

طبقه‌بندی خودکار BI-RADS در گزارش‌های ماموگرافی با استفاده از تلفیق داده

محمد ذهابی^۱، محمد ابراهیم شیرینی^{۲*}، حمید حاج سید جوادی^۳، مصطفی برومندزاده^۴

^۱ گروه مهندسی کامپیوتر، واحد بروجرد، دانشگاه آزاد اسلامی، بروجرد، ایران، ^۲ گروه ریاضی و علوم کامپیوتر، دانشگاه صنعتی امیرکبیر، تهران، ایران، ^۳ گروه ریاضی و علوم کامپیوتر، دانشگاه شاهد، تهران، ایران، ^۴ گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام نور، تهران، ایران
تاریخ وصول: ۱۴۰۲/۰۸/۱۷ تاریخ پذیرش: ۱۴۰۲/۱۲/۰۷

چکیده

زمینه و هدف: سرطان پستان یکی از شایع‌ترین سرطان‌ها در زنان و عامل اصلی مرگ و میر در بیماری‌های سرطانی است و ماموگرافی روش تصویربرداری اولیه برای تشخیص زودهنگام توده‌های پستان است. تشخیص سریع با دقت بالا یکی از دغدغه‌های جدی پزشکان و مراکز بهداشتی درمانی در مواجهه با بیماری‌های خاص است، لذا هدف از این مقاله تعیین و طبقه‌بندی خودکار BI-RADS در گزارش‌های ماموگرافی با استفاده از تلفیق داده بود.

روش بررسی: این یک مطالعه توصیفی - تحلیلی و گذشته نگر می‌باشد که در سال ۱۴۰۲ انجام شد، گزارش ماموگرافی و پرونده الکترونیکی بیمار که به ترتیب از سیستم بایگانی و ارتباط تصویر و سوابق بیمار استخراج شدند، از اطلاعات موجود در مرکز آموزشی درمانی بیمارستان شهیدزاده شهرستان بهبهان به دست آمده است که شامل گزارش ماموگرافی و پرونده الکترونیکی ۲۵۰ بیمار است که اطلاعات کامل داشتند. برای مدل‌سازی روش پیشنهادی با استفاده از داده‌های جمع آوری شده، از نرم افزار پایتون در محیط ویژوال استودیو کد استفاده شده است. در نهایت از صحت‌سنجی متقاطع برای ارزیابی کیفیت و اعتبار نتایج استفاده شد.

یافته‌ها: نتایج نشان داد رویکرد پیشنهادی یعنی استفاده از تکنیک تبدیل کلمه به بردار به همراه روش فراوانی وزنی کلمات کلیدی و تلفیق آن‌ها با HIS تأثیر قابل توجهی بر دقت طبقه‌بندی متون پزشکی دارد. برای طبقه‌بندی سطح سیستم گزارش و داده تصویربرداری پستان (BI-RADS) از بردارهای خروجی تکنیک تبدیل کلمه به بردار در زمانی که از روش فراوانی وزنی کلمات کلیدی استفاده می‌شود و یا بدون استفاده از روش فراوانی وزنی کلمات کلیدی و همچنین با تلفیق HIS و نیز بدون HIS برای طبقه‌بندی کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و k- نزدیک‌ترین همسایه استفاده شد و نتایج به وسیله معیارهای ارزیابی دقت، ویژگی، حساسیت، ارزش پیش‌بینی شده مثبت، ارزش پیش‌بینی شده منفی و امتیاز f1 با هم مقایسه شدند. نتایج نشان می‌دهد که بهترین دقت در روش پیشنهادی با طبقه‌بندی کننده پرسپترون چندلایه برابر با ۹۸/۷۴ درصد می‌باشد، اما بدون HIS دقت همین طبقه‌بندی کننده برابر با ۹۲/۲۳ درصد به دست آمد.

نتیجه‌گیری: ترکیب تکنیک تبدیل کلمه به بردار با روش فراوانی وزنی کلمات کلیدی می‌تواند دقت طبقه‌بندی متن را افزایش دهد، اما سابقه پزشکی که در تشخیص بیماری مهم است، می‌تواند دقت را بهبود ببخشد. نتایج نشان می‌دهد نباید صرفاً بر روی بررسی‌های پزشکی تمرکز کرد و از سایر اطلاعات بالینی و سابقه بیماران نیز باید استفاده کرد. بنابراین استفاده از HIS در کنار گزارش‌های پزشکی می‌تواند طبقه‌بندی سیستم گزارش و داده تصویربرداری پستان را بهبود ببخشد و تأثیر مثبتی بر تشخیص و فرآیندهای درمانی داشته باشد.

واژه‌های کلیدی: طبقه‌بندی متون پزشکی، سرطان پستان، استخراج ویژگی، سیستم گزارش و داده تصویربرداری پستان، سیستم اطلاعات بیمارستان

*نویسنده مسئول: محمد ابراهیم شیرینی، تهران، دانشگاه صنعتی امیرکبیر، گروه ریاضی و علوم کامپیوتر

Email: shiri@aut.ac.ir

"نشریه علمی پژوهشی ارمغان دانش وابسته به دانشگاه علوم پزشکی یاسوج، یک نشریه با دسترسی آزاد است و تمامی مقالات منتشر شده در این نشریه به صورت دسترسی آزاد منتشر می‌شوند."

مقدمه

کند. رویکردهای متن‌کاوی در یافته‌های پژوهش‌های پزشکی با اهداف متفاوتی از جمله طبقه‌بندی خودکار بیماری‌ها (۶) استفاده شده است. متن‌کاوی شکاف بین شکل ساختار یافته اطلاعات و متن را پوشش می‌دهد و از تکنیک‌های پردازش زبان طبیعی^(۴)، یادگیری ماشین و مدیریت دانش برای پردازش اسناد متنی استفاده می‌کند. از سوی دیگر، از متن‌کاوی برای تبدیل داده‌های ضروری از متن به قالب منطقی و عددی استفاده شده است تا بتوان آنها را تجزیه و تحلیل کرد (۷).^۱

در سال‌های اخیر، پژوهش‌های زیادی برای ترکیب رویکردهای متن‌کاوی با متون پزشکی انجام شده است. برخی از این پژوهش‌ها در تکنیک‌های یادگیری ماشین تخصص داشتند، برخی دیگر نیز مستقیماً به سیستم‌های پشتیبانی تصمیم، مهندسی ویژگی و استخراج ویژگی گزارش‌های متنی مرتبط بودند. سان و همکاران الگوریتمی به نام *فراوانی وزنی در فاکتورهای کلمات (WF-TF-IDF)*^(۵) برای استخراج کلمات کلیدی بر اساس فاکتورهای کلمات از صفحات وب پزشکی چینی ارائه دادند که این مطالعه با هدف بهینه‌سازی *فراوانی وزنی (TFIDF)*^(۱) و بهبود دقت و یادآوری انجام شد (۸). درسیباخ و همکاران سیستم استخراج علایمی برای ترکیب ادبیات با استفاده از

سرطان پستان یکی از شایع‌ترین سرطان‌ها در زنان و عامل اصلی مرگ و میر در بیماری‌های سرطانی است (۱). ماموگرافی روش تصویربرداری اولیه برای تشخیص زود هنگام توده‌های پستان است. تشخیص سریع با دقت بالا یکی از دغدغه‌های جدی پزشکان و مراکز بهداشتی درمانی در مواجهه با بیماری‌های خاص است (۱). با این حال، دقت در تشخیص و همچنین کیفیت و سرعت، می‌تواند مرز بین مرگ و زندگی باشد (۲). کالج آمریکایی رادیولوژی، سیستم گزارش و داده تصویربرداری پستان (BI-RADS)^(۱) را برای کاهش تنوع در توصیفات رادیولوژیست‌ها از یافته‌های مورد استفاده برای تشخیص معرفی کرد (۳).

در دو دهه اخیر، پژوهش‌های زیادی نشان داده است که سرعت و دقت تشخیص سرطان پستان می‌تواند نتایج مؤثری در درمان داشته باشد. بنابراین خودکار نمودن و بهبود دقت پردازش داده‌های پزشکی در هر مرحله از فرآیند درمان می‌تواند نتایج مؤثری داشته باشد. برخی از این پژوهش‌ها در تکنیک‌های چارچوب یادگیری عمیق^(۴) تخصص داشتند، اما مجموعه پارامترها و دانش توپولوژی در چنین چارچوب‌هایی سخت است. فرض بر این است که استفاده از متن‌کاوی^(۲) برای تجزیه و تحلیل سوابق متنی از مراکز تشخیص و درمان و سیستم بایگانی و ارتباط تصویر^(۳) ممکن است داده‌های قابل اعتماد بالینی (۵) را برای پیش‌بینی سطوح BI-RADS فراهم

1-Breast Imaging Reporting & Data System (BI-RADS)
2-Text Mining
3-Picture Archiving and Communication System (PACS)
4-Natural Language Processing (NLP)
5-Word Factors Term Frequency and Inverse Document Frequency (WF-TF-IDF)
6-Term Frequency Inverse Document Frequency (TFIDF)

بررسی قرار دادند(۱۴). لیما و همکاران (۱۵) توصیف گره‌های بافت و شکل را برای تعیین این که آیا بافت طبیعی، ضایعه خوش خیم یا بدخیم است، ترکیب و سپس از یک مدل شبکه عصبی هسته‌گرا استفاده نمودند.^۲

