مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره هفتم شماره دوم. ص ۱۱۲-۱۰۲، ۱۳۹۹

# یک مدل طبقه‌بندی ترکیبی برای تشخیص سرطان پستان با استفاده از پشته تعمیم یافته

**ماهیار عشایری[۱]*، امین رضایی‌پناه[۲]**

• **دریافت مقاله:** ۹۸/۱/۲۲ • **پذیرش مقاله:** ۹۸/۵/۱۳

**مقدمه:** سرطان پستان یکی از رایج‌ترین انواع سرطان‌ها است و رشد قابل ملاحظه‌ای از آن در سال‌های اخیر گزارش شده است. به منظور تشخیص این بیماری، پارامترهای زیادی باید بررسی گردد که خطاهای انسانی و یا عوامل محیطی امکان اشتباه را ممکن می‌کند. به همین دلیل در چند دهه اخیر از هوش مصنوعی برای تشخیص این بیماری در جهت کمک به پزشکان استفاده می‌شود.

**روش:** در این مطالعه توصیفی-کاربردی، تشخیص بیماری سرطان پستان با استفاده از پشته تعمیم یافته در قالب یک مدل ترکیبی مبتنی بر سه روش شبکه عصبی MLP، درخت تصمیم ID3 و ماشین بردار پشتیبان ارائه شد. برای بهبود عملکرد مدل طبقه‌بندی ترکیبی از یک رویکرد جدید تحت عنوان بلاک جداکننده استفاده شد. این بلاک وظیفه تشخیص نمونه‌هایی را دارد که باعث ایجاد خطا در مدل طبقه‌بندی می‌شوند.

**نتایج:** به منظور ارزیابی دقت روش پیشنهادی از پایگاه داده ویسکانسین مرتبط با بیماری سرطان پستان استفاده شد. نتایج آزمایش‌ها برتری روش پیشنهادی را در مقابل سایر روش‌های مشابه نشان داد. دقت مدل طبقه‌بندی ارائه شده روی مجموعه داده‌های WBCD، WDBC و WPBC از پایگاه داده ویسکانسین به ترتیب ۹۹/۵۴٪، ۹۹/۵۸٪ و ۹۹/۸۴٪ بود.

**نتیجه‌گیری:** با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوین و با صرفه‌تری در نظام سلامت و درمان ارائه کرد که با دقت بالایی قادر به تشخیص سرطان پستان باشند. در این تحقیق ضمن تشخیص بیماری به کمک روش‌های داده‌کاوی، توانست با استفاده از تکنیک پشته تعمیم یافته به دقت بالایی در تشخیص بیماری دست یابد.

**کلید واژه‌ها:** پشته تعمیم یافته، طبقه‌بندی داده‌ها، پایگاه داده ویسکانسین، داده کاوی، سرطان پستان

**•ارجاع:** عشایری ماهیار، رضایی‌پناه امین. یک مدل طبقه‌بندی ترکیبی برای تشخیص سرطان پستان با استفاده از پشته تعمیم‌یافته. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۹؛ ۷(۲): ۱۲-۱۰۲.

۱. کارشناسی ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، واحد بوشهر، دانشگاه آزاد اسلامی، بوشهر، ایران
۲. کارشناسی ارشد مهندسی کامپیوتر، گروه مهندسی کامپیوتر، مؤسسه آموزش عالی رهجویان دانش برازجان، بوشهر، ایران

*** نویسنده مسئول:** ماهیار عشایری
**آدرس:** بوشهر، عالیشهر، دانشگاه آزاد اسلامی، واحد بوشهر
**•شماره تماس:** ۰۹۱۷۳۷۷۵۶۴۸
• **Email:** m.ashayeri1988@gmail.com

## مقدمه

سرطان پستان یکی از سرطان‌های شایع زنان می‌باشد. احتمال ابتلاء به سرطان پستان با بالارفتن سن افزایش می‌یابد، به طوری که ۸۰ درصد موارد بعد از سن ۵۰ سالگی اتفاق می‌افتد [۱]. علی‌رغم این‌که میزان شیوع سرطان پستان در زنان آمریکایی بیشتر از زنان آسیایی است؛ ولی مرگ‌ومیر ناشی از آن در کشورهای آسیایی بیشتر است که علت آن تشخیص دیرهنگام بیماری است [۲]. علت اصلی این تشخیص دیرهنگام، نداشتن آگاهی از بیماری و بدون علامت بودن بیماری است. اگرچه سرطان پستان اخیراً از جنبه‌های مختلفی موضوع تحقیقات گسترده در اغلب مراکز پژوهشی سرطان در جهان قرار گرفته، ولی با این حال تحقیقات همچنان ادامه دارد. در ایران سرطان پستان اولین نوع سرطان تشخیص داده شده در میان زنان است که ۲۴/۴٪ از همه انواع بدخیمی‌ها را به خود اختصاص می‌دهد. از این رو ارائه مدل‌های تشخیصی و یاری رساندن به پزشکان در ارائه تصمیمات مناسب برای درمان این بیماری اهمیت دارد و لزوم انجام تحقیقات مرتبط است [۴، ۳].

با توجه به زمان‌گیر بودن فرآیند تشخیص توسط افراد کارشناس، استفاده از سیستم‌های کامپیوتری می‌تواند به سرعت بخشیدن و کاهش قابل‌توجه حجم کار عامل انسانی منجر شود. در این راستا در سال‌های اخیر کوشش‌های بسیاری در این زمینه صورت گرفته است که مهم‌ترین آن‌ها شامل بررسی تصاویر پزشکی و آسان‌سازی فرآیند تشخیص برای پزشک به کمک کامپیوتر و نیز تشخیص مستقیم به وسیله تکنیک‌های داده‌کاوی بر اساس فاکتورهای مؤثر در این زمینه می‌باشد. در سال‌های اخیر استفاده از تکنیک‌های داده‌کاوی به منظور تشخیص بیماری‌ها، مورد توجه بسیاری از محققین قرار گرفته است [۳].

