تشخیص در کشورهای غربی و ایران است به‌طوری‌که میانگین سن تشخیص در کشورهای غربی ۵۶ و در ایران ۴۵ سال است (۶). ازطرفی تشخیص زودهنگام در مراحل اولیه، از عوامل مهم و اساسی در درمان این بیماری است زیرا هنگامی‌که سرطان پستان زود تشخیص داده شود احتمال درمان و زنده ماندن بسیار زیاد است (۹–۷). بنابراین سیستم‌های پشتیبان تصمیم پزشکی (CDSS: clinical decision support system)، که نتیجه همکاری متقابل پزشکان و مهندسین هستند و در واقع برای کمک و پشتیبانی کارکنان مراقبت‌های بهداشتی در تصمیم‌گیری‌های بالینی ساخته می‌شوند، در این زمینه نقش مهمی را ایفا می‌نمایند (۱۲–۱۰). امروزه استفاده از CDSS در مراقبت از سرطان پستان در مراکز مراقبت‌های بهداشتی به شدت در حال افزایش است (۱۱). نتایج تحقیقات حاکی از آن است که به‌واسطه سیستم‌های تصمیم‌یار ایجاد شده و با کمک تجسم داده‌های بیمار، پزشکان اجازه یافته‌اند تا به سرعت به اطلاعات لازم جهت تعین درمان مناسب دسترسی پیدا کنند (۱۳). یکی از مواردی که می‌تواند به‌عنوان ورودی در یک سیستم CDSS برای کمک به تشخیص و درمان سرطان سینه به‌کار گرفته شود، گزارشات مربوط به ماموگرافی است (۱۴). رادیولوژیست‌ها بر اساس فاکتورهای مشاهده شده در ماموگرافی و بر اساس تشخیص، از یک سیستم طبقه‌بندی به نام BI-RADS (Breast Imaging-Reporting and Data System) (ایجادشده توسط کالج رادیولوژی آمریکا) برای توصیف نتایج ماموگرافی در گزارش‌های پزشکی استفاده می‌کنند (۱۵). طبقه‌بندی BIRADS در ماموگرافی، یک روش مهم و قابل اعتماد برای ارزیابی و تخمین ریسک بدخیمی در ضایعات پستانی به‌شمار می‌آید (۱۶). BI-RADS در واقع یک برچسب است که در گزارشات ماموگرافی، در ۷ سطح، بین ۰ تا ۶ تعریف می گردد و هر یک از این اعداد دارای تفسیری مشخص هستند (۱۷). از طرفی با توجه به مطالعات اولیه انجام شده در این حوزه مشخص گردید که تا به حال هیچگونه سیستم تصمیم‌یار بالینی برای تشخیص سرطان پستان و طبقه‌بندی بیماران بر اساس ترکیب اطلاعات گزارش‌های ماموگرافی، سوابق پرونده الکترونیک بیمار (در اینجا HIS: Hospital information system) و زیرگروه‌های مولکولی ارایه نشده است اما مقالات متعددی در این حوزه با رویکرد مهندسی نوشته شده است (۲۴–۱۸) که در ادامه به برخی مقالات این حوزه با رویکرد طراحی مدل پرداخته می‌شود.

گائو و همکاران در سال ۲۰۱۵ (۱۹) از Natural language processing (NLP) برای استخراج اطلاعات متون بدون‌ساختار ماموگرافی استفاده کردند روش آنها محدود به تشخیص چهارنوع عارضه پستان بود و صرفاً از گزارش‌های پزشکی استفاده شد. اسماعیلی و همکاران در سال ۲۰۲۰ سیستم تصمیم پشتیبان تصمیم جهت کمک به پزشکان در تفسیر گزارش‌های متن ماموگرافی را ضمن ایجاد مدلی با قابلیت پیش‌بینی نیازمندی بیمار به بیوپسی با دقت ۸۴/۰۶٪ ارایه کردند (۱۴). ژانگ و همکاران در سال ۲۰۱۹ (۲۴) از یادگیری عمیق برای استخراج اطلاعات بالینی سرطان پستان استفاده کردند. مقدار F1 بالغ بر ۹۳٪، اما پیچیدگی بالا بود. کاسترو و همکاران در سال ۲۰۱۷ (۱۸) روشی از NLP بر پایه rule برای طبقه‌بندی گزارش‌های رادیولوژی ارایه نمودند. مقدار F1، ۹۵٪ بود ولی تنها از یک نوع داده متنی استفاده شد. گوپتا و همکاران در سال ۲۰۱۷ (۲۰) روشی مبتنی‌بر درخت تجزیه و معناشناسی برای تولید اطلاعات ساخت‌یافته از گزارش‌های ماموگرافی ارایه نمودند. مقدار F1 یا میانگین هارمونیک، ۹۴٪ به‌دست آمد و صرفاً از گزارش‌های پزشکی استفاده شد. ساپیو و همکاران در سال ۲۰۱۳ (۲۳) توسط NLP و با ایجاد الگوریتم BROK (BI-RADS Observation Kit) به استخراج خودکار سطوح BI-RADS از گزارش‌های ماموگرافی پرداختند.

پرچا و همکاران در سال ۲۰۱۲ (۲۲) با پردازش گزارش‌ها آنها را به یک کلاس BI-RADS اختصاص دادند؛ دقت ۹۹٪ گزارش شد اما تمرکز صرفاً بر بافت پستان بود. نسیف و همکاران در سال ۲۰۱۲ (۲۱) از متون بالینی ویژگی‌های BI-RADS را استخراج و با گزارش‌نویسی دستی مقایسه کردند؛ دقت ۹۶٪ گزارش شد اما BI-RADS کلاسه‌بندی نگردید. بوزورت و همکاران در سال ۲۰۱۶ سیستم پشتیبان تصمیم مبتنی بر NLP جهت تشخیص بدخیمی از

گزارش‌های BI-RADS و متن رادیولوژی با دقت ۹۷/۵۸٪ ارایه کردند (۲۵).

این مقاله، به دنبال ایجاد یک سیستم پشتیبان‌تصمیم مبتنی‌بر پیش‌بینی مقادیر BI-RADS و زیرگروه‌های مولکولی است. برای این منظور ابتدا گزارشات ماموگرافی با استفاده از NLP، پردازش‌شده و با استفاده word2vec (۲۶) تبدیل به بردار شد. ۱۸ ویژگی Hospital information system (HIS) از پرونده الکترونیک بیماران استخراج گردید. این متغیرها شامل دو متغیر عددی و شانزده متغیر اسمی می‌باشند، که در کنار بردار مستخرج از گزارش ماموگرافی، قرار گرفتند. همچنین با استفاده از روش بدون نظارت c–میانگین، کلاس زیرگروه‌های مولکولی نمونه‌ها خوشه‌بندی گردید و به داده‌های هر خوشه مقدار زیرگروه مولکولی تخصیص داده‌شد. برای کلاسه‌بندی و تعیین BI-RADS از ماشین بردار پشتیبان چندکلاسه استفاده شد. در ادامه، از همان بردار ویژگی (متن‌کاوی و HIS) با کمک بیزین ساده، جهت تعیین زیرگروه‌های مولکولی استفاده شد.

باتوجه به مطالعه تحقیقات گذشته که بیان گردید در تشخیص BI-RADS از تصاویر و متون ماموگرافی استفاده می‌گردد ولی از داده‌های HIS و زیرگروه‌های مولکولی برای تشخیص BI-RADS استفاده نمی‌شود لذا در این پژوهش اطلاعات پرونده سلامت الکترونیک بیماران و زیرگروه‌های مولکولی در کنار اطلاعات تصاویر و گزارش ماموگرافی قرار گرفت تا ایجاد تفاوت معنادار در تشخیص BI-RADS با افزودن این اطلاعات افزوده، مشخص گردد. برای انجام این‌کار از ترکیب دو الگوریتم بیزین ساده و الگوریتم ماشین بردار پشتیبان استفاده گردید.