با توجه به مقالات بررسی شده، اگر چه اکثر محققان بر طبقه‌بندی *BI-RADS* و تشخیص سرطان پستان با استفاده از متن‌کاوی پزشکی تمرکز کردند و طبقه‌بندی سرطان پستان را مورد بحث قرار دادند، هم‌چنین برخی از محققین مانند موییدی و همکاران(۱۶)، التوخی و همکاران (۱۷)، ریاد و همکاران (۱۸) فقط تصاویر ماموگرافی را جهت طبقه‌بندی بیماران مورد بررسی قرار دادند، اما تاریخچه خانوادگی و اطلاعات بیمارستانی که برای پیش‌بینی سرطان پستان مهم است و هم‌چنین ترکیبی از تکنیک‌های استخراج ویژگی *Word2vec* و *TFIDF* با سیستم اطلاعات بیمارستان (*HIS*)^(۱) را در نظر نمی‌گیرند. بنابراین، با استفاده از روش‌های یادگیری ماشین و تلفیق داده، یک سیستم طبقه‌بندی خودکار توسعه داده شد که قادر است با دقت بالا و به طور یکپارچه درجه *BI-RADS* را مشخص کند. این روش یک رویکرد نوآورانه است که با استفاده از یک بردار وزنی که همان *TFIDF* است و ترکیب آن با *Word2vec* و در نهایت تلفیق آنها با *HIS*، می‌تواند میزان خطاهای موجود در فرآیند تفسیر را به حداقل رسانده، زمان

پردازش زبان طبیعی و متن‌کاوی به منظور پردازش الکترونیکی متن به وسیله بیمار ارایه دادند(۹). دینگ و همکاران روشی برای کشف کلمات کلیدی جدید بر اساس جاسازی کلمه ارایه دادند، هم‌چنین از تکنیک *Word2vec* برای نگاشت کلمات به فضای برداری انتزاعی *n*- بعدی و در نهایت، استخراج روابط معنایی پنهان بین کلمات استفاده نمودند(۱۰). کاسترو و همکاران یک سیستم پردازش زبان طبیعی توسعه یافته برای استخراج تمام دسته‌های *BI-RADS* از گزارش‌های رادیولوژی متنی، با استفاده از روش‌های بیزین، ماشین بردار پشتیبان و درختان تصمیم جزئی ارایه دادند(۲). نی و همکاران طرح طبقه‌بندی متون درمان بیمارستانی برای طبقه‌بندی گزارش‌های پزشکی ژاپنی را مورد بررسی قرار دادند. این مطالعه با هدف ارتقای کیفیت پرستاری انجام شد که تکنیک پایه آن *Word2vec* بود(۱۱). ناروائز و همکاران یک رویکرد خودکار برای انجام یک توصیف *BI-RADS* از چگالی را با استفاده از ماشین بردار پشتیبان سلسله مراتبی چند هسته و استراتژی بازیابی مبتنی بر شکل ارایه دادند(۱۲). اُسْتِرِس و همکاران یک سیستم مقایسه‌کننده ترکیبی برای مقایسه نظرات رادیولوژیست‌ها و هم‌چنین خروجی یک سیستم ارزیابی تصویر مبتنی بر چگالی را با استفاده از آستانه چگالی و بوت استرپینگ ارایه دادند(۱۳). دیاب و همکاران یک سیستم طبقه‌بندی متن مبتنی بر استفاده از الگوریتم یادگیری بیزین ساده برای تبدیل مسئله تخمین احتمال به یک طرح بهینه‌سازی را مورد

1-Hospital Information System (HIS)
2-Dataset

تشخیص را کاهش داده و به بهبود کلی مراقبت‌های بیماران منتهی می‌شود.

روش بررسی

این یک مطالعه توصیفی - تحلیلی و گذشته نگر می‌باشد که در سال ۱۴۰۲ انجام شد، مجموعه داده‌ای^(۲) که استفاده شد شامل دو منبع اصلی است، گزارش ماموگرافی و پرونده الکترونیکی بیمار(زیر مجموعه HIS) که به ترتیب از سیستم بایگانی و ارتباط تصویر و سوابق بیمار استخراج شده‌اند. این مجموعه داده‌ها از اطلاعات موجود در مرکز آموزشی درمانی بیمارستان شهیدزاده شهرستان بهبهان به دست آمده است که شامل گزارش ماموگرافی و پرونده الکترونیکی ۴۰۰ بیمار است و از آنجایی که اطلاعات برخی از بیماران کامل نبود، در نهایت تنها از اطلاعات ۲۵۰ بیمار که اطلاعات کامل داشتند استفاده شد. با توجه به این که شرح حال بیماران در سیستم HIS وابسته به فرد است و همچنین گزارش‌های ماموگرافی استفاده شده در این تحقیق به وسیله دستگاه‌های مختلف ماموگرافی و به وسیله افراد متفاوت تهیه گردیده که باعث بروز ناهمگنی در مجموعه داده شده است، لذا برای رفع این مشکل در ماژول پیش پردازش، داده‌های گمشده در مجموعه داده شناسایی شده و پاکسازی شدند، نرمالسازی بر روی داده‌ها انجام گرفت و برای بهبود عملکرد الگوریتم‌ها از کلمات کلیدی تخصصی استفاده گردید. در نهایت از صحت‌سنجی متقاطع^(۱) برای ارزیابی کیفیت و اعتبار

نتایج استفاده شد. از این رو، تأکید می‌شود که با توجه به ناهمگنی اولیه داده‌ها، رویکردهای کنترل کیفیت دقیق و متدولوژیک استفاده شده، تضمین‌کننده کیفیت و اعتبار داده‌های تحلیلی است. این فرآیندها اطمینان می‌دهند که اختلافات ناشی از دستگاه‌های متفاوت ماموگرافی یا گزارش‌دهندگان، بر نتایج نهایی اثر نمی‌گذارد. بر اساس معادله ۱ که به فرمول کوکران معروف است و به وسیله ویلیام کوکران ارایه شد و با توجه به آمار ابتلاء بیماری سرطان پستان در ایران(۱۹)، برای رسیدن به دقتی با خطای ۵ درصد در انجام محاسبات، به حداقل ۱۳۸ نمونه نیاز خواهد بود و با توجه به این که در این پژوهش حجم نمونه‌ها ۲۵۰ مورد است، نشان دهنده مناسب بودن حجم نمونه‌ها برای انجام این پژوهش با دقت خطای ۵ درصد است.^۳

$$n = \frac{z^2 p(1-p)}{d^2} \quad (1)$$

در این رابطه p درصد سرطان پستان در جامعه می‌باشد، برای دستیابی به دقت مورد نظر مقدار z برابر ۱/۹۶ و z^2 برابر ۳/۸۴۱۶ در نظر گرفته شد و مقدار d نیز که دقت نمونه‌گیری به این عامل بستگی دارد، برابر با ۰/۰۵ در نظر گرفته شده است و در نهایت مقدار n که حداقل حجم نمونه است برابر با ۱۳۸ نمونه به دست آمد.

1-K-fold cross-validation

نشانه‌گذاری، انتخاب کلمه‌کلیدی و تبدیل به مجموعه کلمات.

در این بخش رکوردها به قسمت‌های مربوطه و اطلاعات مفید تقلیل یافته و نرمال‌سازی انجام شد، علائم نگارشی و کلمات توقف حذف شدند، حروف بزرگ جایگزین شده و کلمات به کلمه‌های اصلی تقلیل یافتند. سپس، نشانه‌گذاری انجام شد. در این مرحله، متن‌های پیوسته به جملات یا کلمات تجزیه شدند (۲۰).

در این تحقیق، از هر دو تکنیک *Word2vec* و *TFIDF* به عنوان یک سیستم ترکیبی برای استخراج بردار ویژگی استفاده شده است.^۴

Word2vec یک روش مبتنی بر شبکه‌های عصبی است و برای نمایش کلمات به صورت برداری استفاده می‌شود. *Word2vec* یک توزیع برداری در فضای معنایی برای هر کلمه ایجاد می‌کند که این بردار می‌تواند چندین بعد داشته باشد. *Word2vec* ترکیبی از دو شبکه عصبی به نام‌های مجموعه لغات پیوسته^(۱) و کلمه هدف^(۲) است. این شبکه‌ها به گونه‌ای آموزش داده شدند که کلمات با زمینه متنی^(۳) مشترک دارای بردارهای عددی یکسانی در فضای معنایی باشند. دنباله‌ای از کلمات، به نام زمینه متن، در معادله ۲ نشان داده شده است. در اینجا، k تعداد کلمات است.

$$Context = \{w_1, w_2, \dots, w_k\} \quad (2)$$

در شبکه مجموعه لغات پیوسته، کلمات مهم با استفاده از کلمات کلیدی در همسایگی آن پیش‌بینی

برخی از ویژگی‌ها و عناصر کلیدی استخراج شده در گزارش‌های متنی پزشکی و واژه‌های کلیدی عبارتند از: چگالی، عدم تقارن، انقباض ساختار، نحوه توزیع، اندازه، سابقه جراحی پستان و شکل. همچنین برخی از ویژگی‌های بررسی شده به وسیله کارشناسان پزشکی و سوابق بیماران (مربوط به HIS) عبارت بودند از: یائسگی، سابقه شیردهی، فعالیت‌های ورزشی، سابقه بارداری، وضعیت تأهل، سن، سابقه خانوادگی سرطان و غیره.

در مرحله اول، تکنیک‌های پردازش متن دقیقاً بر روی گزارش‌های پزشکی انجام می‌شود و اطلاعات استخراج‌شده از این عمل، بر اساس واژگان ماموگرافی، به روش پیشنهادی (با استفاده از تکنیک‌های *Word2vec* و *TFIDF*) انجام می‌شود. خروجی این مرحله برای تولید یک بردار ویژگی برای هر گزارش متنی استفاده می‌شود. سپس به منظور طبقه‌بندی بردارهای حاصل با استفاده از الگوریتم‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و k -نزدیک‌ترین همسایه. از طریق مهندسی ویژگی، HIS اضافه شده و بردارهای ویژگی بهینه می‌شوند. روش پیشنهادی به پنج ماژول اصلی تقسیم می‌شود که در شکل ۱ نشان داده شده است؛ ۱- پیش پردازش گزارش متنی، ۲- استخراج ویژگی، ۳- مهندسی ویژگی، ۴- پیش‌بینی و ۵- ارزیابی.