به‌طورکلی روش‌هایی که در این حیطه ارائه شده‌اند از دو شیوه برای تشخیص سرطان پستان استفاده می‌کنند. روش اول مبتنی بر پردازش تصویر است که با استفاده از تصاویر ماموگرافی و یا ترموگرافی اقدام به تشخیص سرطان می‌شود [۵، ۴]. روش دوم استفاده از مدل‌های داده‌کاوی نظیر طبقه‌بندی برای تشخیص سرطان مبتنی بر پایگاه‌های داده‌ای است که با تست FNA(Fine Needle Aspiration) تولید می‌شوند [۷، ۶]. Nilashi و همکاران، سیستمی مبتنی بر دانش برای تشخیص سرطان پستان با استفاده از روش منطق فازی توسعه دادند [۸]. آن‌ها از روش حداکثر انتظار برای خوشه‌بندی داده‌ها و قرار دادن آن‌ها در گروه‌های مشابه استفاده کردند، سپس از درخت رگرسیون CART برای قوانین فازی در هر خوشه بهره گرفتند.

Diz و همکاران، یک رویکرد مبتنی بر داده‌کاوی برای غده‌شناسی در سرطان پستان ارائه دادند [۹]. این روش با استفاده از طبقه‌بندی چگالی پستان به شناسایی تمایزات در مجموعه داده‌ها کمک می‌کند از الگوریتم بیزین برای کار طبقه‌بندی بهره می‌گیرد. Devi و همکارش، یک الگوریتم سه مرحله‌ای را برای تشخیص زودهنگام سرطان پستان ارائه دادند [۱۰]. در مرحله اول داده‌ها با استفاده از الگوریتم خوشه‌بندی Farthest First گروه‌بندی می‌شوند. در مرحله دوم از الگوریتم ODA (Outlier Detection Algorithm) برای طبقه‌بندی و در مرحله سوم، شناسایی خوش‌خیم یا بدخیم سرطان با پیش‌پردازش مجموعه داده با استفاده از الگوریتم طبقه‌بندی J48 انجام می‌شود. نتایج دقت ۹۹/۹٪ را برای مجموعه داده (Wisconsin Breast Cancer Diagnosis) نشان می‌دهد. Vaidehi و همکاران، تشخیص سرطان پستان را با استفاده از طبقه‌بندی KNN ترکیبی ارائه دادند [۱۱]. آن‌ها از ترکیب سه ماتریس فاصله متفاوت یوکلیدان، کوزین و CITY-BLOCK برای طبقه‌بندی استفاده کردند. Onan، یک مدل طبقه‌بندی نزدیک‌ترین مجاور گنگ – سخت ترکیبی را با زیرمجموعه سازگار و گزینش موردی برای تشخیص اتوماتیک سرطان پستان مورد مطالعه قرار دادند [۱۲].

Shen و همکاران، یک مدل هوشمند پیش‌بینی سرطان پستان با استفاده از تکنیک‌های داده‌کاوی ارائه دادند [۱۳]. آن‌ها برای انتخاب ویژگی‌های مؤثر در تشخیص سرطان پستان از روش خصیصه انتخابی و برای کار طبقه‌بندی از SVM استفاده کردند. Aalaei و همکاران، انتخاب ویژگی را با استفاده از الگوریتم ژنتیک برای تشخیص سرطان پستان مطرح کردند و از مرتب‌سازی PS برای بهبود دقت بهره گرفتند [۱۴]. در تحقیقی مشابه Rao و همکاران، مطالعه‌ای در مورد تشخیص DNA در بیماران مبتلا به سرطان پستان با استفاده از الگوریتم ژنتیک ارائه دادند [۱۵]. Ahmad و همکاران، یک مدل تشخیص سرطان پستان مبتنی بر الگوریتم ژنتیک و شبکه عصبی (Network Artificial Neural) ANN ارائه دادند [۱۶] و از الگوریتم ژنتیک به طور هم‌زمان

به منظور انتخاب ویژگی و بهینه‌سازی پارامترهای شبکه عصبی ANN استفاده شده است. نتایج دقت بهترین ۹۹/۲۴٪ و میانگین ۹۸/۲۹٪ را در مجموعه داده WBCD نشان می‌دهد. Kabir و همکاران، مدل (Constructive Approach Feature Selection)CAFS را بر مبنای شبکه عصبی ANN ارائه دادند [۱۷]. CAFS به طور خودکار به تعیین تعداد نودهای لایه مخفی در طول فرآیند انتخاب ویژگی می‌پردازد.

این پژوهش بر اساس مجموعه داده‌های استاندارد از مخزن UCI(University of California, Irvine) که با استفاده از تست FNA ایجاد شده است، جهت تشخیص سرطان پستان ارائه گردید [۱۸]. در عین حال در این تحقیق به بررسی مهم‌ترین روش‌های ارائه شده در این زمینه، یعنی روش‌های شبکه عصبی [۱۹]، ماشین بردار پشتیبان [۲۰] و درخت تصمیم [۲۱] پرداخته و ضمن مقایسه این روش‌ها با یکدیگر، روش پشته تعمیم‌یافته را در قالب یک مدل طبقه‌بندی ترکیبی ارائه داد. هدف نهایی این تحقیق ترکیب مدل‌های طبقه‌بندی مختلف بر اساس تکنیک پشته تعمیم‌یافته به منظور بهبود تشخیص سرطان پستان بود.