مواد و روش‌ها

در این پژوهش یک مدل پیش‌بینی تشخیص BI-RADS ارایه گردید. مجموعه داده‌ها شامل دو منبع گزارش‌های ماموگرافی و سوابق پرونده الکترونیک بیماران (مستخرج از HIS) است. این مجموعه داده، شامل گزارش‌های ماموگرافی و سوابق الکترونیک ۱۲۰۰ بیمار از مرکز آموزشی درمانی نمازی در بازه زمانی سال‌های ۱۳۹۴ تا ۱۳۹۶ می‌باشد. گزارش‌های متون ماموگرافی دارای ۱۶۰ ویژگی و سایر سوابق پرونده الکترونیک دارای ۱۸ ویژگی است. این ۱۸ ویژگی در بخش‌های الف و ب، از جدول ۱ مشاهده می‌گردد. جدول ۱ الف شامل ۲ متغیر، مربوط به ویژگی‌های عددی و جدول ۱ ب شامل شانزده متغیر، مربوطبه ویژگی‌های اسمی است که به انضمام ۱۶۰ ویژگی مربوطبه گزارش‌های متن ماموگرافی، مجموعاً ۱۷۸ ویژگی برای هر بیمار استخراج گردید.

همچنین اطلاعات ۱۲۰۰ بیمار، شامل ۷ طبقه BI-RADS است که در جدول ۲، گزارش تعداد این بیماران در هر طبقه آمده است.

شکل ۱ در پنج فاز مراحل مختلف فرآیند پیشنهادی را نمایش داده است. در فاز اول مجموعه‌داده به‌دست می‌آید که شامل گزارش‌های ماموگرافی و اطلاعات HIS هر فرد است. از آنجاکه گزارشات ماموگرافی، متون آزاد می‌باشند بنابراین با استفاده از روش‌های NLP، پردازش و تبدیل به بردار شدند. در فاز دوم با مشاوره پزشک ویژگی‌های مهم در سیستم اطلاعات بیمارستانی انتخاب شدند. در فاز سوم، از آنجایی‌که در مجموعه داده تنها BI-RADS مشخص شده است لذا در ابتدا باید کلاس زیرگروه‌های مولکولی مشخص شود از این‌رو تمام داده‌ها با روش بدون نظارت c– میانگین در چهارخوشه مربوطبه جدول ۲، خوشه‌بندی گریدند و مقادیر زیرگروه‌های مولکولی به هر خوشه تخصیص داده شد. در فاز چهارم یک مدل آموزش دیده‌شده برای پیش‌بینی مقادیر BI-RADS با استفاده از SVM و یک مدل آموزش دیده‌شده برای پیش‌بینی مقادیر زیرگروه‌های مولکولی با استفاده از بیزین ساده ارایه شد و نتایج در فاز پنجم توسط شاخص‌های ارزیابی اعتبارسنجی شد.

مجموعه داده‌های ما شامل دو منبع اصلی است: گزارش‌های ماموگرافی و سوابق پرونده الکترونیک بیماران (زیرمجموعه HIS). این داده‌ها از اطلاعات موجود در مرکز آموزشی درمانی نمازی، مربوطبه سال‌های ۱۳۹۴ تا ۱۳۹۶، دریافت شد. این مجموعه داده شامل گزارش‌های ماموگرافی سوابق پرونده الکترونیک ۱۵۰۰ بیمار است. از آنجایی‌که اطلاعات برخی از بیماران کامل نبود، در نهایت تنها از ۱۲۰۰ استفاده شد.

شکل ۲ بلوک‌های روش پیشنهادی برای طبقه‌بندی گزارش‌های پزشکی و نحوه استخراج یک بردار از یک گزارش ماموگرافی را نشان می‌دهد. لازم به ذکر است که در اینجا تنها فرآیند پردازش متن نشان داده شده‌است.

در مرحله پیش‌پردازش، محتوای گزارش‌های ماموگرافی توسط کتابخانه NLTK (۲۷) ریشه‌یابی شده و حروف اضافی، علایم نگارشی به استثنای کلمات منفی‌ساز، حذف گردید. برای مثال «توده قابل لمس در پستان و زیر بغل مشاهده نشد» جمله‌ای منفی است و نشان منفی‌ساز در آن حذف نمی‌شود. اعداد صحیح و اعشاری به رشته مربوطه تبدیل شدند. جهت حفظ وابستگی‌های محلی، جمع‌آوری بایگرام (Bi-gram) از جفت کلمه‌های ممکن براساس اطلاعات متقابل محاسبه می‌شود. برای بهبود صحت تعبیه‌بندی (Embedding) کلمات، بایگرام‌ها بارخداد کمتر از ۵۰ مورد، حذف و بارخداد بالای ۱۰۰۰ مورد به‌عنوان کلمه‌ای واحد در نظر گرفته شدند. در ادامه کلمات کلیدی که در فرهنگ لغت وجود دارد از متن انتخاب گردید. اگر در

جمله کلمه منفی‌ساز وجود داشت، متضاد شده و یا بردار آن معکوس شد. به‌عنوان مثال در جمله بالا، احتمال دارد فرد سرطان پستان نداشته باشد. بنابراین «توده قابل لمس» از آنجایی‌که در دیکشنری وجود دارد، متضاد می‌شود و در صورتی‌که کلمه متضادی برای آن پیدا نشد حاصل بردار Word2vec آن معکوس می‌شود. به‌منظور کاهش ابهامات و بهبود دقت معنایی گزارشات از هستی‌شناسی دامنه در بخش پردازش متن استفاده شد. اینکار توسط یک پویشگر واژگانی (۲۸) انجام شد که کار آن تشخیص اصطلاحات به‌دست‌آمده‌ای است که با اصطلاحات از پیش تعریف‌شده، ریشه مشترک دارند که ما آنها را به اصطلاح‌های کنترل شده (اصطلاحات کلیدی) نگاشت کردیم. علاوه‌بر دیکشنری، از اصطلاحات عمومی در دسترس (CLEVER) (۲۸) و مورد استفاده در شناسایی زمینه‌های بالینی و نگاشت استفاده کردیم.

پس از ترکیب اصطلاحات کلیدی و اصطلاحات به‌دست‌آمده از CLEVER، مجموعاً ۳۵۰ کلید به‌دست آمد که عمدتاً برای دو هدف استفاده می‌شوند: (الف) گزارش‌ها را از طریق نگاشت کاهش می‌دهد؛ (ب) به تولید بردارهای آگاه از متن کمک می‌کند. برای ایجاد تعبیه‌برداری کلمات از روش بدون نظارت، با استفاده از مدل Word2vec استفاده شده است (۲۶). برای آموزش Word2vec از Skipgram با طول بردار ۱۶۰ و عرض پنجره ۸، که این مقادیر بر اساس آزمون و خطا به‌دست آمد، استفاده شد. در هر گزارش از اصطلاحات کلیدی منتخب برای توصیف آن متن استفاده کردیم. سپس میانگین تمام برداری‌های به‌دست‌آمده نشان‌دهنده بردار آن متن است. هر بردار گزارش بر اساس رابطه ۱ محاسبه شد.