ماژول پیش پردازش گزارش متنی پس از دریافت گزارش‌های پزشکی دارای پنج مرحله است، از جمله؛ بازیابی متن اطلاعات، نرمال‌سازی،

1-Continuous Bag-of-Words (CBOW)
2-Skip-gram
3-Context word

تکمقداری برای کلمه (یا کلمات) می‌باشد. بر این اساس، محتوای متن یک کلمه داده شده w با اندازه پنجره برابر با c به وسیله معادله ۴ تعیین می‌شود. اندازه پنجره نیز فاصله همسایگی نامیده می‌شود.

$$\text{Context}_c(w_j) = \{w_{j-c}, w_{j-c+1}, \dots, w_{j-1}, w_{j+1}, w_{j+2}, \dots, w_{j+c}\} \quad (4)$$

هدف شبکه مجموعه لغات پیوسته به حداکثر رساندن $P(w_j | \text{Context}_c(w_j))$ است. برای رسیدن به این هدف، دنباله‌ای از کلمات آموزشی مانند w_1, w_2, \dots, w_k مورد نیاز است. با استفاده از معادله ۵ و اعمال این دنباله، تابع لگاریتم درست‌نمایی^(۲) به حداکثر می‌رسد.

$$l_{CBOW} = \frac{1}{k} \sum_{j=1}^k \sum_{-c \leq i \leq c, i \neq 0} \log P(w_j | w_{j+i}) \quad (5)$$

در روش کلمه هدف کلمات در $c(w_j)$ به وسیله w_j پیش‌بینی می‌شوند و هدف به حداکثر رساندن $P(\text{Context}_c(w_j) | w_j)$ است. برای آموزش این شبکه، دنباله‌ای از کلمات مانند w_1, w_2, \dots, w_k مورد نیاز است و برای رسیدن به نتیجه در روش کلمه هدف، طبق معادله ۶، میانگین تابع لگاریتم درست‌نمایی باید حداکثر شود.

$$l_{\text{skip-gram}} = \frac{1}{k} \sum_{j=1}^k \sum_{-c \leq i \leq c, i \neq 0} \log P(w_{j+i} | w_j) \quad (6)$$

به طور کلی $P(w_0 | w_i)$ به وسیله تابعی به نام بیشینه هموار^(۳) و بر اساس معادله ۷ محاسبه می‌شود.

می‌شوند. این کلمات که در همسایگی قرار دارند محدود هستند و باید مشخص شوند. برعکس، در کلمه هدف، بر اساس یک کلمه، کلماتی که در همسایگی آن قرار دارند پیش‌بینی می‌شوند^(۲).

$TFIDF$ یک روش آماری است که در آن به هر کلمه مربوط به متن، وزن اختصاص داده می‌شود. سپس وزن کلمه i در متن j با استفاده از معادله ۳ محاسبه می‌شود.

$$w_{ij} = tf_{ij} \times \log \frac{N}{df_i} \quad (3)$$

در این رابطه w_{ij} وزن کلمه i در متن j را نشان می‌دهد، tf_{ij} فراوانی کلمه i را در سند j نشان می‌دهد، N نشان دهنده تعداد اسناد موجود می‌باشد و df_i تعداد اسناد حاوی کلمه i را نشان می‌دهد^(۲). همان‌طور که در رابطه ۳ مشخص است، از آنجایی که $TFIDF$ فقط بر تکرار کلمات در اسناد تمرکز دارد، نمی‌تواند به درستی رابطه ساختاری کلمات را در متون مفهومی پزشکی تشخیص دهد. بنابراین، در این کار سعی شد روش یادگیری دیگری با $TFIDF$ ترکیب شود.

در $Word2vec$ برای نمایش یک کلمه، فقط عنصر مربوط به همان کلمه مقدار "۱" و سایر عناصر مقدار "۰" می‌گیرند. لایه ورودی هر شبکه حاوی بردارهای کدبندی شده به روش تک‌مقداری^(۱) است. در لایه پنهان هر شبکه، بردارهای ورودی به فضای دیگری نگاشت می‌شوند. تعداد نورون‌ها در لایه پنهان ابعاد بردارهای کلمه (d بعد) را نشان می‌دهد. لایه خروجی نیز یک بردار کدبندی شده به روش

1-One-hot Encoding
2-Log-likelihood
3-Softmax

تلاش برای استفاده از *Word2vec* و *TFIDF* به طور هم‌زمان است. در واقع، بردارهای ویژگی با هر گزارشی متن‌ی بر اساس دو نکته مهم و هم‌زمان مرتبط است، محتوای گزارش از نظر ترتیب کلمات (*Word2vec*) و همچنین تعداد تکرارها و وزن کلمات کلیدی در گزارش (*TFIDF*).

به منظور افزایش کیفیت تشخیص و طبقه‌بندی، علاوه بر بردارهای استخراج شده از گزارش‌های متن‌ی پزشکی، *HIS* نیز از سیستم بایگانی و ارتباط تصویر استخراج گردید. در این راستا ۳۰ ویژگی در *HIS* وجود داشت که از ۷ پزشک متخصص درخواست شد که به هر ویژگی بر اساس اهمیت تأثیری که در تشخیص سرطان پستان دارند، از یک تا ده امتیاز دهند. سپس بر اساس میانگین امتیازات داده شده، ۲۴ ویژگی مهم برای انجام محاسبات و انجام طبقه‌بندی انتخاب شدند.

تلفیق *HIS* و *Word2vec+TFIDF* برای به دست آوردن بردارهای ویژگی انجام می‌شود. این تلفیق در شکل ۲- ب نشان داده شده است. بخش اول بردارهای عددی استخراج شده از گزارش‌های متن‌ی پزشکی با استفاده از روش‌های *TFIDF* و *Word2vec* است. در بخش دوم، مقادیر مرتبط با ویژگی‌های مشخص شده در اطلاعات *HIS* به همان بیماران اشاره دارد. بدیهی است که بردارهای ویژگی تولید شده در هر دو روش به طور مستقل به الگوریتم‌های طبقه‌بندی ارسال می‌شوند.

$$P(w_o | w_l) = \frac{\exp(V_{w_o}^T \cdot V_{w_l})}{\sum_{w=1}^W \exp(V_{w_o}^T \cdot W_{w_l})} \quad (7)$$

V_{w_o} و V_{w_l} بردار خروجی و ورودی کلمه w را نشان می‌دهند و W تعداد کلمات موجود در واژگان را نشان می‌دهد. در این تحقیق با توجه به روش آزمون و خطا (۲۲)، برای اندازه پنجره از عدد ۵ و از روش بیشینه هموار سلسله مراتبی استفاده شد. بر اساس شکل ۲-الف، ترکیبی از هر دو روش *TFIDF* و *Word2vec* برای تولید بردارهای عددی برای هر گزارش پزشکی استفاده شده است.

همان‌طور که توضیح داده شد، *Word2vec* یک بردار عددی برای هر کلمه در گزارش متن پزشکی ایجاد می‌کند و *TFIDF* نیز وزن هر کلمه را محاسبه می‌کند. اگر گزارش متن‌ی حاوی کلمات w_1, w_2, \dots, w_k باشد و بردارهای تولید شده برای این کلمات در d بعد v_1, v_2, \dots, v_k باشد (هر v دارای d بعد است)، اگر وزن کلمه i به وسیله *TFIDF* با t_i نمایش داده شود، بردار هر گزارش متن‌ی حاصل از ترکیب روش‌های *Word2vec* و *TFIDF* مطابق معادله ۸ محاسبه می‌شود.

$$\text{Vector}(\text{report}) = \frac{\sum_{i=1}^k t_i \times v_i}{k} = \frac{\sum_{i=1}^k f_i}{k} \quad (8)$$

بنابراین، برای هر گزارش پزشکی، یک برداری تولید شد که به عنوان ورودی به الگوریتم‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و k -نزدیک‌ترین همسایه به منظور پیش‌بینی مقدار *BI-RADS* داده شد. یکی از مهم‌ترین نکاتی که این مقاله را متمایز می‌کند، انتخاب واژه‌های کلیدی و سپس

پرسپترون چندلایه برای طبقه‌بندی و تشخیص BI-RADS استفاده شده است.

درخت تصمیم^(۱) از الگوریتم‌های یادگیری ماشین تحت نظارت می‌باشد که در مسائل طبقه‌بندی و رگرسیون کاربرد دارد. در این ساختارهای درختی، برگ‌ها نشان‌دهنده برچسب‌های کلاس و شاخه‌ها نشان‌دهنده ترکیبی از ویژگی‌هایی هستند که به آن برچسب‌های کلاس منتهی می‌شوند. هدف درخت تصمیم این است که بتواند مقدار یک متغیر را بر اساس مقادیر متغیرهای ورودی پیش‌بینی کند. در اینجا نیز از درخت تصمیم برای طبقه‌بندی و تشخیص BI-RADS استفاده شده است.^(۶)

الگوریتم k -نزدیک‌ترین همسایه یکی از الگوریتم‌های یادگیری ماشین با ناظر است که برای حل مسائل طبقه‌بندی و رگرسیون استفاده می‌شود. این الگوریتم برای مسائل طبقه‌بندی، k -نزدیک‌ترین همسایه را پیدا و با اکثریت آرا نزدیک‌ترین همسایگان کلاس را پیش‌بینی می‌کند. در این تحقیق نیز از این برای طبقه‌بندی و تشخیص BI-RADS استفاده شده است که با روش آزمون و خطا مقدار k برابر با ۵ قرار داده شد.

برای اجرای روش پیشنهادی از کامپیوتری با پردازنده نسل چهارم از شرکت اینتل و مقدار حافظه

در این بخش، بردارهای ویژگی تولید شده از مرحله قبل به الگوریتم‌های طبقه‌بندی ارسال می‌شوند. الگوریتم‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و k -نزدیک‌ترین همسایه برای پیش‌بینی مقدار BI-RADS در گزارش‌های بیمار مورد استفاده قرار گرفته‌اند.