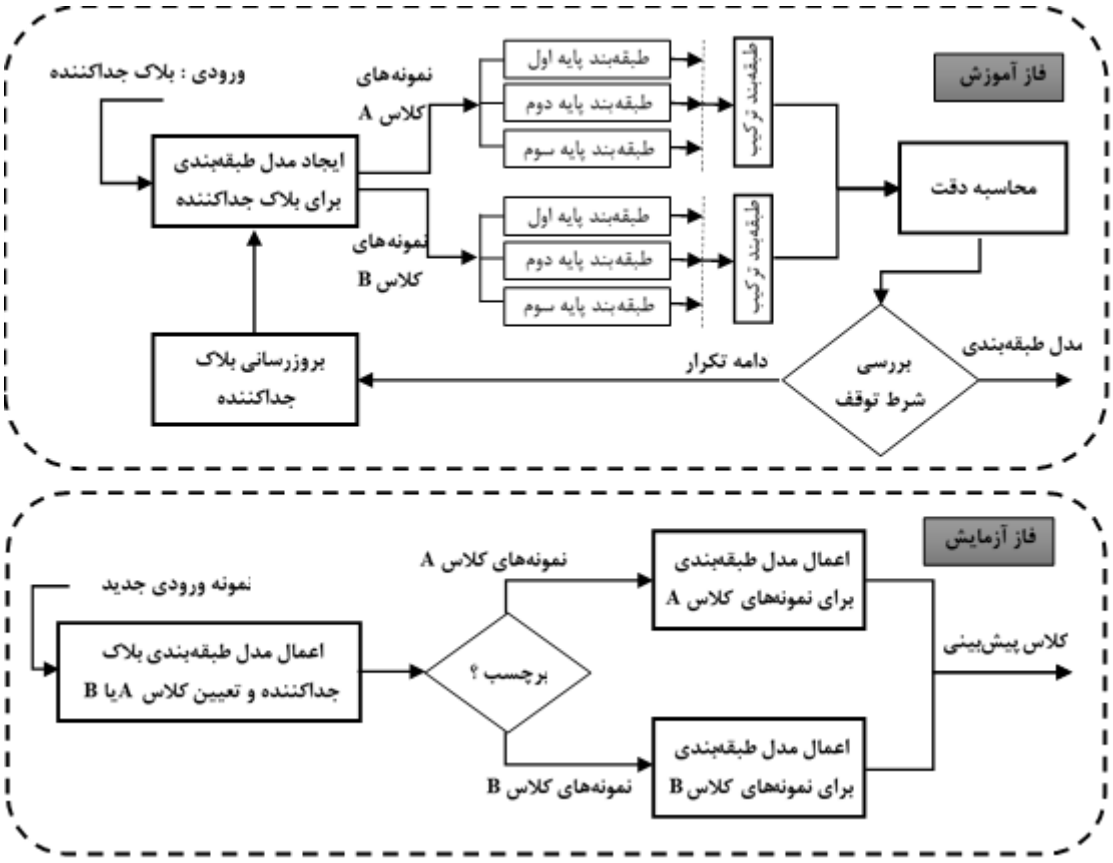
## روش

در این مطالعه توصیفی-کاربردی، برای بهبود تشخیص و پیش‌بینی سرطان پستان از ترکیب سه الگوریتم شبکه عصبی NN (Neural Network)، ماشین بردار پشتیبان SVM (Support Vector Machine) و درخت تصمیم DT (Decision Tree) بر مبنای تکنیک پشته تعمیم یافته [۲۲] استفاده شد. مفهوم پشته تعمیم‌یافته در حیطه داده‌کاوی، پیش‌بینی دقیق‌تر بر اساس ترکیب چند مدل طبقه‌بندی است. روش پیشنهادی ترکیبی بوده و از چندین مرحله تشکیل می‌شود.

در ابتدا داده‌های سرطان پستان در فاز آموزشی پیش‌پردازش شده و پس از آن یک بلاک جداکننده به منظور تقسیم‌بندی نمونه‌ها به k کلاس (برچسب) جدید طراحی شد. در این تحقیق k= ۲ در نظر گرفته شد، از این رو بلاک جداکننده به هر نمونه کلاس جدید A یا B را تخصیص می‌دهد. کلاس تخصیص داده شده به صورت مجازی می‌باشد و با کلاس اصلی نمونه‌ها (سرطان دارد یا سرطان ندارد) متفاوت است. در واقع نمونه‌های مجموعه داده به دو گروه تقسیم شده که نمونه‌های گروه اول مربوط به کلاس A و نمونه‌های گروه دوم مربوط به کلاس B هستند. به منظور جداسازی اولیه و تخصیص کلاس مجازی به نمونه‌ها از الگوریتم خوشه‌بندی k-means استفاده شد. در مرحله بعد یک مدل طبقه‌بندی برای نمونه‌های بلاک جداکننده با کلاس جدید A و B ایجاد شد. این مدل تصمیم می‌گیرد که در مرحله بعد نمونه‌ها چگونه مدل شوند. نمونه‌های خروجی بلاک جداکننده با کلاس A و B در بخش‌های مستقل و متفاوتی بر مبنای تکنیک پشته تعمیم‌یافته آموزش داده می‌شوند. از آن یک مدل طبقه‌بندی ترکیبی بر مبنای پشته تعمیم‌یافته ایجاد می‌شود. مدل طبقه‌بندی ارائه شده بر اساس ترکیب سه الگوریتم شبکه عصبی، ماشین بردار پشتیبان و درخت تصمیم ایجاد می‌شود. روش ترکیب بر مبنای پشته تعمیم‌یافته است و هر یک از الگوریتم‌های طبقه‌بندی نقش متفاوتی در مدل طبقه‌بندی نهایی ایفا می‌کنند.

مدل طبقه‌بندی مرتبط با بلاک جداکننده باعث ایجاد دو گروه نمونه با کلاس A و B شده که برای هر یک از گروه‌ها یک مدل طبقه‌بندی متفاوت طراحی می‌شود. نمونه‌های موجود در هر یک از گروه‌ها می‌توانند شامل هر دو کلاس اصلی مجموعه داده (سرطان دارد یا سرطان ندارد) باشند؛ بنابراین در ساختار ارائه شده سه مدل متفاوت ایجاد شده که دقت هر سه مدل برای ارزیابی نهایی محاسبه می‌شود. در نهایت بلاک جداکننده به منظور بهبود گروه‌بندی نمونه‌ها به دو کلاس A و B در یک فرایند تکراری به‌روزرسانی می‌شود.

برای تعیین کلاس نمونه ورودی در فاز آزمایش، ابتدا کلاس مجازی A یا B توسط مدل طبقه‌بندی مرتبط با بلاک جداکننده تشخیص داده می‌شود. سپس با تعیین کلاس مجازی نمونه ورودی، جهت‌دهی نمونه به سمت یکی از مدل‌های طبقه‌بندی (گروه A یا گروه B) انجام شده و کلاس نهایی نمونه با مدل طبقه‌بندی مربوطه پیش‌بینی می‌شود. خروجی این مدل طبقه‌بندی در واقع کلاس خروجی تشخیص داده شده برای نمونه ورودی است و از آن در محاسبه دقت نهایی استفاده می‌شود. شکل ۱ روند جریان روش پیشنهادی را نشان می‌دهد.