رابطه ۱

$$V_{MTR} = \frac{1}{N}\sum_{i=1}^{N} V_{W_i}$$

جدول ۱ الف- ویژگی‌های کمی مستخرج از HIS

نام متغیر	توضیحات متغیر	افراد سالم (n=۱۰۱)	افراد بیمار (n=۱۰۹۹)
		انحراف معیار±میانگین	میانگین±انحراف معیار
سایز	سایز سینه	۱۷/۶۹±۳/۷۳	۲۳/۰۳±۵/۴۸
سن	سن مراجعه‌کنندگان/بیماران	۴۷/۹±۱۴/۴۴	۵۴/۳۷±۱۲/۶۵

جدول ۱ب- ویژگی‌های کیفی استخراج‌شده از HIS

نام متغیر	توضیحات متغیر	افراد سالم (n=۱۰۱)	افراد بیمار (n=۱۰۹۹)
		تعداد	تعداد
ترشح			
	وجود/ عدم وجود سابقه ترشحات غیرطبیعی از پستان		
		No=۴۱	No=۵۹۴
		Yes=۶۰	Yes=۵۰۵
جهت			
	با سه وضعیت: سینه سمت چپ، سمت راست یا هر دو جهت		
		Left=۵۴	Left=۵۴۳
		Right=۳۲	Right=۴۷۹
		Bilateral=۱۵	Bilateral=۷۷
دخانیات			
	وجود/ عدم وجود سابقه استعمال سیگار		
		No=۹۴	No=۹۹۱
		Yes=۷	Yes=۱۰۸
درد پستان			
	وجود/ عدم وجود سابقه درد در ناحیه پستان‌ها		
		No=۱۹	No=۸۳
		Yes=۸۲	Yes=۱۰۱۶
سابقه بارداری			
	وجود/ عدم وجود سابقه بارداری		
		No=۱۶	No=۲۰۱
		Yes=۸۵	Yes=۸۹۸
سابقه بیماری			

وجود/عدم وجود سابقه سایر بیماری مانند دیابت، تیروئید، خونریزی‌های داخلی، جراحی و غیره

No=۳۷۸
Yes=۷۲۱

No=۴۵
Yes=۵۶

سابقه شیردهی

وجود/عدم وجود سابقه شیردهی

No=۷۹
Yes=۱۰۲۰

No=۲۲
Yes=۷۹

شکل

شکل سینه، که بر اساس ژنتیک، سن، وزن و سطح هورمون فرد می‌تواند متفاوت باشد با سه وضعیت: بیضی، گرد و بی‌قاعده

Oval=۴۰۵
Round=۲۰۳
Irregular=۴۹۱

Oval=۳۶
Round=۴۱
Irregular=۲۴

فعالیت ورزشی

بیمار سابقه هرگونه فعالیت ورزشی دارد یا خیر

No=۸۶۴
Yes=۲۳۵

No=۷۱
Yes=۳۰

قاعدگی منظم

وجود/عدم وجود قاعدگی‌های منظم با توجه به سن

No=۶۰۱
Yes=۴۹۸

No=۴۸
Yes=۵۳

مصرف الکل

وجود/عدم وجود سابقه سوء مصرف الکل

No=۳۲
Yes=۱۰۶۷

No=۴
Yes=۹۶

مصرف قرص ضدبارداری

مصرف/ عدم مصرف قرص ضدبارداری

No=۵۶۸
Yes=۵۳۱

No=۳۶
Yes=۶۵

وراثت

وراثت در سه گروه تقسیم شد. افرادی که هیچ‌گونه سابقه سرطان خانوادگی ندارند. افرادی که سابقه سرطان‌های دیگر دارند و افرادی که سابقه خانوادگی در سرطان سینه دارند

No=۱۰۹
Yes Breast=۲۵۴
Yes=۶۹۶

No=۳۲
Yes Breast=۵
Yes=۶۲

وضعیت تأهل

وجود/عدم وجود سابقه تأهل؛ بیمار مجرد است یا متأهل

Single=۳۹۷
Married=۷۰۲

Single=۱۴
Married=۸۷

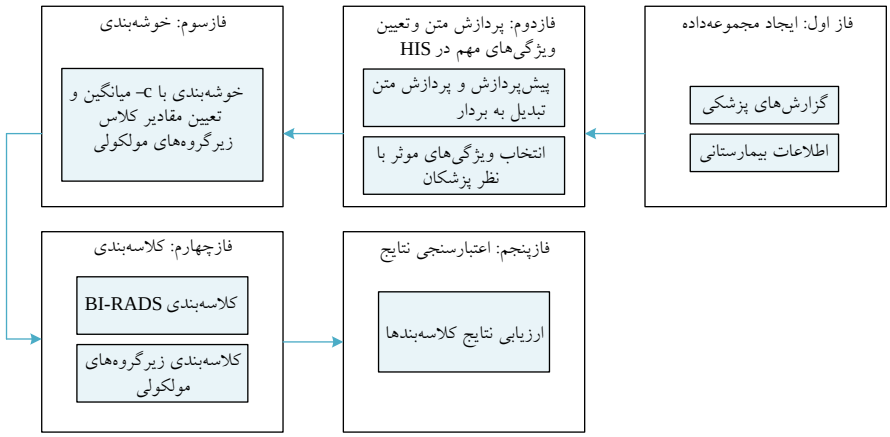
ویژگی‌های مرتبط

وجود/عدم وجود هریک از موارد زیر به‌عنوان ویژگی‌های وابسته در سوابق مراجعه‌کننده
ضخیم‌شدگی پوست، جمع‌شدگی پوست، جمع‌شدگی نوک‌پستان، اعوجاج ساختاری، آدنوپاتی زیر بغل و توده‌های کلسیمی

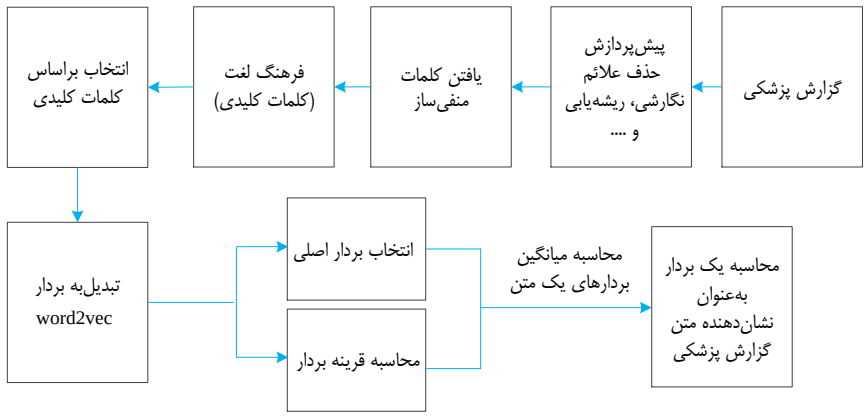
Skin thickening=۱۴۹
Skin retraction=۲۰۱
Nipple retraction=۱۰۲
Architectural distortion=۱۴۵
Axillary adenopathy=۲۷۴
Calcification=۲۲۸

Skin thickening=۱۲
Skin retraction=۵
Nipple retraction=۱۸
Architectural distortion=۲۳

Axillary adenopathy=۲۷ Calcification=۱۶	
یائسگی	
ورود/ عدم ورود به دوره یائسگی	
No=۸۱۵	No=۷۴
Yes=۲۸۴	Yes=۲۷



شکل ۱- فازبندی روش پیشنهادی



شکل ۲- تبدیل گزارش به بردار

جدول ۲- تقسیم‌بندی تعداد بیماران برحسب طبقه BI-RADS

طبقه‌بندی	تعداد بیماران
BI-RADS 0	۱۰۱
BI-RADS 1	۳۷۶
BI-RADS 2	۳۳۳
BI-RADS 3	۴۳
BI-RADS 4	۱۵۱
BI-RADS 5	۸۲
BI-RADS 6	۱۱۴
جمع کل	۱۲۰۰ نفر

که در این رابطه، V_{MTR} بردار نشان‌دهنده گزارش، N تعداد کلمات انتخاب‌شده از گزارش و V_{Wi} بردار هر کلمه به‌دست آمده از Word2vec است.