شبکه‌های عصبی پیچشی^(۱) مجموعه‌ای از شبکه‌های عصبی عمیق هستند که معمولاً برای تجزیه و تحلیل تصویر، گفتار یا متن در یادگیری ماشین استفاده می‌شوند (۲۳). در این مقاله، از شبکه عصبی پیچشی به عنوان یک طبقه‌بندی‌کننده برای تشخیص BI-RADS استفاده شده است. تابع فعال‌ساز در لایه‌های پیچشی اول و دوم از نوع واحد یک‌سو شده^(۲) است. همچنین تابع فعال‌ساز لایه خروجی از نوع بیشینه هموار و تابع زیان^(۳) مجذور میانگین مربعات خطا است. لازم به ذکر است که در اینجا از الگوریتم بهینه‌سازی *Adam*^(۴) استفاده شده است که یک الگوریتم بهینه‌سازی نرخ یادگیری تطبیقی است (۲۴).

مدل پرسپترون چندلایه^(۵) شامل یک گره ورودی، گره‌های پنهان و یک بردار خروجی است که مقادیر پیش‌بینی شده را ارائه می‌دهد. لایه ورودی به هر متغیر ورودی یک نورون اختصاص می‌دهد. قابلیت محاسباتی واقعی مدل پرسپترون چندلایه در مجموع واحدهای پنهان اضافه شده بین لایه‌های ورودی و خروجی می‌باشد. داده‌ها در یک مدل پرسپترون چندلایه، مشابه یک شبکه فیدفوروارد، به جلو می‌روند و نورون با استفاده از رویکرد پس انتشار آموزش داده می‌شود. در این مقاله از مدل

1-Convolutional Neural Network (CNN)
2-Rectified Linear Unit (ReLU)
3-Loss function
4-Adaptive Moment Estimation (Adam)
5-Multi-Layer Perceptron (MLP)
6-Decision Tree (DT)

برای بیماران تشخیص داده شود که دارای ۷ کلاس است. در اینجا، TP_i مقدار مثبت واقعی کلاس i است که به مواردی اشاره دارد که کلاس واقعی i و کلاس تشخیص داده شده نیز i است. TP_i با استفاده از معادله ۹ به دست می‌آید. دو مفهوم دیگر وجود دارد، مثبت کاذب (FP) و منفی کاذب (FN)، که FP_i مقدار مثبت کاذب کلاس i است که به مواردی اشاره دارد که در آن کلاس واقعی i است، اما کلاس تشخیص داده شده غیر از i است. FP_i با استفاده از معادله ۱۰ محاسبه می‌شود. همچنین FN_i که مقدار منفی کاذب کلاس i است، نشان می‌دهد که کلاس تشخیص داده شده کلاس i است، اما کلاس واقعی غیر از i است. FN_i با استفاده از معادله ۱۱ محاسبه می‌شود. TN_i مقدار منفی واقعی کلاس i است، که به مواردی اشاره دارد که کلاس واقعی i نیست و کلاس شناسایی شده نیز i نیست. TN_i با استفاده از معادله ۱۲ به دست می‌آید.

$$TP_i = C_{ii}, \quad i = 0, 1, \dots, 6 \quad (9)$$

$$FP_i = \sum_{i \neq j=0}^6 C_{ij}, \quad i = 0, 1, \dots, 6 \quad (10)$$

$$FN_i = \sum_{i \neq j=0}^6 C_{ji}, \quad i = 0, 1, \dots, 6 \quad (11)$$

$$TN_i = \sum_{i \neq j=0}^6 \sum_{i \neq k=0}^6 C_{jk}, \quad i = 0, 1, \dots, 6 \quad (12)$$

-
- 1-Confusion matrix
 - 2-True Positive (TP)
 - 3-True Negative (TN)
 - 4-False Positive (FP)
 - 5-False Negative (FN)

اصلی ۱۶ گیگابایت و کارت گرافیک مدل جی فورس ۱۲۸ بیتی با ۲ گیگابایت حافظه استفاده شد.

در این تحقیق از سیستم عامل در این تحقیق از سیستم عامل Microsoft Windows 10 64bit و Python 3.8.7 در محیط Visual Studio Code برای مدل سازی روش پیشنهادی استفاده شده است. به منظور مقایسه کمیت پارامترهای $Word2vec+TFIDF$ از سه الگوریتم اساسی استفاده شد: $Word2vec$ (بدون HIS)، $Word2vec$ (با HIS) و $Word2vec + TFIDF$ (بدون HIS).

در ابتدا علاوه بر موارد فوق، دو الگوریتم برای حالت‌های ($TFIDF$ با HIS) و ($TFIDF$ بدون HIS) در نظر گرفته شد که به دلیل دقت طبقه‌بندی بسیار پایین حذف شدند. در نهایت، روش پیشنهادی ($Word2vec+TFIDF$ با HIS) با سایر روش‌های مطرح شده در این تحقیق و در چهار طبقه‌بندی کننده مختلف مورد ارزیابی و مقایسه قرار گرفت.

ماتریس درهم‌ریختگی^(۱) یکی از معیارهای ارزیابی طبقه‌بندی کننده‌ها است و ماتریس مربع $N \times N$ است. که در آن N تعداد کلاس‌ها را نشان می‌دهد، که در اینجا ۷ کلاس برای $BI-RADS$ وجود دارد. قطر اصلی تعداد تشخیص‌های صحیح را نشان می‌دهد و بقیه آرایه‌های ماتریس تشخیص‌های نادرست را بیان می‌کنند.

در مدل‌های طبقه‌بندی باینری که فقط مثبت یا منفی بودن بیماری تشخیص داده می‌شود، در ماتریس سردرگمی مفهوم مثبت واقعی^(۲) (TP)، منفی واقعی^(۳) (TN)، مثبت کاذب^(۴) (FP) و منفی کاذب^(۵) (FN) وجود دارد، اما در این تحقیق باید مقدار $BI-RADS$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (13)$$

$$Specificity = \frac{TN}{TN + FP} \quad (14)$$

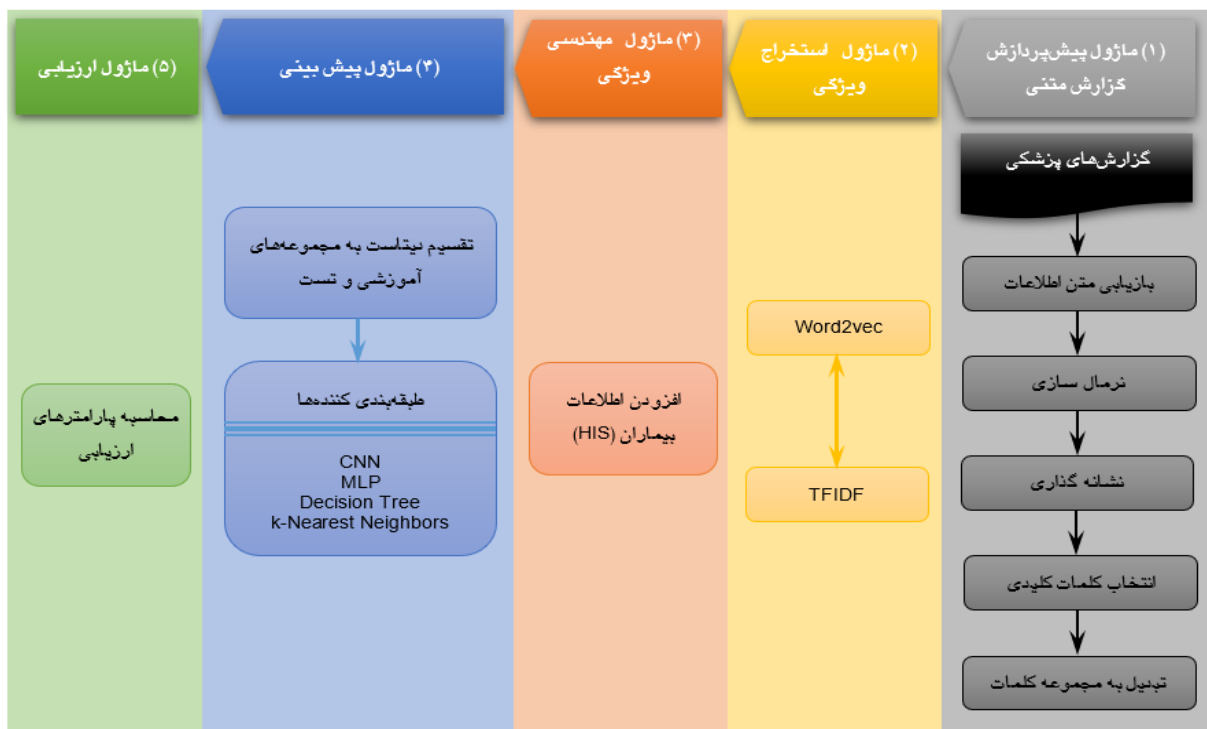
$$Sensitivity = \frac{TP}{TP + FN} \quad (15)$$

$$PPV = \frac{TP}{TP + FP} \quad (16)$$

$$NPV = \frac{TN}{TN + FN} \quad (17)$$

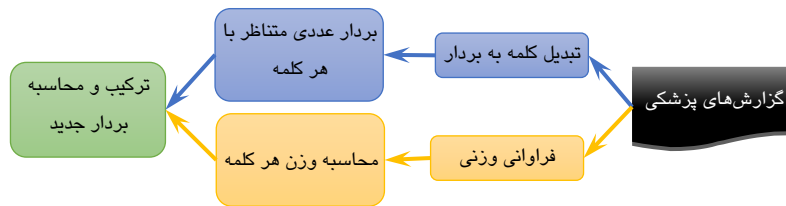
$$F1 - measure = \frac{2 \times PPV \times Sensitivity}{PPV + Sensitivity} \quad (18)$$

پارامترهایی مانند دقت^(۱)، ویژگی^(۲)، حساسیت^(۳)، ارزش پیش‌بینی شده مثبت^(۴)، ارزش پیش‌بینی شده منفی^(۵) و امتیاز $f1$ ^(۶) به ترتیب با استفاده از معادلات ۱۳ تا ۱۸ محاسبه می‌شوند که در آنها TP ، TN ، FP و FN به ترتیب به عنوان میانگین TP_i ، TN_i ، FP_i و FN_i نشان داده می‌شوند^(۲۵).