**شکل ۱: روند جریان روش پیشنهادی**

جزئیات هر بخش از روش پیشنهادی

**پیش‌پردازی داده‌ها:** پیش‌پردازش داده‌ها به منظور آماده‌سازی و بهبود کیفیت داده‌های اولیه انجام می‌شود. در ابتدا فیلدهایی از مجموعه داده نظیر شناسه بیمار که به دانش نهفته در اطلاعات مرتبط نمی‌باشد، حذف گردید. همچنین نمونه‌هایی که دارای مقادیر گم شده [۲۳] هستند (با نماد «؟» مشخص شده‌اند) نیز از مجموعه داده حذف شدند. علاوه بر این، از روش Z-score برای نرمال‌سازی استفاده شده است [۲۴]. Z-score روی مرکزیت صفر قرار می‌دهد و میانگین و انحراف معیار نمونه‌ها را برای هر ویژگی به ترتیب ۰ و ۱ می‌کند. رابطه (۱) نرمال‌سازی Z-score را تعریف می‌کند.

$$e_{i,j}^{z-score} = \frac{e_{i,j} - \mu_j}{\sigma_j} \qquad (۱)$$

جایی که $e_{i,j}$ مقدار ویژگی $i$ را برای نمونه $j$ نشان می‌دهد. $\mu_j$ و $\sigma_j$ به ترتیب میانگین و انحراف معیار مقادیر ویژگی $i$ برای تمام نمونه‌ها می‌باشد.

**بلاک جداکننده:** در این تحقیق از یک بلاک جداکننده به منظور تقسیم‌بندی نمونه‌ها به $k$ کلاس مجازی (در این تحقیق $k = 2$) استفاده شد. در واقع نمونه‌های مجموعه داده به دو گروه تقسیم شده که نمونه‌های گروه اول مربوط به کلاس A و نمونه‌های گروه دوم مربوط به کلاس B هستند. فرض کنید یک مجموعه داده با n نمونه روی یک روش طبقه‌بندی انفرادی اعمال شود. در اکثر مواقع مدل طبقه‌بندی موفق به تشخیص و آموزش صحیح برخی از نمونه‌ها نمی‌شود. دلایل ایجاد خطا در طبقه‌بندی‌ها را می‌توان به صورت زیر توصیف کرد.

– مقادیر برخی از ویژگی‌ها در بعضی نمونه‌ها ممکن است در حالات بسیار نادر و خاصی رخ داده شده باشند که در مقایسه با سایر نمونه‌ها مقادیر متفاوتی هستند.

– ممکن است تعداد برخی از نمونه‌ها با مقادیر ویژگی خاص برای یک کلاس در داده آموزشی بسیار کم باشد.

– شرایط تقسیم‌بندی داده‌های آموزش و آزمایش ممکن است منجر به کاهش فرکانس یک یا چند کلاس خاص برای نمونه‌ها شود.

– برخی از نمونه‌ها نویز هستند و در واقعیت نیز مشاهده نمی‌شوند.

تکنیک بلاک جداکننده پیشنهادی به منظور آموزش دو مدل طبقه‌بندی به جای یک مدل ارائه شد. در این تحقیق به نمونه‌های ورودی، در داده‌های آموزشی دو برچسب A و B به‌صورت مجازی اختصاص داده شد. این برچسب‌ها بدون توجه به برچسب‌های واقعی نمونه‌ها می‌باشند. وظیفه بلاک جداکننده تعمیم نمونه‌های آموزشی به دو کلاس A و B است به طوری که هر برچسب با یک مدل مجزا طبقه‌بندی شود. تشخیص و نحوه جداسازی نمونه‌ها به دو برچسب A و B چالش این تحقیق است که در یک فرآیند تکراری بهینه‌سازی می‌شود. به منظور جداسازی اولیه و تخصیص برچسب مجازی به نمونه‌ها از الگوریتم خوشه‌بندی K-means [۲۵] با سه خوشه استفاده شد. به نمونه‌های موجود در خوشه‌هایی با تعداد اعضاء بیشتر کلاس A و B و به نمونه‌های موجود در خوشه با کمترین اعضاء هر دو برچسب A و B تخصیص داده شد. دلیل استفاده از این روش جداسازی حفظ شباهت بین اعضاء دو کلاس A و B و همچنین عدم کاهش نمونه در هر یک از برچسب‌ها و ایجاد اشتراک بین آن‌ها است. در واقع نمونه‌هایی که با یک روش طبقه‌بندی مستقل به راحتی مدل شده و باعث افزایش کارایی می‌شوند در هر دو گروه A و B قرار می‌گیرند. علاوه بر این، نمونه‌هایی که بنا به دلایل مختلف تشخیص صحیح کلاس آن‌ها با استفاده از یک طبقه‌بندی مستقل ممکن نمی‌باشد را جداسازی کرده و بین دو گروه تقسیم می‌شوند.

برای مدل کردن این جداسازی از یک مدل طبقه‌بندی انفرادی استفاده شد. بررسی‌های انجام شده نشان داد که الگوریتم شبکه عصبی عملکرد بهتری در این بخش داشته است، از این رو مدل طبقه‌بندی مرتبط با بلاک جداکننده شبکه عصبی می‌باشد. بهینگی این مدل نقش مؤثری در افزایش کارایی مدل طبقه‌بندی نهایی دارد، لذا نحوه جداسازی و تخصیص برچسب‌های مجازی باید بهینه‌سازی شود. علاوه بر این، نمونه‌های خروجی بلاک جداکننده با کلاس A و B در بخش‌های مستقل و متفاوتی بر مبنای تکنیک پشته تعمیم‌یافته آموزش داده می‌شوند. از این رو برای هر گروه از نمونه‌ها (نمونه‌ها با کلاس A و نمونه‌ها با کلاس B) یک مدل طبقه‌بندی ایجاد می‌شود.