در اینجا HIS از PACS و پرونده الکترونیک بیماران در مرکز آموزشی درمانی نمازی بین سال‌های ۱۳۹۴ تا ۱۳۹۶ استخراج شده است. PACS شامل سوابق الکترونیکی برای ذخیره و بازیابی تصاویر پزشکی و اسناد و گزارش‌های مرتبط است. HIS یک سیستم اطلاعاتی یکپارچه برای پوشش تمام جنبه‌های عملکرد بیمارستانی

مانند خدمات مالی، اداری، سلامت بیماران، حقوقی و غیره است. بانک اطلاعاتی از اطلاعات مربوطبه سیستم PACS در مراکز آموزشی درمانی استفاده می‌کند.

سرطان پستان، بیماری ناهمگن با چندین زیرگروه مولکولی مجزا براساس وضعیت گیرنده و ایمونوشیمی است، که شامل گیرنده استروژن (ER)، گیرنده پروژسترون (PR)، گیرنده عامل رشد اپیدرمی 2-neu(HER2)، نشانگر تکثیر Ki67 و گیرنده عامل رشد اپیدرمی (EGFR) است. چهار مجموعه اصلی زیرگروه مولکولی وجود دارد: لومینA، لومینB، بیانگر فاکتور رشد اپیدرمی انسانی (HER2 over-expression) و سرطان پستان با رده مولکولی Basal-like (BLBC). هر نوع از زیرگروه‌های مولکولی میزان عود و بقا را نشان می‌دهد که مهمترین عامل در انتخاب تکنیک‌های مختلف درمانی است (۲۹). در این پژوهش از روش بدون نظارت c–میانگین نمونه‌ها برای تعیین زیرگروه‌های مولکولی استفاده شد. بدین‌صورت که در ابتدا تمام بیماران با استفاده از c–میانگین (و ویژگی‌های حاصل از فاز دوم) در چهارخوشه قرار گرفتند و پس از اتمام مراحل خوشه‌بندی، براساس مقادیر مراکز خوشه، برای هرخوشه زیرگروه‌های مولکولی تعیین گردید.

در الگوریتم خوشه‌بندی c–میانگین، نمونه‌ها به C (تعداد زیرگروه مولکولی) خوشه تقسیم می‌شوند که تعداد c از قبل مشخص شده‌است. تابع هدف به‌صورت رابطه ۲ می‌باشد.

$$J = \arg\min\left(\sum_{i=1}^{n}\sum_{j=1}^{C} u_{ij}^{m} \| x_i - c_j \|^2\right) \qquad \text{رابطه ۲}$$

در این رابطه، m یک عدد حقیقی بزرگ‌تر از ۱ است که در اکثر موارد برای m عدد ۲ انتخاب می‌شود، n تعداد نمونه، c مراکز خوشه، u درجه عضویت و x نمونه‌ها می‌باشند. برای کمینه کردن مقدار j، درجه عضویت و مراکز خوشه در هر تکرار به‌ترتیب با روابط ۳ و ۴ به‌روز می‌شوند.

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\| x_i - c_j \|}{\| x_i - c_k \|}\right)^{\frac{2}{m-1}}} \qquad \text{رابطه ۳}$$

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^{m} . x_i}{\sum_{i=1}^{n} u_{ij}^{m}} \qquad \text{رابطه ۴}$$

معیارهای اصلی بدین‌شرح بود: ۱) تنها افرادی که سرطان پستان داشتند خوشه‌بندی شدند؛ ۲) بر اساس نتایج هیستوشیمیایی ایمنی پس از عمل جراحی یا بیوپسی مطابق با سیزدهمین کنفرانس بین‌المللی سرطان پستان سنت‌گالن سال ۲۰۱۳، هر یک از خوشه‌ها با یکی از چهار زیرگروه مولکولی مختلف تشخیص داده شدند. زیر گروه‌های مولکولی به همراه ایمونوفنوتیپ در جدول ۳ نشان داده شده‌است.

بنابراین ارتباطی منطقی بین تقسیم‌بندی BI-RADS و زیرگروه‌های مولکولی به‌دست آوردیم. حال با استفاده از یک کلاسه‌بند می‌توانیم بر اساس اطلاعات BI-RADS زیرگروه‌های‌مولکولی را به‌صورت احتمالی تشخیص دهیم.

Support vector machine (SVM) ابرصفحه‌ای با حداکثر حاشیه بین دو کلاس را می‌یابد که ابرصفحه جداکننده بهینه نامیده می‌شود. در اینجا از تابع کرنل RBF استفاده شد و پس از استخراج مدل (۳۰)، مقادیر احتمالی برای هر کلاس از BI-RADS به‌دست آمد. در اینجا از نرمال‌سازی به‌روش انحراف معیار (۳۱) استفاده شد. برای تشخیص BI-RADS که دارای هفت کلاس است، از هفت ماشین بردار پشتیبان استفاده گردید. بنابر جدول ۴ به ازای یک نمونه، هفت ماشین‌بردار پشتیبان تصمیم‌گیری کرده‌اند و با توجه به اینکه در این مثال، ماشین‌بردار پشتیبان چهارم بیشترین احتمال را نشان می‌دهد، بنابراین نمونه متعلق به کلاس چهارم یا"Probably benign" (۱۵) است.

جدول ٤- مقادیر بردارهای پشتیبان

SVM 1	SVM 2	SVM 3	SVM 4	SVM 5	SVM 6	SVM 7
۰/۰۵	۰/۰۷	۰/۰۳	۰/۷۹	۰/۰۲	۰/۰۱	۰/۰۳

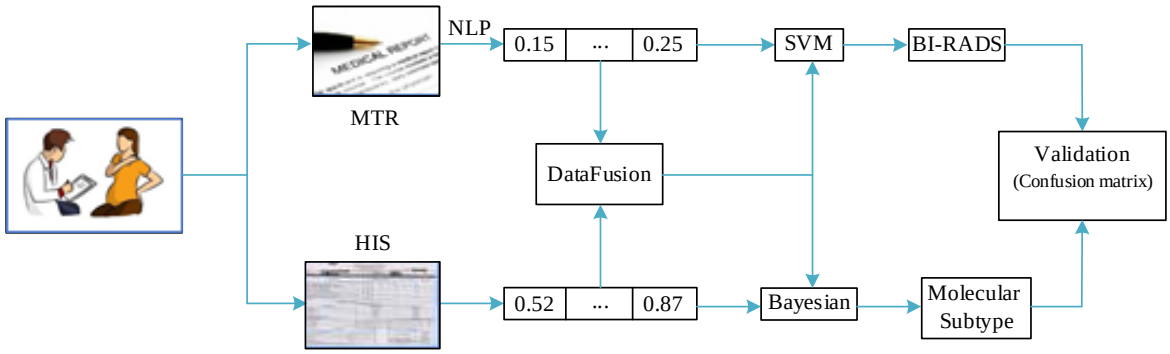
در این مقاله از بیزین ساده به‌عنوان کلاسه‌بند برای تشخیص زیرگروه‌های مولکولی استفاده گردید. از آنجا که طبقه‌بندی بیزین روشی احتمالی برای طبقه‌بندی است، خروجی به‌دست‌آمده به‌صورت احتمال است. برای طبقه‌بندی نمونه جدید، محتمل‌ترین طبقه را با داشتن مقادیر صفات $(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_k})$ که توصیف‌کننده نمونه جدید است، $\mathbf{P(C = c|X = x)}$ را شناسایی می‌کند. قضیه بیز به‌صورت رابطه ۵ می‌باشد.