شکل ۱: نقشه راه روش پیشنهادی^۸

- 1-Accuracy
- 2-Specificity
- 3-Sensitivity
- 4-Positive Predicted Value (PPV)
- 5-Negative Predicted Value (NPV)
- 6-F1-measure



الف) سیستم ترکیبی طراحی شده برای تولید یک بردار عددی از هر گزارش پزشکی



ب) سیستم تلفیق‌داده استخراج شده از سیستم بایگانی و ارتباط تصویر و سابقه پزشکی بیماران برای تولید بردارهای گزارش

شکل ۲: شماتیک روش پیشنهادی

یافته‌ها

در اینجا، گزارش‌های پزشکی برای استخراج *BI-RADS* با استفاده از پردازش زبان طبیعی، *TFIDF* و *HIS* پردازش شدند. در حالی که چندین مقاله در مورد استخراج مفید اطلاعات بالینی از گزارش‌های ماموگرافی از طریق سیستم‌های مختلف پردازش زبان طبیعی منتشر شده است (۲)، اما سیستمی مبتنی بر متون پزشکی و *HIS* که بتواند *BI-RADS* را تشخیص دهد، مورد مطالعه قرار نگرفته است. بنابراین در این کار با استفاده از *Word2vec* و *TFIDF* بردارهای ویژگی از متون پزشکی استخراج شد و سپس ویژگی‌های مهم *HIS* با نظر پزشک متخصص انتخاب شده که برای طبقه‌بندی استفاده شدند. در این بخش، نتایج روش پیشنهادی به تفصیل مورد بحث قرار گرفته است. شکل ۳ بر تأثیر ابعاد مختلف بردارهای ویژگی بر دقت طبقه‌بندی کننده‌ها تمرکز دارد. در واقع، هنگام نگاشت کلمات در فضای برداری، با استفاده از

Word2vec، هیچ قانون واحدی برای تعیین تعداد ابعاد بردارها وجود ندارد (۲۶) که انتخاب ابعاد مناسب بر اساس آزمایش‌های مکرر صورت گرفت. بنابراین برای به دست آوردن بهترین نتایج در مجموعه داده مورد استفاده از روش آزمون و خطا استفاده شد و ابعاد برداری ۱۱۰ تا ۲۰۰ بر اساس مقالات ارزیابی شد. دقت طبقه‌بندی‌کننده‌ها با افزایش ابعاد افزایش یافته تا در ۱۶۰ به حداکثر رسید. سپس با افزایش ابعاد دقت کاهش یافت. طبقه‌بندی‌کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه به حداکثر دقت در ۱۶۰ رسیده‌اند، بنابراین، بردار با ابعاد ۱۶۰ انتخاب گردید. کاهش دقت به دلیل افزایش ابعاد به علت خطای محاسباتی است (۲۶).

همان‌طور که ذکر شد، بعد ۱۶۰ برای بردارها انتخاب شده است زیرا بهترین دقت را در هر الگوریتمی دارد. شکل ۳- الف دقت طبقه‌بندی‌های مختلف

را بدون استفاده از *TFIDF* و دخالت ویژگی‌های *HIS* نشان می‌دهد در حالی که در شکل ۳ همین دقت با *HIS* نمایش داده شده است و از *TFIDF* استفاده نشده است. در شکل ۳-ج نیز بررسی استفاده از *TFIDF* را بدون استفاده از ویژگی‌های *HIS* نشان می‌دهد و شکل ۳-د نیز تأثیر تداخل *HIS* و استفاده از *TFIDF*، روش پیشنهادی، را در افزایش دقت طبقه‌بندی کننده‌ها نسبت به بقیه حالت‌ها نشان می‌دهد. در جدول ۱ نتایج طبقه‌بندی کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه را برای طبقه‌بندی سطوح *BI-RADS* با و بدون مشارکت *HIS* در زمانی که از *TFIDF* استفاده می‌شود و یا استفاده نمی‌شود، به وسیله معیارهای ارزیابی دقت، ویژگی، حساسیت، ارزش پیش‌بینی شده مثبت، ارزش پیش‌بینی شده منفی و امتیاز *f1* با هم مقایسه شدند. نتایج نشان می‌دهد که الگوریتم پرسپترون چندلایه کارآمدتر است و کمی با روش شبکه عصبی پیچشی اختلاف دارد. دقت روش‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه برای *Word2vec* با استفاده از *TFIDF* و بدون استفاده از ویژگی‌های *HIS* به ترتیب ۸۴/۹۱، ۹۲/۲۳، ۸۹/۸۳ و ۹۰/۹۷ درصد است. دقت *Word2vec* به همراه *TFIDF* و با استفاده از ویژگی‌های *HIS* در طبقه‌بندی کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه به ترتیب: ۹۴/۷۴، ۹۸/۷۴، ۹۰/۹۷ و ۹۳/۰۳ درصد است.

همان‌طور که در جدول ۱ نشان داده شد، روش پیشنهادی یعنی *Word2vec* با کمک *TFIDF* و با استفاده از *HIS* بهتر از سایر روش‌ها عمل می‌کند. دقت، نسبت نمونه‌های طبقه‌بندی شده به وسیله طبقه‌بندی کننده در یک کلاس معین، به تعداد کل نمونه‌هایی است که طبقه‌بندی کننده در آن کلاس به درستی یا نادرست طبقه‌بندی کرده است. همان‌طور که از معادله ۱۳ پیداست، دقت نشان می‌دهد که چه نسبتی از کلاس‌های پیش‌بینی شده واقعاً در آن کلاس هستند که دقت طبقه‌بندی کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه در روش پیشنهادی به ترتیب: ۹۴/۷۴، ۹۸/۷۴، ۹۰/۹۷ و ۹۳/۰۳ درصد می‌باشد. پارامتر دیگری که در جدول ۱ نشان داده شده است، حساسیت هر طبقه‌بندی کننده است که نسبت طبقه‌بندی درست نمونه‌ها در کلاس‌های معین به وسیله طبقه‌بندی کننده به تعداد نمونه‌های آن کلاس را نشان می‌دهد، در روش پیشنهادی برای طبقه‌بندی کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه به ترتیب: ۷۹/۵۰، ۹۴/۷۱، ۶۵/۲۸ و ۷۳/۰۳ درصد می‌باشد. ویژگی هر طبقه‌بندی کننده که نشان دهنده نسبت منفی بودن پاسخ به یک کلاس است با توجه به این که آن نمونه واقعاً در آن کلاس نباشد که در روش پیشنهادی برای طبقه‌بندی کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه به ترتیب: ۹۶/۸۷، ۹۹/۲۵، ۹۴/۶۵ و ۹۵/۸۴ درصد

بررسی در این تحقیق، نیاز به یک ماتریس درهم ریختگی داریم. در جدول ۲، ردیف‌ها نشان‌دهنده کلاس‌های پیش‌بینی شده به وسیله طبقه‌بندی‌کننده‌ها و ستون‌ها نشان‌دهنده کلاس‌های واقعی هستند. مقادیر کلاس‌های واقعی، نتایج ماموگرافی هستند که رادیولوژیست‌ها آن را تعیین کرده‌اند و در مجموعه داده ثبت شده است و از آن برای آموزش و تست طبقه‌بندی‌کننده‌ها استفاده می‌شود. با توجه به این جداول، تعداد اشتباهات هر طبقه‌بندی‌کننده برای هر کلاس در مقایسه با سایر کلاس‌ها مشخص است. می‌توان نتیجه گرفت که مرز بین کلاس‌های تعریف شده برای *BI-RADS* بسیار حساس است. با توجه به جدول ۱، روش پیشنهادی یعنی استفاده از *Word2vec* به همراه *TFIDF* و *HIS* دقت بیشتری دارد که ماتریس درهم‌ریختگی طبقه‌بندی‌کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k* نزدیک‌ترین همسایه برای این روش در جدول ۲ نمایش داده شده است.

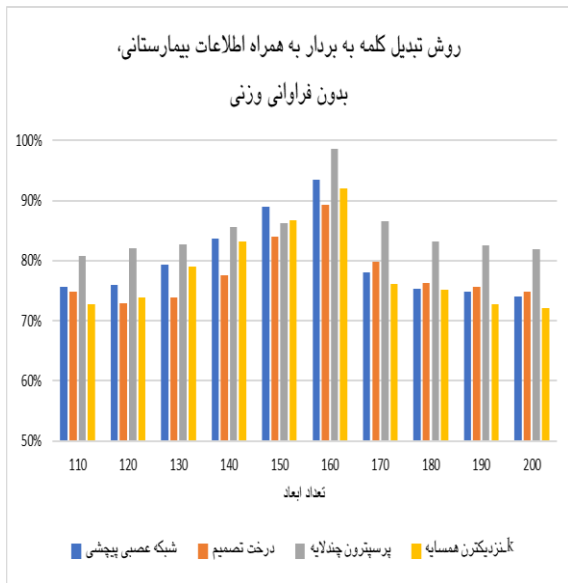
در شکل ۴ پارامترهای مختلفی از جمله دقت، حساسیت، ویژگی، ارزش پیش‌بینی شده مثبت، ارزش پیش‌بینی شده منفی و امتیاز *f1* برای این طبقه‌بندی‌کننده‌ها مورد ارزیابی که قرار گرفتند، نمایش داده شده است که برای روش پیشنهادی و با طبقه‌بندی‌کننده پرسپترون چندلایه به ترتیب مقادیر ۹۸/۷۴، ۹۴/۷۱، ۹۹/۲۵، ۹۵/۸۶، ۹۹/۲۹ و ۹۷/۵۱ درصد به عنوان بهترین نتیجه حاصل شد.