**آموزش مدل طبقه‌بندی با بهره‌گیری از تکنیک پشته تعمیم یافته:** در پشته تعمیم یافته خروجی هر یک از طبقه بندها، یک ورودی برای یادگیر سطح بالاتر در دیگر طبقه‌بندی‌ها می‌باشد که مشخص می‌کند به چه خوبی آن‌ها را ترکیب نموده است. در این تحقیق یک روش ابتکاری بر مبنای پشته تعمیم‌یافته برای ترکیب سه طبقه‌بند شبکه عصبی، ماشین بردار پشتیبان و درخت تصمیم ارائه شد. ساختار الگوریتم پشته تعمیم‌یافته برای ترکیب مدل‌های طبقه‌بندی از دولایه تشکیل شده است. در لایه اول، طبقه‌بندهایی قرار دارند که روی داده‌های مجموعه یادگیری آموزش می‌بینند. پس از اتمام آموزش خروجی طبقه‌بندهای لایه اول به ازای تمام داده‌های مجموعه یادگیری جمع‌آوری شده و در مجموعه داده جدید قرار می‌گیرند. این مجموعه به عنوان ورودی به طبقه‌بند لایه دوم (فراطبقه‌بند) داده می‌شود. سپس فراطبقه‌بند لایه دوم نگاشت میان خروجی‌های هر یک از طبقه‌بندهای معمولی لایه اول را با کلاس‌های خروجی واقعی یاد می‌گیرد. با توجه به تکنیک پشته تعمیم یافته، مجموعه داده جدید ایجاد شده شامل کلاس خروجی هر یک از طبقه‌بندی‌های لایه اول و همچنین ویژگی‌های مجموعه داده اصلی می‌باشد. در لایه اول برای هر گروه از نمونه‌ها به طور مشابه سه طبقه‌بندی درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی وجود دارد. الگوریتم‌های درخت تصمیم و ماشین بردار پشتیبان به ترتیب برای نمونه‌های کلاس A و B در لایه دوم استفاده می‌شود، این دو الگوریتم بهترین عملکرد را در ایجاد مدل این بخش‌ها داشته‌اند. در این بخش کلاس مجازی از نمونه‌ها حذف شده و کلاس واقعی هر نمونه برای ایجاد مدل استفاده می‌شود.

**دقت مدل طبقه‌بندی ترکیبی:** در فرآیند تکرار و به‌روزرسانی بلاک جداکننده، دقت کل مدل طبقه‌بندی پیشنهادی به صورت رابطه (۲) محاسبه می‌شود.

$$TrainAccuracy = Average(C_1 + C_2 + C_3) \quad (۲)$$

در این رابطه $C_1$ دقت طبقه‌بندی بلاک جداکننده، $C_2$ دقت خروجی مدل برای نمونه‌های کلاس A و $C_3$ دقت خروجی مدل برای نمونه‌های کلاس B است. اگر دقت محاسبه شده در طی $\Gamma$ تکرار متوالی تغییر نکند، فرآیند اجرای الگوریتم متوقف می‌شود. در اینجا $\Gamma = 35$ در نظر گرفته شده است.

**فرایند به‌روزرسانی بلاک جداکننده:** با توجه به هدف الگوریتم در بهینه‌سازی بلاک جداکننده و اهمیت جداسازی نمونه‌ها، نیاز به تکرار و به‌روزرسانی بلاک جداکننده می‌باشد. در این تحقیق از یک الگوریتم تپه نوردی برای جستجوی بهینه‌ترین بلاک جداکننده (مناسب‌ترین جداسازی نمونه‌ها) استفاده شد. در هر تکرار از الگوریتم تپه‌نوردی، بلاک جداکننده با هدف بهبود دقت طبقه‌بندی به‌روزرسانی می‌شود. در اینجا

$DS_A$ به نمونه‌های آموزشی کلاس A و $DS_B$ به نمونه‌های آموزشی کلاس B اشاره دارد. به‌روزرسانی بلاک جداکننده به صورت زیر انجام می‌شود.

- هر نمونه از $DS_A$ که به اشتباه طبقه‌بندی شده است، به احتمال $\alpha$ به مجموعه داده $DS_B$ انتقال پیدا می‌کند و به احتمال $1-\alpha$ با یک نمونه از $DS_B$ که به اشتباه طبقه‌بندی شده است، جابه‌جا می‌شود (در اینجا $\alpha = 0.5$ در نظر گرفته شده است).
- هر نمونه از $DS_B$ که به اشتباه طبقه‌بندی شده است، به احتمال $\alpha$ به $DS_A$ انتقال پیدا می‌کند و به احتمال $1-\alpha$ با یک نمونه از $DS_A$ که به اشتباه طبقه‌بندی شده است، جابه‌جا می‌شود.
- اگر در یکی از دو گروه $DS_A$ و $DS_B$ نمونه‌ای با کلاس‌بندی اشتباه یافت نشد، همواره احتمال $\alpha$ برقرار است و جابه‌جایی انجام می‌شود.

**تعیین کلاس نمونه ورودی در فاز آزمایش:** برای پیش‌بینی کلاس نمونه ورودی جدید در فاز آزمایش، ابتدا نمونه ورودی با استفاده از مدل طبقه‌بندی بلاک جداکننده ارزیابی شده و کلاس A یا B آن مشخص می‌شود. کلاس نمونه پیش‌بینی شده نوع طبقه‌بندی جهت تعیین کلاس واقعی را تعیین می‌کند. در اینجا نمونه‌های کلاس A از مدل فراطبقه‌بند درخت تصمیم و نمونه‌های کلاس B از مدل فراطبقه‌بند ماشین بردار پشتیبان جهت پیش‌بینی سرطان استفاده می‌کنند. دقت نهایی داده‌های آموزشی از رابطه (۳) حاصل می‌شود.

$$TestAccuracy = \frac{1}{N}\sum_{i=1}^{N}\rho, \quad \rho = \begin{cases} 1 & c_i = t_i \\ 0 & c_i \neq t_i \end{cases} \qquad (۳)$$

در این رابطه $N$ تعداد نمونه‌های آزمایش، $c_i$ کلاس پیش‌بینی‌شده نمونه ورودی و $t_i$ کلاس واقعی نمونه ورودی می‌باشد.