$$\left(C = c \middle| X = x\right) \ \alpha \ P\left(C = c\right) P\left(X = x \middle| C = c\right) \qquad \text{رابطه ۵}$$

حال با استفاده از داده‌های آموزشی سعی می‌کنیم دو جمله معادله بالا را تخمین بزنیم. محاسبه از روی داده‌های آموزشی به این‌صورت که میزان تکرار c در داده‌ها چقدر است، آسان می‌باشد. اما محاسبه جملات مختلف $\mathbf{P(X = x|C = c)}$ به این‌صورت قابل قبول نخواهد بود مگر اینکه حجم زیادی از داده‌های آموزشی را در اختیار داشته باشیم. بیزین ساده فرض را بر استقلال متغیرهای پیش‌بینی می‌گذارد و روشی احتمالی برای طبقه‌بندی است که برای طبقه‌بندی نمونه جدید، محتمل‌ترین طبقه را با داشتن مقادیر صفات که توصیف‌کننده نمونه جدید است، شناسایی می‌کند (۳۲).

جدول ۳- انواع زیرگروه‌های مولکولی و ایمونوفنوتیپ (۲۹)

زیرگروه مولکولی	ایمونوفنوتیپ
Luminal A	ER+and/or PR+, HER2−, CK5/6±, and Ki67 < 14%
Luminal B	ER+and/or PR+, CK5/6±, HER2+, or Ki67≥14%; or PR < 20%
HER2	ER−, PR−, HER2+, CK5/6±
BLBC	ER−, PR−, HER2− (triple negative), CK5/6+, and/or EGFR+



شکل ۳- فرآیند روش پیشنهادی

جدول ۵- ماتریس درهم‌ریختگی (۳۳)

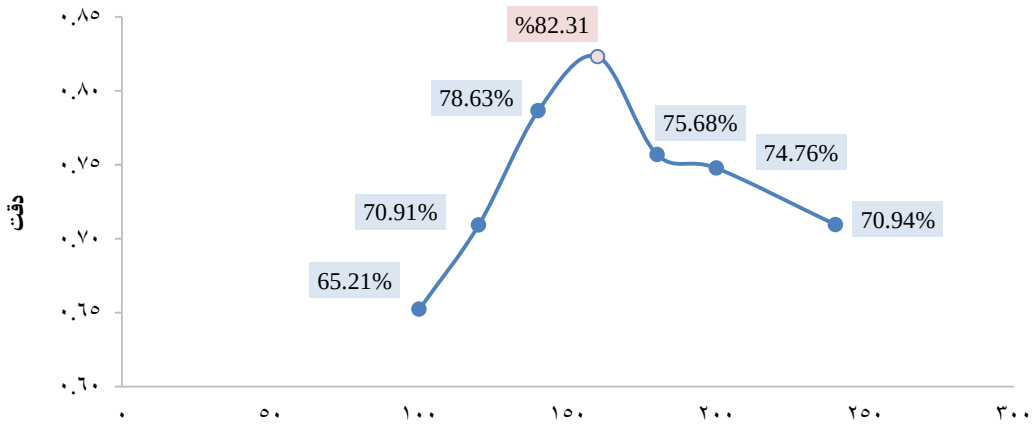
	مقادیر اصلی/ واقعی	
مقادیر پیش‌بینی شده	مثبت کاذب (FPj1)	مثبت صحیح (TP11)
	کلاس j که به اشتباه کلاس ۱ تشخیص داده شده است	کلاس ۱ که به‌درستی کلاس ۱ تشخیص داده شده است
	منفی صحیح (TNji)	منفی کاذب (FNi1)
	کلاس j که به‌درستی کلاس j تشخیص داده شده‌اند	کلاس ۱ که به اشتباه کلاس i تشخیص داده شده است

جدول ٦- مقایسه مقادیر شاخص‌های ارزیابی (مقادیر BI-RADS)

Confusion Matrix							ID Class	Sensitivity(%)	Specificity(%)	PPV(%)	NPV(%)	F1 Measure(%)	Accuracy(%)
۷۵	۹	۲	۳	۹	۳	۰	$Class_1$	۷۰/۷۵	۹۷/۳۴	۷۴/۲۶	۹۶/۸۴	۷۲/۴۶	
۵	۳۴۰	۱۰	۵	۵	۲	۹	$Class_2$	**۹٤/۱۸**	۹۵/۰۱	۹۰/۴۳	۹۷/۰۳	**۹۲/۲۷**	
۴	۵	۳۰۰	۳	۶	۸	۷	$Class_3$	۹۲/۰۲	۹۵/۶۵	۹۰/۰۹	۹۶/۵۴	۹۱/۰۵	
۲	۰	۰	۴۰	۰	۰	۱	$Class_4$	۶۳/۴۹	**۹۹/۷۰**	**۹۳/۰۲**	۹۷/۷۲	۷۵/۴۷	۸۵/۴۲
۸	۴	۳	۲	۱۳۰	۲	۲	$Class_5$	۸۲/۲۸	۹۷/۷۱	۸۶/۰۹	۹۶/۹۷	۸۴/۱۴	
۳	۱	۵	۳	۵	۶۰	۵	$Class_6$	۷۳/۱۷	۹۷/۷۷	۷۳/۱۷	**۹۷/۷۷**	۷۳/۱۷	
۹	۲	۶	۷	۳	۷	۸۰	$Class_7$	۷۶/۹۲	۹۶/۵۳	۷۰/۱۸	۹۷/۵۲	۷۳/۳۹	

جدول ۷- مقایسه مقادیر شاخص‌های ارزیابی (مقادیر زیرگروه مولکولی)

Confusion Matrix				ID Class	Sensitivity	Specificity	PPV	NPV	F1 Measure	Accuracy
۲۲۰	۲۰	۲۵	۱۶	$Class_1$	**٪۸۳/۰۲**	٪۸۳/۵۶	**٪۷۸/۲۹**	٪۸۷/۳۲	**٪۸۰/۵۹**	٪۷۲/۳۱
۱۵	۱۴۰	۱۹	۱۶	$Class_2$	٪۷۴/۸۷	٪۸۸۵/۶۴	٪۷۳/۶۸	٪۸۹/۲۴	٪۷۴/۲۷	
۱۹	۱۵	۸۰	۱۸	$Class_3$	٪۵۶/۷۴	٪۸۹/۶۴	٪۶۰/۶۱	٪۸۸/۰۶	٪۵۸/۶۱	
۱۱	۱۲	۱۷	۹۰	$Class_4$	٪۶۴/۲۹	**٪۹۱/٦۷**	٪۶۹/۲۳	**٪۸۹/۸۰**	٪۶۶/۶۷	



شکل ٤- تغییرات میزان دقت با تغییر ابعاد در بردار حاصل از word2vec

در فاز پنجم نتایج BI-RADS و زیرگروه‌های مولکولی تشخیص داده‌شده توسط شاخص‌های ارزیابی اعتبارسنجی می‌شود. براساس شکل ۳، با فرض مراجعه فرد به سیستم درمان، درابتدا براساس متون پزشکی (MTR: Medical Text Report) که در این کار گزارش‌های ماموگرافی است و همچنین سوابق پرونده الکترونیک بیمار از HIS، مقادیر BI-RADS و زیرگروه‌های مولکولی، به‌ترتیب با استفاده از SVM و بیزین ساده تخمین‌زده شده و خروجی اعتبارسنجی می‌گردد.