می‌باشد. ارزش پیش‌بینی شده مثبت هر طبقه‌بندی‌کننده پارامتر دیگری است و نشان‌دهنده نسبت نتایج مثبتی است که واقعاً مثبت هستند که در روش پیشنهادی برای طبقه‌بندی‌کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k* نزدیک‌ترین همسایه به ترتیب: ۸۴/۹۰، ۹۵/۸۶، ۷۰/۷۰ و ۷۶/۵۳ درصد می‌باشد. پارامتر دیگری که مورد بررسی و ارزیابی قرار گرفت ارزش پیش‌بینی شده منفی هر طبقه‌بندی‌کننده است که نشان‌دهنده نسبت نتایج منفی است که واقعاً منفی هستند که در روش پیشنهادی برای طبقه‌بندی‌کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k* نزدیک‌ترین همسایه به ترتیب: ۹۷/۰۵، ۹۹/۲۹، ۹۴/۸۳ و ۹۶/۰۲ درصد می‌باشد. امتیاز *f1* هر طبقه‌بندی‌کننده پارامتر دیگری است که در روش پیشنهادی برای طبقه‌بندی‌کننده‌های شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k* نزدیک‌ترین همسایه مورد بررسی و ارزیابی قرار گرفت که به ترتیب: مقادیر ۸۹/۹۱، ۹۷/۵۱، ۷۹/۹۰ و ۸۴/۷۲ درصد به دست آمد. بنابراین با بررسی مقادیر مورد ارزیابی قرار گرفته، می‌توان نتیجه گرفت که روش پیشنهادی نسبت به سایر روش‌ها در تشخیص انواع مختلف *BI-RADS* دقیق‌تر است.

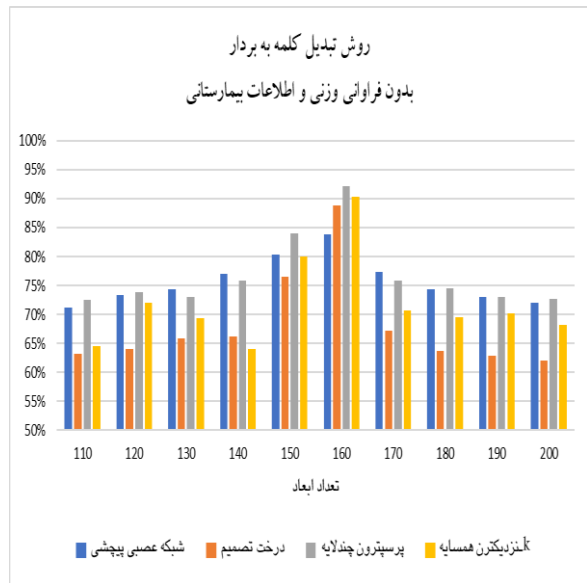
همان‌طور که اشاره شد، با توجه به این که در این تحقیق حالت چند کلاسه مورد بررسی قرار گرفته است، برای به دست آوردن پارامترهای اشاره شده جهت ارزیابی طبقه‌بندی‌کننده‌ها در روش‌های مورد

TFIDF و همچنین استفاده از سابقه پزشکی بیماران، که به صورت خودکار مقدار BI-RADS را پیش بینی می‌کند، نسبت به سایر روش‌ها رقت تشخیص را بهبود بخشیده است.

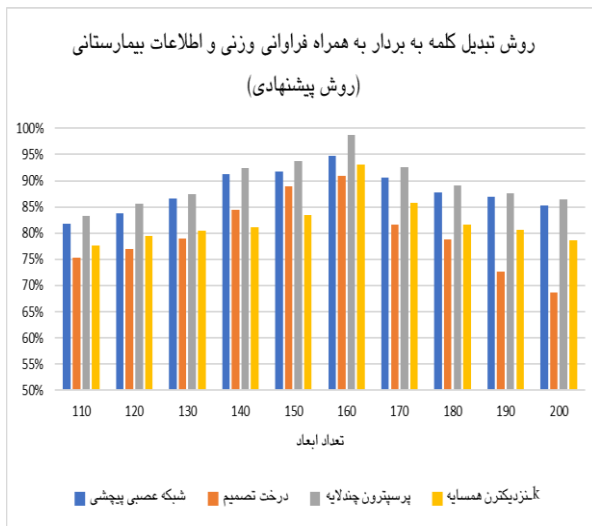
در جدول ۳ نیز رقت روش‌های مشابه دیگر که با روش پیشنهادی مقایسه شده است نمایش داده شده است. همان طور که در این جدول مشخص است روش پیشنهادی یعنی ترکیب تکنیک‌های Word2vec



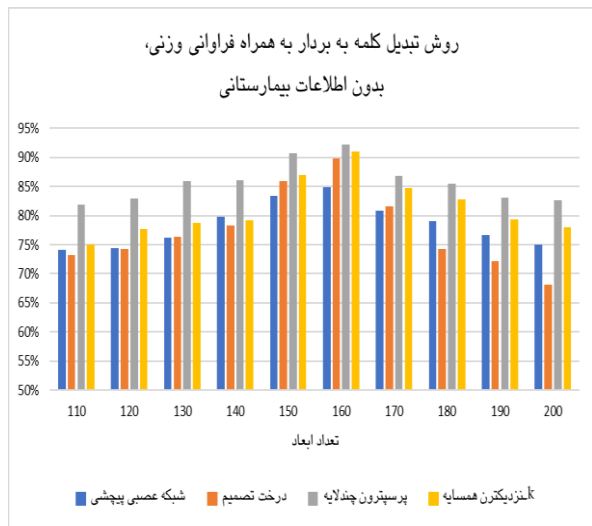
ب) روش تبدیل کلمه به بردار به همراه اطلاعات بیمارستانی، بدون فراوانی وزنی



الف) روش تبدیل کلمه به بردار بدون فراوانی وزنی و اطلاعات بیمارستانی



د) روش تبدیل کلمه به بردار به همراه فراوانی وزنی و اطلاعات بیمارستانی (روش پیشنهادی)



ج) روش تبدیل کلمه به بردار به همراه فراوانی وزنی، بدون اطلاعات بیمارستانی

شکل ۳: مقایسه دقت در طبقه‌بندی‌های مختلف و ابعاد مختلف بردارهای ویژگی

جدول ۱: مقایسه معیارهای ارزیابی در روش‌های بررسی شده در طبقه‌بندی‌کننده‌های مختلف

الف) شبکه عصبی پیچشی

روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	
بدون فراوانی وزنی و اطلاعات بیمارستانی	بدون فراوانی وزنی، همراه اطلاعات بیمارستانی، بدون فراوانی وزنی	بدون اطلاعات بیمارستانی، همراه فراوانی وزنی، بدون اطلاعات بیمارستانی	بدون فراوانی وزنی و اطلاعات بیمارستانی	
۸۳/۷۷	۹۳/۴۹	۸۴/۹۱	۹۴/۷۴	دقت
۳۸/۵۱	۷۵/۴۳	۴۳/۲۰	۷۹/۵۰	حساسیت
۹۰/۱۱	۹۶/۱۳	۹۰/۸۸	۹۶/۸۷	ویژگی
۵۱/۹۹	۸۰/۸۶	۵۲/۹۳	۸۴/۹۰	ارزش پیش بینی شده مثبت
۹۰/۴۶	۹۶/۲۸	۹۱/۲۹	۹۷/۰۵	ارزش پیش بینی شده منفی
۶۰/۱۵	۸۷/۱۱	۶۱/۰۳	۸۹/۹۱	امتیاز f1

ب) درخت تصمیم

روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	
بدون فراوانی وزنی و اطلاعات بیمارستانی	بدون فراوانی وزنی، همراه اطلاعات بیمارستانی، بدون فراوانی وزنی	بدون اطلاعات بیمارستانی، همراه فراوانی وزنی، بدون اطلاعات بیمارستانی	بدون فراوانی وزنی و اطلاعات بیمارستانی	
۸۸/۸۰	۸۹/۳۷	۸۹/۸۳	۹۰/۹۷	دقت
۵۷/۸۸	۶۰/۱۳	۶۱/۰۵	۶۵/۲۸	حساسیت
۹۳/۳۵	۹۳/۷۵	۹۳/۹۵	۹۴/۶۵	ویژگی
۵۹/۸۹	۶۲/۶۷	۶۳/۶۵	۷۰/۷۰	ارزش پیش بینی شده مثبت
۹۳/۴۳	۹۳/۷۴	۹۴/۱۲	۹۴/۸۳	ارزش پیش بینی شده منفی
۷۲/۳۸	۷۳/۹۴	۷۵/۲۱	۷۹/۹۰	امتیاز f1

ج) کنزدیک‌ترین همسایه

روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	
بدون فراوانی وزنی و اطلاعات بیمارستانی	بدون فراوانی وزنی، همراه اطلاعات بیمارستانی، بدون فراوانی وزنی	بدون اطلاعات بیمارستانی، همراه فراوانی وزنی، بدون اطلاعات بیمارستانی	بدون فراوانی وزنی و اطلاعات بیمارستانی	
۹۰/۴۰	۹۲/۱۱	۹۰/۹۷	۹۳/۰۳	دقت
۶۳/۲۵	۷۰/۹۴	۶۵/۴۴	۷۳/۰۳	حساسیت
۹۴/۳۵	۹۵/۳۲	۹۴/۶۷	۹۵/۸۴	ویژگی
۶۶/۳۶	۷۳/۴۴	۷۵/۱۱	۷۶/۵۳	ارزش پیش بینی شده مثبت
۹۴/۴۹	۹۵/۳۹	۹۴/۸۲	۹۶/۰۲	ارزش پیش بینی شده منفی
۷۷/۰۲	۸۲/۶۷	۸۲/۵۵	۸۴/۷۲	امتیاز f1