## نتایج

برای انجام شبیه‌سازی و تجزیه‌وتحلیل روش پیشنهاد شده از نرم‌افزار Matlab نسخه ۲۰۱۷ روی مجموعه داده‌های WBCD (۶۹۹ نمونه و ۹ ویژگی)، WDBC (۵۶۹ نمونه و ۳۱ ویژگی) و WPBC (۱۹۸ نمونه و ۳۲ ویژگی) از پایگاه داده ویسکانسین [۲۶] استفاده شده است. همه ویژگی‌ها به صورت مقادیر عدد صحیح هستند. علاوه بر این، همه نمونه‌ها دارای دو کلاس «سرطان پستان دارد» و «سرطان پستان ندارد» می‌باشند. جدول ۱ چند نمونه از رکوردهای مجموعه داده WBCD را با ۹ ویژگی نشان می‌دهد.

**جدول ۱: چند نمونه از رکوردهای مجموعه داده WBCD**

| شناسه بیمار | ضخامت توده | یکنواختی اندازه سلول | یکنواختی شکل سلول | چسبندگی حاشیه‌ای | اندازه سلول‌های اپیتلیال | هسته خالی | کروماتین مطلوب | هستک نرمال | تقسیم میتوز | برچسب کلاس |
|---|---|---|---|---|---|---|---|---|---|---|
| ۱۰۰۰۰۲۵ | ۵ | ۱ | ۱ | ۱ | ۲ | ۱ | ۳ | ۱ | ۱ | ۲ |
| ۱۰۰۲۹۴۵ | ۵ | ۴ | ۴ | ۵ | ۷ | ۱۰ | ۳ | ۲ | ۱ | ۲ |
| ۱۰۱۵۴۲۵ | ۳ | ۱ | ۱ | ۱ | ۲ | ۲ | ۳ | ۱ | ۱ | ۲ |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |

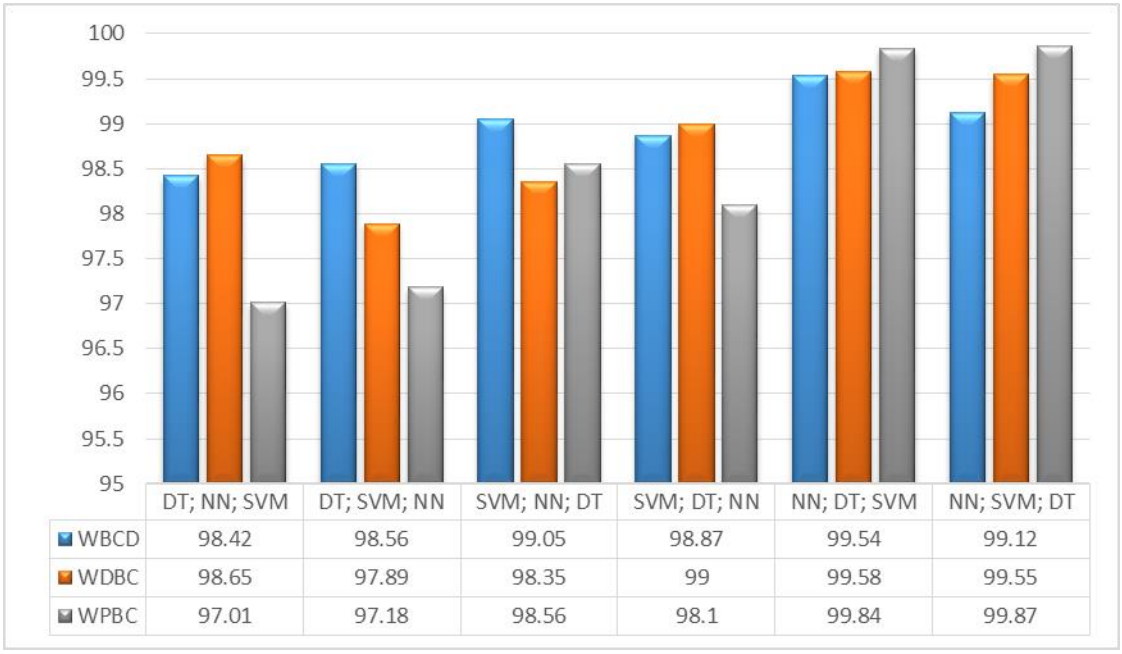
در این تحقیق از تکنیک اعتبارسنجی 10-Fold برای ارزیابی مدل استفاده می‌شود [۲۷]. در هر مرحله از اعتبارسنجی، مجموعه داده اصلی به دو بخش $E^T$ (آموزش) و $E^P$ (آزمایش) تقسیم شده، جایی که ۹۰٪ نمونه‌ها برای $E^T$ و ۱۰٪ دیگر برای $E^P$ استفاده می‌شود. به منظور حصول اطمینان از نتایج ارائه شده میانگین ۲۰ بار اجرای الگوریتم در تمام آزمایش‌ها محاسبه شده است. جدول ۲ نتایج روش پیشنهادی را برای سه مجموعه داده WBCD، WDBC و WPBC نشان می‌دهد.

جدول ۲: نتایج روش پیشنهادی برای سه مجموعه داده WBCD، WDBC و WPBC

| مجموعه داده | دقت مدل بلاک جداکننده (٪) | دقت مدل برچسب A (٪) | دقت مدل برچسب B (٪) | دقت روش پیشنهادی (٪) میانگین | بهترین |
|---|---|---|---|---|---|
| WBCD | ۹۷/۹۳ | ۹۸/۹۰ | ۹۷/۰۰ | ۹۹/۵۴ | ۹۹/۸۷ |
| WDBC | ۹۹/۰۴ | ۹۹/۳۴ | ۹۹/۸۳ | ۹۹/۵۸ | ۹۹/۹۳ |
| WPBC | ۹۴/۶۵ | ۹۷/۳۸ | ۹۸/۰۵ | ۹۹/۸۴ | ۹۹/۹۵ |

در روش پیشنهادی برای ساخت مدل بلاک جداکننده از طبقه‌بندی شبکه عصبی (NN)، برای مدل کردن نمونه‌های برچسب A از طبقه‌بندی درخت تصمیم (DT) و برای مدل کردن نمونه‌های برچسب B از طبقه‌بندی ماشین بردار پشتیبان (SVM) بهره گرفته شد. بررسی صحت انتخاب این الگوریتم‌ها برای بخش‌های مختلف در شکل ۲ انجام شده است. نتایج این بررسی نشان می‌دهد که ترکیب NN, DT, SVM بهترین عملکرد را برای مدل طبقه‌بندی دارد.