برای پیاده‌سازی این طرح، از رایانه با مشخصات زیر استفاده شد.

Processor:	Intel Skylake Core i7-6700 K
Installed memory (RAM):	4*8 GB DDR RAM
VGA:	GTX 1080
HDD:	256GB SSD +1TB SATA

سیستم عامل مورداستفاده از شرکت مایکروسافت به‌نام Windows 10 64bit بود و برای مدلسازی برنامه از Python 3.7 در محیط Spyder استفاده شد.

بنابر جدول ۵، ماتریس درهم‌ریختگی (matrix Confusion) از معیارهای ارزیابی کلاسه‌بندها و یک ماتریس مربعی N در N است؛ که N بیانگر تعداد کلاس‌هاست. قطر اصلی تعداد تشخیص‌های درست و سایر آرایه‌ها تشخیص‌های اشتباه را نشان می‌دهد. از آنجایی‌که در اینجا برای BI-RADS، هفت کلاس و برای زیرگروه‌های مولکولی، چهار کلاس داریم بنابراین برای هر کلاس به‌طور مجزا شاخص‌های ارزیابی به‌دست آمد.

با استفاده از این ماتریس، شاخص‌هایی همچون Sensitivity، Specificity، PPV، NPV، F1 و Accuracy به‌دست می‌آید (۳۳ و ۳۴).

نتایج

شکل ۴ میزان دقت برای ماشین بردار پشتیبان جهت تشخیص BI-RADS و صرفاً با استفاده از متن‌کاوی را نشان می‌دهد. می‌بینیم که با افزایش ابعاد در بردار حاصل‌از متن، دقت کلاسه‌بندی افزایش‌یافته و این مقدار در ابعاد بالاتر از ۱۶۰ سیر نزولی دارد. در بسیاری از بررسی‌ها (۳۵) نیز با افزایش ابعاد، کیفیت word2vec و متعاقباً دقت کاهش‌یافته که این مسأله با کاهش و افزایش ابعاد نیز بررسی شد. نهایتاً از آنجا که بیشینه دقت در ابعاد ۱۶۰ و برابر با ۸۲.۳۱٪ به‌دست آمد لذا این مقدار به‌عنوان ابعاد پایه برای سایر پردازش‌ها مورد استفاده قرار گرفت.

جدول ۶ به‌ترتیب میزان حساسیت (Sensitivity)، خاصیت (Specificity)، ارزش اخباری مثبت (PPV)، ارزش اخباری منفی (NPV)، F1 و دقت را برای کلاسه‌بندی BI-RADS و به همین شکل جدول ۷ همین مقادیر را برای زیرگروه‌های مولکولی نشان می‌دهند. در جدول ۶ کلاس‌های سه الی هفت نشان‌دهنده مقادیر متناظر در BI-RADS دو الی شش است. بیشتر کلاس‌های بیماری با دقت بالای ۷۰٪ تشخیص داده‌شد. پیگیری بیماری متناظر با BI-RADS=2 و BI-RADS=5 بیشترین میزان حساسیت به‌ترتیب برابر با ۹۲/۰۲٪ و ۸۲/۲۸٪ را دارند و کمترین میزان، کلاس چهارم (۶۳/۴۹٪) است که می‌تواند به‌دلیل تعداد کم این نمونه‌ها در مجموعه‌داده باشد. بیشترین میزان حساسیت برای زیرگروه‌های مولکولی (جدول ۷) مربوط به کلاس یک (۸۳/۰۲٪) و کمترین آن مربوط‌به کلاس سوم (۵۶/۷۴٪) است.

مقدار خاصیت برای افراد سالم در جدول ۶ برابر ۹۵/۰۱٪ است که نشان از عملکرد بالای تشخیص افراد سالم دارد. به‌طورکلی به‌ازای هر

کلاس BI-RADS مقدار خاصیت بیش از ۹۵٪ است و بیشینه آن مربوطبه کلاس چهارم (۹۹/۷۰٪) است. بیشترین مقدار خاصیت در زیرگروه‌های مولکولی (جدول ۷) مربوطبه کلاس یک (۸۳/۰۲٪) و کمترین آن مربوطبه کلاس سوم (۵۶/۷۴٪) است. مقادیر نشان می‌دهد عملکرد روش پیشنهادی از نظر مقادیر خاصیت مناسب است.

مقدار PPV در جدول ۶ برای افراد سالم برابر ۹۰/۴۳٪ است که نشان از عملکرد خوب روش پیشنهادی دارد و برای سایر کلاس‌ها نیز مناسب و کمینه آن ۷۰/۱۸٪ (کلاس هفت) است. این مقدار در جدول ۷ و برای تشخیص زیرگروه‌های مولکولی برای تمام کلاس‌ها بین ۶۰/۶۱٪ (کمینه در کلاس سوم) و ۷۸/۲۹٪ (بیشینه در کلاس اول) است. که نشان از تشخیص مناسب هر کلاس دارد.

مقدار NPV در جدول ۶ برای افراد سالم برابر ۹۷/۰۳٪ است که نشان از عملکرد خوب روش پیشنهادی دارد و برای سایر کلاس‌ها نیز مناسب و کمینه آن ۹۶/۵۴٪ (کلاس سوم) است. کمینه این مقدار برای تشخیص زیرگروه‌های مولکولی در جدول ۷، برابر ۸۷/۳۲٪ (کلاس اول) و بیشینه آن، ۸۹/۸۰٪ (کلاس چهارم) می‌باشد.

میانگین هارمونیک (harmonic) (F1) مربوطبه نتایج BI-RADS و زیرگروه مولکولی به‌ترتیب در جداول ۶ و ۷ آمده است. بیشینه F1 برای تشخیص BI-RADS برابر ۹۲/۲۷٪ (کلاس دوم) و کمینه آن برابر ۷۲/۴۶٪ (کلاس اول) است که نشان از تشخیص مناسب روش پیشنهادی دارد. از طرفی بیشینه F1 برای تشخیص زیرگروه‌های مولکولی، ۸۰/۵۹٪ (کلاس اول) و کمینه آن ۵۸/۶۱٪ (کلاس سوم) می‌باشد.

میزان دقت (Accuracy) یا توانایی آزمون، در افتراق صحیح موارد بیمار و سالم، در کلاسه‌بندی BI-RADS و زیرگروه‌های مولکولی به‌ترتیب مقادیر ۸۵/۴۲٪ و ۷۲/۳۱٪ می‌باشد. از طرفی مقایسه دقت کلاسه‌بندی BI-RADS با استفاده از متون پزشکی، نسبت‌به زمانی‌که از HIS استفاده نشده (شکل ۴)، نشان‌دهنده تأثیر HIS در افزایش این شاخص است.

درنتیجه با کنارهم قراردادن شاخص‌های ارزیابی مشخص می‌شود که روش پیشنهادی در تشخیص کلاس‌های BI-RADS، زیرگروه‌های مولکولی و نهایتاً کمک به تشخیص تعیین روال بیماری عملکرد خوبی داشته است. ازآنجایی‌که در اینجا روشی جدید برای پردازش متن ارایه شده است و همچنین از مقادیر HIS در کنار نتایج پردازش متن استفاده می‌شود، لذا عملکرد روش پیشنهادی نسبت‌به کارهای مشابه از نظر تشخیص BI-RADS و زیرگروه‌های مولکولی بهبودیافته است.