د) پرسپترون چندلایه

روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	روش تبدیل کلمه به بردار	
بدون فراوانی وزنی و اطلاعات بیمارستانی	بدون فراوانی وزنی، همراه اطلاعات بیمارستانی، بدون فراوانی وزنی	بدون اطلاعات بیمارستانی، همراه فراوانی وزنی، بدون اطلاعات بیمارستانی	بدون فراوانی وزنی و اطلاعات بیمارستانی	
۹۲/۲۳	۹۸/۶۳	۹۲/۲۳	۹۸/۷۴	دقت
۶۹/۳۹	۹۴/۳۵	۶۹/۱۲	۹۴/۷۱	حساسیت
۹۵/۳۷	۹۹/۱۸	۹۵/۳۵	۹۹/۳۵	ویژگی
۷۸/۳۷	۹۵/۶۰	۷۹/۴۸	۹۵/۸۶	ارزش پیش بینی شده مثبت
۹۵/۵۷	۹۹/۲۲	۹۵/۶۰	۹۹/۲۹	ارزش پیش بینی شده منفی
۸۵/۴۸	۹۷/۳۳	۸۶/۱۸	۹۷/۵۱	امتیاز f1

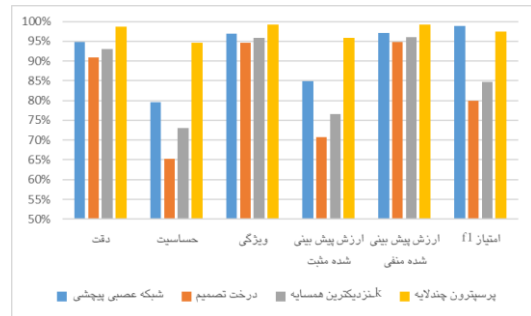
جدول ۲: ماتریس‌های درهم‌ریختگی برای طبقه‌بندی‌کننده‌های مورد نظر در روش پیشنهادی

مقادیر اصلی/واقعی														BI-RADS
VI	V	IV	III	II	I	0	VI	V	IV	III	II	I	0	
درخت تصمیم							شبکه عصبی پیچشی							
۱	۳	۱۰	۰	۰	۲	۹	۰	۰	۰	۰	۰	۱۳	۱۲	0
۰	۵	۱	۰	۰	۱۶	۰	۰	۰	۰	۰	۳	۱۹	۰	I
۴	۵	۶	۰	۲۰	۲	۰	۰	۰	۹	۲۵	۳	۰	۰	II
۰	۱۰	۲	۲۲	۱	۱	۲	۰	۰	۳۷	۱	۰	۰	۰	III
۰	۱	۵۳	۲	۰	۳	۰	۰	۰	۵۷	۲	۰	۰	۰	IV
۰	۳۱	۰	۰	۰	۴	۵	۰	۲۷	۱۳	۰	۰	۰	۰	V
۲۰	۲	۰	۰	۰	۶	۱	۲۷	۱	۱	۰	۰	۰	۰	VI
پرسپترون چندلایه							کنزدیک‌ترین همسایه							
۰	۰	۰	۰	۱	۰	۲۴	۲	۴	۶	۱	۰	۱	۱۱	0
۰	۰	۰	۰	۳	۱۹	۰	۲	۱	۱	۰	۰	۱۷	۱	I
۰	۱	۱	۰	۳۵	۰	۰	۳	۴	۵	۰	۲۴	۱	۰	II
۱	۲	۱	۳۴	۰	۰	۰	۰	۵	۵	۲۴	۱	۱	۲	III
۰	۰	۵۹	۰	۰	۰	۰	۲	۱	۵۵	۰	۰	۱	۰	IV
۰	۴۰	۰	۰	۰	۰	۰	۲	۳۳	۰	۰	۰	۳	۲	V
۲۸	۰	۰	۰	۰	۱	۰	۲۵	۲	۰	۰	۱	۰	۱	VI

مقادیر پیش‌بینی شده

بحث

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 توسعه یافته و سومین عامل مرگ در کشورهای کمتر توسعه یافته است و به تنهایی بیش از بیماری‌های سل، ایدز و مالاریا افراد را به کام مرگ می‌کشاند(۲۷). بر اساس آمار ارایه شده، از هر چهار مورد مرگ بر اثر سرطان در زنان یک نفر با دلیل ابتلاء به سرطان پستان می‌میرد(۲۸)، که بیان‌کننده عاملیت مهم سرطان پستان در مرگ و میر زنان است. همچنین مطابق با آمارهای منتشره به وسیله سازمان جهانی بهداشت، از هر ۸ تا ۱۰ زن، یک نفر به این نوع از سرطان مبتلا



شکل ۴: مقایسه نتایج روش پیشنهادی که با

طبقه‌بندی‌کننده‌های بررسی شده در این تحقیق اجرا شده‌اند

جدول ۳: مقایسه روش پیشنهادی با روش‌های دیگر

نویسنده	دقت (درصد)
کاسترو و همکاران (۲)	۷۴/۰۰
دینگ و همکاران (۱۰)	۹۲/۰۰
ناروائز و همکاران (۱۲)	۹۶/۸۰
دیاب و همکاران (۱۴)	۸۶/۰۵
لیما و همکاران (۱۵)	۹۴/۱۱
روش پیشنهادی	۹۸/۷۴

می‌شود. این آمار برای کشور ایران، احتمال ابتلای یک نفر به ازای هر ۱۰ تا ۱۵ زن را نشان می‌دهد (۱۹). بنابر اعلام نظر محققین، پیشگیری از سرطان پستان به دلیل شناخته نبودن عوامل ابتلاء، تقریباً غیرممکن است، اما یکی از عوامل مهم و اساسی در درمان این نوع سرطان، تشخیص زود هنگام این بیماری بخصوص در مراحل اولیه آن است که می‌تواند احتمال درمان کامل و قطعی را برای بیمار به ارمغان بیاورد. روش‌های تصویر برداری مختلفی وجود دارد که با استفاده از آن پزشک می‌تواند بافت پستان را بررسی نموده و در مورد احتمال وجود انواع توده‌ها نظر خود را بیان نماید. از مهم‌ترین روش‌های مورد استفاده جهت تشخیص سرطان پستان، تحلیل گزارش‌های پزشکی مانند: نتایج ماموگرافی، سونوگرافی، *MRI* و همچنین تعیین طبقه‌بندی زیر مجموعه‌های مولکولی پس از نمونه‌برداری است (۲۹) که هر کدام از این گزارش‌ها می‌توانند دارای سطحی از خطا باشند. به دلیل هزینه‌های بالا در درمان و نیز گسترش این بیماری، شناسایی زود هنگام و به‌موقع این عارضه، گام بسیار مهمی در بهبود میزان سلامتی و همچنین کاهش عوارض اجتماعی جامعه خواهد داشت. جهت تشخیص بیماری سرطان پستان، عوامل و فاکتورهای مختلفی به صورت هم‌زمان برای سرعت بخشی و ارتقای کیفی تشخیص در نظر گرفته می‌شود که دارای جنبه‌های گوناگون و متعددی است و بررسی هم‌زمان همه این عوامل، تصمیم‌گیری پزشک در تشخیص و فرآیند درمان را حساس‌تر و مشکل‌تر

می‌نماید. بسیاری از موارد مربوط به تصمیم‌گیری در پزشکی مدرن بسیار پیچیده است و تعریف مشخص و دلیل منطقی برای تصمیم‌گیری دقیق وجود ندارد، به همین علت تصمیم و نظر پزشکان معمولاً با هم اختلاف دارد. با توجه به این که تصمیم‌گیری در مسائل پزشکی حتی در یک محدوده کوچک، به دلیل حجم بسیار اطلاعات پزشکی مفید و مناسب که با آن محدوده در ارتباط است، معمولاً با عدم قطعیت مواجه می‌شود، به همین دلیل است که تصمیم‌گیری‌های سریع و دقیق را پیچیده می‌کند (۳۰) و کمک به افزایش کیفی تصمیم‌ها، وظیفه‌ای بسیار مهم است که می‌تواند به پزشکان در گرفتن تصمیم دقیق‌تر کمک فراوانی نماید، لذا هدف این مقاله ارایه یک سیستم طبقه‌بندی خودکار با استفاده از روش‌های یادگیری ماشین و تلفیق‌داده می‌باشد.

در این روش با ترکیب *Word2vec* و *TFIDF* می‌توان دقت طبقه‌بندی متن را افزایش داد، اما سابقه پزشکی بیماران در تشخیص بیماری مهم است و می‌تواند دقت را بهبود بخشد. بنابراین، تلفیق *HIS* در کنار *Word2vec* و *TFIDF* با استفاده از مهندسی ویژگی پیشنهاد گردید. به منظور مقایسه روش پیشنهادی، از چهار طبقه‌بندی‌کننده شبکه عصبی پیچشی، پرسپترون چندلایه، درخت تصمیم و *k*-نزدیک‌ترین همسایه برای تشخیص خودکار مقادیر *BI-RADS* استفاده شد. پارامترهای مختلفی از جمله دقت، حساسیت، ویژگی، ارزش پیش‌بینی شده مثبت، ارزش پیش‌بینی شده منفی و امتیاز *f1* برای این طبقه‌بندی‌کننده‌ها مورد ارزیابی

تقدیر و تشکر

نویسندگان این مقاله بر خود لازم دانستند از مسئولین و کارشناسان مرکز آموزشی درمانی بیمارستان شهیدزاده شهرستان بهبهان که در جمع‌آوری مجموعه داده با این مقاله همکاری لازم را داشتند صمیمانه قدردانی نمایند.

تعارض منافع

این مقاله حاوی هیچ مطالعه‌ای به وسیله نویسندگان با شرکت کنندگان انسانی نیست.

حمایت مالی

این تحقیق با حمایت مالی نویسنده اول به انجام رسیده است.

ملاحظات اخلاقی

این مقاله برگرفته از رساله دکتری رشته مهندسی کامپیوتر از دانشگاه آزاد اسلامی، واحد بروجرد با کد اخلاق IR.IAU.B.REC.1401.050 می‌باشد.

مشارکت نویسندگان

تمام نویسندگان در طراحی، اجرا و نگارش همه بخش‌های پژوهش حاضر مشارکت داشته‌اند.

که قرار گرفتند که برای روش پیشنهادی و با طبقه‌بندی‌کننده پرسپترون چند لایه به ترتیب مقادیر ۹۸/۷۴، ۹۴/۷۱، ۹۹/۲۵، ۹۵/۸۶، ۹۹/۲۹ و ۹۷/۵۱ درصد به عنوان بهترین نتیجه حاصل شد.