در ادامه نتایج روش پیشنهادی با تعدادی از روش‌های طبقه‌بندی کلاسیک مقایسه شد. روش‌های طبقه‌بندی مورد مقایسه شبکه عصبی (Multi-Layer Perceptron) MLP، درخت تصمیم ID3 و ماشین بردار پشتیبان هستند. نتایج این مقایسه برای سه مجموعه داده WBCD، WDBC و WPBC در جدول ۳ گزارش شده است.




شکل ۲: نتایج استفاده از سه طبقه‌بند مستقل در بخش‌های مختلف روش پیشنهادی

جدول ۳: نتایج مقایسه دقت تشخیص سرطان پستان در روش پیشنهادی و سه طبقه‌بند MLP، ID3 و SVM

| مجموعه داده | شبکه عصبی MLP | درخت تصمیم ID3 | ماشین بردار پشتیبان | روش پیشنهادی |
|---|---|---|---|---|
| WBCD | ۸۹/۳۳ | ۹۲/۸۵ | ۹۳/۵۸ | ۹۹/۵۴ |
| WDBC | ۵/۶۰ | ۹۳/۱۷ | ۹۰/۲۳ | ۹۹/۵۸ |
| WPBC | ۸۸/۱۰ | ۹۰/۷۰ | ۹۱/۴۷ | ۹۹/۸۴ |

برای بررسی عملکرد روش پیشنهادی، نتایج این تحقیق با تعدادی از جدیدترین روش‌های تشخیص سرطان پستان که از پایگاه داده ویسکانسین در آزمایش‌های خود بهره گرفته‌اند، مقایسه شده است. جدول ۴ نتایج این مقایسه را نشان می‌دهد.

**جدول ٤: مقایسه دقت تشخیص سرطان پستان مدل پیشنهادی با سایر روش‌ها مشابه روی پایگاه داده ویسکانسین**

| مجموعه داده | EM-Fuzzy, 2017 [۸] | ODA, 2016 [۱۰] | RS-BPNN, 2015 [۲۸] | SMO-IBK, 2012 [۲۹] | روش پیشنهادی |
|---|---|---|---|---|---|
| WBCD | ۹۳/۲۰ | ۹۹/۹۰ | ۹۷/۳۰ | ۹۷/۲۸ | ۹۹/۵۴ |
| WDBC | ۹۴/۱۰ | ۹۹/۶۰ | ۹۸/۶۰ | ۹۷/۰۱ | ۹۹/۵۸ |
| WPBC | - | - | ۹۰/۴ | ۹۴/۲۲ | ۹۹/۸۴ |

## بحث و نتیجه‌گیری

با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوین و با صرفه‌تری در نظام سلامت و درمان ارائه کرد که با دقت بالایی قادر به تشخیص سرطان پستان باشند. تحقیق حاضر با هدف تشخیص سرطان خوش‌خیم و بدخیم پستان با استفاده از یک مدل طبقه‌بندی ترکیبی دو لایه مبتنی بر تکنیک پشته تعمیم یافته انجام شد. تشخیص و جداسازی نمونه‌هایی که باعث ایجاد خطا می‌شوند، با استفاده از یک بلاک جداکننده در راستای افزایش دقت طبقه‌بندی رویکرد جدیدی در روش پیشنهادی است. نتایج این بررسی به وضوح نشان از اثربخشی مدل طبقه‌بندی ترکیبی و تکنیک پشته تعمیم یافته در تشخیص سرطان پستان را دارد؛ بنابراین استفاده از این روش می‌تواند در کنار سایر روش‌های تشخیصی غیرتهاجمی، به عنوان یک سیستم پشتیبان تشخیص با دقت بالا برای تشخیص این بیماری مورد استفاده قرار گیرد. علاوه بر این، استفاده از این روش می‌تواند موجب کاهش آسیب‌های احتمالی روش‌های تهاجمی و عمل‌های غیرضروری شود و دقت تشخیص سرطان پستان را بهبود بخشد.

نتایج حاصل از آزمایش‌ها نشان می‌دهد که استفاده از شبکه عصبی (NN) برای مدل کردن کلاس‌های مجازی بلاک جداکننده بهترین دقت را فراهم می‌کند. این دقت برای مجموعه داده WBCD حدود ۹۸٪ و برای مجموعه داده‌های WDCB و WPBC حدود ۹۹٪ و ۹۵٪ است. با توجه به این که تقسیم‌بندی صحیح نمونه‌ها باعث کاهش خطا در مدل طبقه‌بندی نهایی خواهد شد؛ لذا کارایی مدل طبقه‌بندی این بخش از اهمیت بالایی برخوردار می‌باشد. به همین دلیل است که شبکه‌های عصبی مصنوعی به عنوان روش نوین در تشخیص بیماری‌ها مورد توجه بسیاری از محققین در سال‌های اخیر قرار گرفته است. علاوه بر این برای گروه نمونه‌هایی با کلاس‌های مجازی A و B به ترتیب الگوریتم‌های طبقه‌بندی درخت تصمیم (DT) و ماشین بردار پشتیبان (SVM) بهترین دقت‌ها را گزارش می‌دهند.

به‌طورکلی نتایج مقایسه در جدول ۳ نشان می‌دهد که روش پیشنهادی در مقایسه با سایر مدل‌های طبقه‌بندی کلاسیک برتری قابل‌توجهی دارد. تجربه نشان داده است که ترکیب پیشگویی‌های انجام شده توسط چند روش معمولاً پیشگویی‌های دقیق‌تری را نسبت به مدل‌های انفرادی حاصل می‌کند. برتری روش پیشنهادی به طور میانگین در مجموعه داده WBCD نسبت به شبکه عصبی MLP حدود ۱۱٪ است. برای مجموعه داده‌های WDCB و WPBC این برتری حدود ۱۶٪ و ۱۳٪ است. همچنین برتری روش پیشنهادی در مقایسه با درخت تصمیم ID3 روی مجموعه داده‌های WBCD، WDCB و WPBC به ترتیب ۷٪، ۷٪ و ۱۰٪ می‌باشد. علاوه بر این، به طور مشابه روش پیشنهادی برتری ۷٪، ۱۰٪ و ۹٪ را نسبت به روش SVM گزارش می‌دهد.