بحث

کالج رادیولوژی آمریکا برای یکسان‌سازی گزارش‌های ماموگرافی استانداردی بنام BI-RADS را ارایه نمود. این سیستم سبب همگون‌سازی گزارش‌ها شد و نقشی اصلی در پیشبرد برنامه‌ریزی درمانی استاندارد ایفا کرد، چراکه می‌توان از آن برای اولویت‌بندی دقیق پیشبرد درمان استفاده نمود. اما این رویکرد معایبی همچون تفاوت نظر بین پزشکان برای نتیجه‌گیری مقدار BI-RADS داشت. از این‌رو در این مقاله پیشنهاد شد تا از اطلاعات سوابق پرونده الکترونیک افراد نیز استفاده شود. بنابراین رویکردی ترکیبی از داده‌های بدون‌ساختار (گزارش‌های ماموگرافی) و داده‌های ساختاریافته (پرونده سوابق الکترونیک از HIS) استفاده شده است. بدین‌شکل که پس از پیش پردازش و پردازش متون، کلمات کلیدی با استفاده از Word2vec تبدیل‌به بردار شدند که در هر متن میانگین بردارهای کلمات کلیدی، نشان‌دهنده آن متن بودند. برای هر متن یک بردار ۱۶۰ بُعدی (۱۶۰ ویژگی) به‌دست آمد. سپس ۱۸ ویژگی از پرونده الکترونیک بیماران استخراج شده است. این متغیرها شامل ۲ متغیر عددی و شانزده متغیر اسمی می‌باشند که در کنار بردار مستخرج از گزارش ماموگرافی، قرار گرفتند و به انضمام ۱۶۰ویژگی مربوطبه گزارش‌های ماموگرافی، در مجموع ۱۷۸ ویژگی برای کلاسه‌بندی مورد استفاده قرار گرفت. همچنین از ماشین‌بردار پشتیبان برای تعیین کلاس‌های BI-RADS و از بیزین ساده برای تعیین کلاس زیرگروه‌های مولکولی استفاده شد. نتایج در قالب شاخص‌های ارزیابی مختلف همچون حساسیت، خاصیت، ارزش اخباری مثبت، ارزش اخباری منفی، مقدار F1 و دقت مورد ارزیابی قرار گرفت. نتایج بیشینه شاخص‌های ارزیابی برای تخمین BI-RADS به‌ترتیب ۹۴/۱۸٪، ۹۹/۷۰٪، ۹۳/۰۲٪، ۹۷/۷۷٪ و ۹۲/۲۷٪ و نتایج کمینه شاخص‌های ارزیابی به‌ترتیب ۶۳/۴۹٪، ۹۵/۰۱٪، ۷۰/۱۸٪، ۹۶/۵۴٪ و ۷۲/۴۶٪ می‌باشد. این نتایج، برای برای تخمین زیرگروه‌های مولکولی در حالت بیشینه به‌ترتیب ۸۳/۰۲٪، ۹۱/۶۷٪، ۷۸/۲۹٪، ۸۹/۸۰٪ و ۸۰/۵۹٪ و در حالت کمینه به‌ترتیب ۵۶/۷۴٪، ۸۳/۵۶٪، ۶۰/۶۱٪، ۸۷/۳۲٪ و ۵۸/۶۱٪ می‌باشد. دقت تشخیص مقادیر BI-RADS برابر ۸۵/۴۲٪ و در زیرگروه‌های مولکولی ۷۲/۳۱٪ می‌باشد. روش پیشنهادی برای تبدیل متن به بردار و استفاده از سوابق پرونده الکترونیک بیماران برای تشخیص BI-RADS و همچنین تشخیص زیرگروه‌های مولکولی بر اساس گزارش‌های ماموگرافی متن پزشکی و ویژگی‌های HIS، می‌تواند پزشک را برای تصمیم‌گیری بهتر کمک می‌کند. این رویکرد نسبت‌به کارهای مشابه، سبب بهبود تشخیص وجود یا عدم‌وجود بیماری،

همچنین تعیین سطح بیماری شده‌است؛ لذا پزشک می‌تواند روال درمانی فرد را با دقت بیشتری تعیین کند.

در این کار، از تلفیق‌داده برای بهبود دقت استفاده شد. پیشنهاد می‌شود در تحقیقات بعدی از ترکیب شواهد برای افزایش کارآیی سیستم استفاده شود. همچنین از آنجا که تصاویر ماموگرافی اطلاعات مفیدی در اختیار پزشک قرار می‌دهند، پیشنهاد دیگر استفاده از تکنیک‌های تلفیق تصمیم و یادگیری عمیق، در تخمین دقیق‌تر سطح‌بیماری و زیرگروه‌های مولکولی است تا در تصمیم‌سازی دقیق‌تر روال درمان کمک بیشتری به پزشکان نمود.

تشکر و قدردانی

مقاله حاضر بخشی از رساله دکترای نویسنده اول، مصوب با کد ۱۶۳۴۸۲۲۹۷۵۱۷۰۵۱۱۳۹۸۷۱۷۱۱ پایان‌نامه دانشگاه آزاد اسلامی واحد شیراز می‌باشد و نویسندگان از تمامی حمایت‌های آن مجموعه تشکر و قدردانی می‌نمایند.