نتیجه‌گیری

نتایج نشان می‌دهد که همه طبقه‌بندی‌کننده‌ها در روش پیشنهادی نسبت به سایر روش‌ها بهتر و دقیق‌تر عمل کردند، لذا نباید صرفاً بر روی گزارش‌های پزشکی تمرکز کرد و از سایر اطلاعات بالینی و سابقه بیماران نیز باید استفاده نمود. بنابراین استفاده از HIS در کنار گزارش‌های متنی پزشکی می‌تواند دقت در طبقه‌بندی BI-RADS را بهبود ببخشد و تأثیر مثبتی بر تشخیص و کیفیت فرآیندهای درمانی داشته باشد.

در این تحقیق، فقط از بسته‌های استاندارد در پردازش زبان طبیعی استفاده شده است. این روش ممکن است این ظرفیت را داشته باشد که به عنوان مبنایی برای پژوهش‌های آینده باشد که برچسب‌گذاری BI-RADS را برای کمک به رادیولوژیست‌ها و پزشکان به عنوان بخشی از سیستم سلامت یادگیری تقویت می‌کند. همچنین استفاده از زیر گروه‌های مولکولی و استفاده از تصاویر پزشکی در یادگیری عمیق برای بهبود دقت برای کارهای آینده پیشنهاد می‌شود.

REFERENCES

- 1.Redaniel MT, Martin RM, Ridd MJ, Wade J, Jeffreys M. Diagnostic intervals and its association with breast, prostate, lung and colorectal cancer survival in England: historical cohort study using the Clinical Practice Research Datalink. *PLoS ONE* 2015; 1(5): e0126608.
- 2.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.
- 3.Mendonca SC, Abel GA, Saunders CL, Wardle J, Lyratzopoulos G. Pre-referral general practitioner consultations and subsequent experience of cancer care: evidence from the english cancer patient experience survey. *European Journal of Cancer Care* 2016; 25(3): 478-90.
- 4.Gao F, Yoon H, Wu T, Chu X. A feature transfer enabled multi-task deep learning model on medical imaging. *Expert Systems with Applications* 2020; 143: 112957.
- 5.Gong J, Bai X, Li D-a, Zhao J, Li X. Prognosis Analysis of Heart Failure Based on Recurrent Attention Model. *Innovation and Research in BioMedical engineering (IRBM)* 2019; 41(2): 71-9.
- 6.Yang M, Kiang M, Shang W. Filtering big data from social media—Building an early warning system for adverse drug reactions. *Journal of biomedical informatics*. 2015; 54: 230-40.
- 7.Genes N, Chandra D, Ellis S, Baumlin K. Validating emergency department vital signs using a data quality engine for data warehouse. *The Open Medical Informatics Journal* 2013; 7: 34-9.
- 8.Sun P, Wang L, Xia Q. The Keyword Extraction of Chinese Medical Web Page Based on WF-TF-IDF Algorithm. *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE 2017; 193-198.
- 9.Dreisbach C, Koleck TA, Bourne PE, Bakken S. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics* 2019; 125: 37-46.
- 10.Ding X, Zhang X. Research on text structuralization in medical field. *2nd International Conference on Cloud Computing and Internet of Things (CCIOT)*: IEEE; 2016; 155-61.
- 11.Nii M, Tuchida Y, Iwamoto T, Uchinuno A, Sakashita R. Nursing-care text evaluation using word vector representations realized by word2vec. *International Conference on Fuzzy Systems (FUZZ-IEEE)*; 2016: IEEE; 2165-2169.
- 12.Narváez F, Díaz G, Poveda C, Romero E. An automatic BI-RADS description of mammographic masses by fusing multiresolution features. *Expert Systems with Applications* 2017; 74: 82-95.
- 13.Østerås BH, Martinsen ACT, Brandal SHB, Chaudhry KN, Eben E, Haakenaasen U, et al. BI-RADS density classification from areometric and volumetric automatic breast density measurements. *Academic Radiology* 2016; 23(4): 468-78.
- 14.Diab DM, El Hindi KM. Using differential evolution for fine tuning naïve Bayesian classifiers and its application for text classification. *Applied Soft Computing* 2017; 54: 183-99.
- 15.de Lima SML, da Silva-Filho AG, Dos Santos WP. Detection and classification of masses in mammographic images in a multi-kernel approach. *Computer Methods and Programs in Biomedicine* 2016; 134: 11-29.
- 16.Moayedi F, Azimifar Z, Boostani R, Katebi S. Contourlet-based mammography mass classification using the SVM family. *Computers in Biology and Medicine* 2010; 40(4): 373-83.
- 17.Eltoukhy MM, Faye I, Samir BB. A statistical based feature extraction method for breast cancer diagnosis in digital mammogram using multiresolution representation. *Computers in Biology and Medicine* 2012; 42(1): 123-8.
- 18.Reyad YA, Berbar MA, Hussain M. Comparison of Statistical, LBP, and Multi-Resolution Analysis Features for Breast Mass Classification. *Journal of medical systems* 2014; 38: 100.
- 19.Isfahani P, Hossieni Zare SM, Shamsaii M. The Prevalence of Depression in Iranian Women With Breast Cancer: A Meta-Analysis. *Journal of Internal Medicine Today* 2020; 26(2): 170-181.
- 20.Bouvry C, Tvardik N, Kergourlay I, Bittar A, Arnod-Prin P, Segond F, et al. The SYNODOS Project: System for the Normalization and Organization of Textual Medical Data for Observation in Healthcare. *Innovation and Research in BioMedical engineering (IRBM)* 2016; 37(2): 109-15.
- 21.Yang L, Liu B, Lin H, Lin Y. Combining local and global information for product feature extraction in opinion documents. *Information Processing Letters* 2016; 116(10): 623-7.
- 22.Kim D, Seo D, Cho S, Kang P. Multi-co-training for document classification using various document representations: TF-IDF, LDA, and Doc2Vec. *Information Sciences* 2019; 477: 15-29.
- 23.Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences. *arXiv:1404.2188* 2014; 655-65.

24. Jais IKM, Ismail AR, Nisa SQ. Adam optimization algorithm for wide and deep neural network. Knowledge Engineering and Data Science 2019; 2(1): 41-6.
25. Tharwat A. Classification assessment methods. Applied Computing and Informatics 2020; 17(1): 168-92.
26. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) 2014; 1532-1543.
27. Balakumar P, Maung-U K, Jagadeesh G. Prevalence and prevention of cardiovascular disease and diabetes mellitus. Pharmacological Research 2016; 113: 600-9.
28. 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(6): 394-424.
29. Siu AL, US Preventive Services Task Force. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine 2016; 164(4): 279-96.
30. Shahmoradi L, Farzanehnejad AR. Guideline-based clinical decision support systems as an inseparable tool for better cancer care management. Iranian journal of public health 2016; 45(7): 962.

Automatic Classification of BI-RADS in Mammography Reports Using Data Fusion

Zahabi M¹, Shiri ME^{2,1*}, Haj Seyed Javadi H³, Broumandzadeh M⁴

¹Department of Computer Engineering, Borujerd Branch, Islamic Azad University, Borujerd, Iran, ²Department of Mathematics and Computer Science, Amir-Kabir University of Technology, Tehran, Iran, ³Department of Mathematics and Computer Science, Shahed University, Tehran, Iran, ⁴Department of Computer Engineering and Information Technology, Payam-e-Nour University, Tehran, Iran

Received: 08 Nov 2023 Accepted: 26 Feb 2024

Abstract

Background & aim: Breast cancer is one of the most common cancers in women and the main cause of death in cancer diseases, and mammography is the primary imaging method for early detection of breast masses. Rapid diagnosis with high accuracy is one of the serious concerns of doctors and healthcare centers when facing certain diseases, so the purpose of this article was to determine the automatic classification of BI-RADS in mammography reports using data fusion.

Methods: The present descriptive, analytical, and retrospective study was conducted in 2023, the mammography report and the electronic file of the patients were extracted from the archiving and communication system of the patient's image and records obtained from the available information in the medical training center of Shahidzadeh hospital in Behbahan, Iran, which includes the mammography reports and the electronic record of 250 patients who had ample information. To model the proposed method using the collected dataset, Python software was used in the Visual Studio Code environment. Finally, cross-validation was used to evaluate the quality and validity of the results.

Results: The results confirmed that the proposed approach, namely the use of Word2vec combined with TFIDF, and their integration with HIS, had a significant impact on the accuracy of medical text classification. The output vectors of Word2vec were used for BI-RADS level classification when TFIDF was applied or not applied, as well as with and without the integration of HIS, for classifiers such as CNN, MLP, DT, and k-NN, and the results were compared using evaluation measures such as accuracy, precision, sensitivity, positive predictive value, negative predictive value, and F1 score. The results indicated that the best accuracy with the proposed method using the multilayer perceptron classifier was 98.74%, but without HIS, the accuracy for the same classifier was 92.23%.

Conclusion: By combining Word2vec with TFIDF, the accuracy of text classification could be increased, but the medical history of patients was important in the diagnosis of disease and could improve the accuracy. The results indicated that one should not focus only on medical reports and other clinical information and patients' history should also be used. Therefore, the use of HIS along with medical text reports could improve BI-RADS classification and have a positive effect on diagnosis and treatment processes.

Keywords: medical text classification, breast cancer, feature extraction, BI-RADS, HIS

***Corresponding author:** Shiri ME, Department of Mathematics and Computer Science, Amir-Kabir University of Technology, Tehran, Iran.

Email: shiri@aut.ac.ir

Please cite this article as follows: Zahabi M, Shiri ME, Haj Seyed Javadi H, Broumandzadeh M. Automatic Classification of BI-RADS in Mammography Reports Using Data Fusion. *Armaghane-danesh* 2024; 29(3): 365-385.

The scientific research journal *Armaghan Danesh*, affiliated with Yasuj University of Medical Sciences, is an open-access publication. All articles published in this journal