References

1. Balakumar P, Maung-U K, Jagadeesh G. Prevalence and prevention of cardiovascular disease and diabetes mellitus. Pharmacological research 2016;113:600-9. doi: 10.1016/j.phrs.2016.09.040
2. Institute for research education and treatment of cancer (n d). Available from: http://ncii.ir/about-us.
3. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians 2018;68:394-424. doi: 10.3322/caac.21609
4. Centers for disease control and prevention. cancer. bearst. cancer statistics data visualizations tool 2017. Available from: https://gis.cdc.gov/Cancer/USCS/DataViz.html.
5. Isfahani P, Hossieni Zare SM, Shamsaii M. The prevalence of depression in iranian women with breast cancer: a meta-analysis. Quarterly of Horizon of Medical Sciences 2020;26:170-81. doi: 10.32598/hms.26.2.3207.1
6. Tabarestani S, Noori-Daloii mr. Molecular genetics, diagnosis and treatment of breast cancer: review article. Journal of Sabzevar University of Medical Sciences 2010;17:74-87.
7. Dehghan P, Mogharabi M, Zabbah I, Layeghi K, Maroosi A. Modeling Breast cancer using data mining methods. Journal of Health and Biomedical Informatics 2018;4:266-78.
8. Ginsburg O, Yip CH, Brooks A, Cabanes A, Caleffi M, Dunstan Yataco JA, et al. Breast cancer early detection: a phased approach to implementation. Cancer 2020;126:2379-93. doi:10.1002/cncr.32887
9. Sadeghi S, Golabpour A. An Algorithm for Predicting Recurrence of Breast Cancer Using Genetic Algorithm and Nearest Neighbor Algorithm. Journal of Health and Biomedical Informatics 2020;6:309-19.
10. Alaa AM, Moon KH, Hsu W, Van Der Schaar M. Confidentcare: A clinical decision support system for personalized breast cancer screening. IEEE Transactions on Multimedia 2016;18:1942-55. doi:10.1109/TMM.2016.2589160
11. Mazo C, Kearns C, Mooney C, Gallagher WM. Clinical Decision Support Systems in Breast Cancer: A Systematic Review. Cancers 2020;12:369. doi:10.3390/cancers12020369
12. Sim LLW, Ban KHK, Tan TW, Sethi SK, Loh TP. Development of a clinical decision support system for diabetes care: A pilot study. PloS one 2017;12:e0173021. doi:10.1371/journal.pone.0173021
13. Park J, Rho MJ, Moon HW, Park YH, Kim C-S, Jeon SS, et al. Prostate cancer trajectory-map: clinical decision support system for prognosis management of radical prostatectomy. Prostate International 2020. doi:10.1016/j.prnil.2020.06.003
14. Esmaeili M, Ayyoubzadeh SM, Ahmadinejad N, Ghazisaeedi M, Nahvijou A, Maghooli K. A decision support system for mammography reports interpretation. Health Information Science and Systems 2020;8:1-8. doi:10.1007/s13755-020-00109-5
15. Gossman W, Shikhman R, Keppke AL. Breast, Imaging, Reporting and Data System (BI RADS). StatPearls [Internet]: StatPearls Publishing; 2019.
16. Farrokh D, Alamdaran SA, Feizy A, Soleimany H. Diagnostic value of BIRADS method using sonography in evaluating the level of malignancy of breast masses compared with biopsy. The Iranian Journal of Obstetrics, Gynecology and Infertility 2019;22:1-6. doi:10.22038/ijogi.2019.13738
17. Vanderheyden B, Xie Y. Mammography Image BI-RADS Classification Using OHPLall. 2020.
18. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, et al. Automated annotation and classification of BI-RADS assessment from radiology reports. Journal of biomedical informatics 2017;69:177-87. doi: 10.1016/j.jbi.2017.04.011
19. Gao H, Bowles EJA, Carrell D, Buist DS. Using natural language processing to extract mammographic findings. Journal of biomedical informatics 2015;54:77-84. doi:10.1016/j.jbi.2015.01.010
20. Gupta A, Banerjee I, Rubin DL. Automatic information extraction from unstructured mammography reports using distributed semantics. Journal of biomedical informatics 2018;78:78-86. doi:10.1016/j.jbi.2017.12.016
21. Nassif H, Cunha F, Moreira IC, Cruz-Correia R, Sousa E, Page D, et al., editors. Extracting BI-RADS features from Portuguese clinical texts. 2012 IEEE International Conference on Bioinformatics and Biomedicine; 2012: IEEE.
22. Percha B, Nassif H, Lipson J, Burnside E, Rubin D. Automatic classification of mammography reports by BI-RADS breast tissue composition class. Journal of the American Medical Informatics Association 2012;19:913-6. doi: 10.1136/amiajnl-2011-000607
23. Sippo DA, Warden GI, Andriole KP, Lacson R, Ikuta I, Birdwell RL, et al. Automated extraction of BI-RADS final assessment categories from radiology reports with natural language processing. Journal of Digital Imaging 2013;26:989-94. doi: 10.1007/s10278-013-9616-5
24. Zhang X, Zhang Y, Zhang Q, Ren Y, Qiu T, Ma J, et al. Extracting comprehensive clinical information for breast cancer using deep learning methods. International Journal of Medical Informatics 2019;132:103985. doi: 10.1016/j.ijmedinf.2019.103985
25. Bozkurt S, Gimenez F, Burnside ES, Gulkesen KH, Rubin DL. Using automatically extracted information from mammography reports for decision-support. Journal of Biomedical Informatics 2016;62:224-31.
26. Guo D, Wang Q, Liang M, Liu W, Nie J. Molecular cavity topological representation for pattern analysis: A NLP analogy-based word2vec method. International Journal of Molecular Sciences 2019;20:6019. doi: 10.3390/ijms20236019
27. Loper E, Bird S. NLTK: the natural language toolkit. arXiv preprint cs/0205028 2002.
28. Banerjee I, Bozkurt S, Alkim E, Sagreiya H, Kurian AW, Rubin DL. Automatic inference of BI-RADS final assessment categories from narrative mammography report findings. Journal of Biomedical Informatics 2019;92:103137. doi: 10.1016/j.jbi.2019.103137

29. Kao K-J, Chang K-M, Hsu H-C, Huang AT. Correlation of microarray-based breast cancer molecular subtypes and clinical outcomes: implications for treatment optimization. BMC Cancer 2011;11:143. doi: 10.1186/1471-2407-11-143

30. Chang C-C, Lin C-J. LIBSVM: A library for support vector machines. ACM transactions on intelligent systems and technology (TIST) 2011;2:1-27. doi: 10.1145/1961189.1961199

31. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome biology 2019;20:1-15. doi: 10.1186/s13059-019-1874-1

32. Sen PC, Hajra M, Ghosh M. Supervised classification algorithms in machine learning: A survey and review. Emerging Technology in Modelling and Graphics: Springer; 2020. p. 99-111.

33. Tharwat A. Classification assessment methods. Applied Computing and Informatics 2020. doi: 10.1016/j.aci.2018.08.003

34. Shahabi M, Hassanpour H. Using the artificial intelligence techniques for diagnosing of intensity of non-alcoholic fatty liver disease by clinical parameters. Journal of Knowledge & Health 2016;11:69-75. doi: 10.22100/jkh.v11i3.1369

35. Li B, Drozd A, Guo Y, Liu T, Matsuoka S, Du X. Scaling word2vec on big corpus. Data Science and Engineering 2019;4:157-75. doi: 10.1007/s41019-019-0096-6

Journal of Knowledge & Health
in Basic Medical Sciences 2020; Vol 15, No 3

DOI: 10.22100/jkh.v15i3.2454

Original Article

Proposing a Clinical Decision Support System for Breast Cancer Diagnosis

Mostafa Boroumandzadeh (Ph.D. Student)[1], Elham Parvinnia (Ph.D.)[1*]

1- Dept. of Computer Engineering, Shiraz Branch, Islamic Azad University, Shiraz, Iran.

Received: 5 October 2020**, Accepted:** 7 December 2020

Abstract:

Introduction: *Breast cancer is one of the leading causes of death in women. Therefore, the accuracy and speed of diagnosis are crucial in the treatment procedure. In this regard, the BI-RADS classification system has been used for the standardization of mammography reports. However, there is much disagreement among physicians about the BI-RADS values. The aim of this paper is to diagnose BI-RADS by natural language processing of mammography reports and clinical information from the electronic health records and combining them to identify molecular subtypes and help patient follow-up.*

Methods: *In this study, 1200 mammography reports and electronic health records obtained from Namazi Educational and Medical Center for years between 2015-2017. After text processing, the vector with 160 features was obtained, then 18 features were extracted by referring to the electronic health records. Finally, 178 features were used by SVM and naïve Bayesian to predict BI-RADS and molecular subtypes, respectively.*

Results: *The values of Accuracy, Positive Prediction Value, Negative Prediction Value, Sensitivity, and Specificity were calculated to evaluate the results. Accuracy was 85.42% for BI-RADS and 72.31% for molecular subtypes.*

Conclusion: *The proposed decision support system was an appropriate model to help the physician to diagnose breast cancer and categorize patients. It was also found that the combined information, including electronic medical records of patients and designated molecular subtypes along with mammography reports, can be useful in diagnosing the disease and defining the treatment follow-up.*

Keywords: Decision support system, Molecular subtypes, Clinical information, Breast cancer.

Conflict of Interest: No

***Corresponding author:** E. Parvinnia**, Email:** parvinnia@iaushiraz.ac.ir*

Citation: Boroumandzadeh M, Parvinnia E. Proposing a clinical decision support system for breast cancer diagnosis. Journal of Knowledge & Health in Basic Medical Sciences 2020;15(3):54-66.