روش پیشنهادی در مقایسه با سایر روش‌های مشابه نیز نتایج قابل قبولی را ارائه داده است. با توجه به جدول ۴ روش پیشنهادی پس از الگوریتم (Outlier Detection ODA(Algorithm در رتبه دوم قرار دارد. ODA یک الگوریتم سه مرحله‌ای است که از گروه‌بندی ویژگی‌ها و ترکیب الگوریتم‌های ODA و J48 برای شناسایی خوش‌خیم یا بدخیم بودن نمونه‌های سرطانی استفاده می‌کند. دلیل برتری این الگوریتم استفاده از فرآیند انتخاب ویژگی‌های مؤثر می‌باشد

که با این کار افزونگی داده‌ها را در مدل طبقه‌بندی کاهش می‌دهد. این روش انتخاب ویژگی‌ها را بر اساس میزان همبستگی انجام داده و بر گروه ویژگی که دارای همبستگی کمتر است تأکید دارد.

نتایج همچنین نشان می‌دهد که روش پیشنهادی نسبت به الگوریتم‌های EM-Fuzzy، RS-BPNN و SMO-IBK و به ازای همه مجموعه داده‌ها برتری دارد. الگوریتم EM-Fuzzy از یک سیستم مبتنی بر فازی برای تشخیص سرطان پستان استفاده می‌کند. در این روش نیز ویژگی‌ها توسط EM گروه‌بندی شده و برای کار طبقه‌بندی و ایجاد پایگاه قوانین از درخت رگرسیون CART استفاده شده است. برتری روش پیشنهادی نسبت به این روش در مجموعه داده‌های WBCD و WDBC به ترتیب حدود ۷٪ و ۶٪ است. الگوریتم RS-BPNN از رابطه همبستگی ناهنجار برای انتخاب ویژگی‌ها و برای کار طبقه‌بندی از یک شبکه عصبی استفاده می‌کند. این روش عملکرد مناسبی در مدل کردن سایر مجموعه داده‌های بالینی نظیر هپاتیت و بیماری قلبی نیز داشته است. برتری روش پیشنهادی نسبت به این روش در مجموعه داده‌های WBCD، WDBC و WPBC به ترتیب حدود ۲٪، ۱٪ و ۱۰٪ است. در نهایت الگوریتم SMO-IBK در تحقیق Salama و همکاران [۲۹] با توجه به مقایسه تجربی مدل‌های مختلف طبقه‌بندی برای تشخیص سرطان پستان عملکرد بهتری نشان داده است. در این روش IBK بر پایه k نزدیک‌ترین همسایه است و در یک مدل ترکیبی با روش SMO(Sequential Minimal Optimization) تلفیق می‌شود. نتایج جدول ۴ برتری روش پیشنهادی را نسبت به این الگوریتم نیز برای همه مجموعه داده‌ها نشان می‌دهد. این برتری برای مجموعه داده WBCD حدود ۲٪ است و برای مجموعه داده‌های WDCB و WPBC حدود ۳٪ و ۶٪ می‌باشد.

روش ارائه شده با توجه به تخصیص کلاس جدید و هدایت مسیر نمونه‌ها به سمت طبقه‌بندی با ارزیابی دقیق‌تر توانسته است دقت بهتری نسبت به روش‌های طبقه‌بندی کلاسیک گزارش دهد. به‌طورکلی روش تشخیص سرطان پیشنهادی در مقایسه با سایر روش‌های مورد بررسی و به ازای برخی از پایگاه‌های داده دقت بیشتری داشته و در بقیه موارد نیز دقت مناسبی را ارائه می‌دهد.

روش پیشنهادی بر مبنای بلاک جداکننده، نمونه‌هایی که باعث ایجاد خطای طبقه‌بندی می‌شوند را شناسایی کرده و با ایجاد مدل‌های طبقه‌بندی متفاوت سعی در کاهش این خطا دارد؛ بنابراین این روش برای مجموعه داده‌هایی که احتمال وجود نویز و نمونه‌های با مقادیر خاص در آن‌ها بیشتر است عملکرد بهتری خواهد داشت. با توجه به عدم بررسی وجود نویز و پالایش آن‌ها در این تحقیق، پیشنهاد می‌شود این مورد در تحقیقات آینده در نظر گرفته شود. از جمله محدودیت‌های این تحقیق، مقادیر زیاد داده‌های از دست رفته به دلیل حذف نمونه‌ها در بخش پیش‌پردازش است. از دیگر پیشنهادهای این تحقیق اعمال سیاست‌های بهتری نظیر میانگین گرفتن در برخورد با چنین نمونه‌هایی می‌باشد. علاوه بر این، اگرچه به نظر می‌رسد روش پیشنهادی عملکرد خوبی در پایگاه داده ویسکانسین دارد؛ اما هیچ بینش کیفی وجود ندارد که برای سایر مجموعه داده‌ها نیز این عملکرد مناسب باشد. از این رو پیشنهاد می‌شود عملکرد این روش برای سایر مجموعه داده‌ها نیز بررسی شود. یکی دیگر از محدودیت‌های این تحقیق، پیچیدگی محاسباتی نسبتاً بالایی است که این روش برای جداسازی نمونه‌ها دارد. این مسئله برای مجموعه داده‌هایی با تعداد نمونه‌های زیاد بحرانی است و پیشنهاد می‌شود برای به‌روزرسانی بلاک جداکننده از الگوریتم‌های اکتشافی نظیر ژنتیک استفاده شود. همچنین با توجه به این که کاهش ابعاد داده‌های واقعی بر عملکرد مدل‌های طبقه‌بندی مؤثر است، پیشنهاد می‌شود تأثیر انواع روش‌های انتخاب ویژگی نیز مورد بررسی قرار گیرد.

## تعارض منافع

بدین‌وسیله نویسندگان تصریح می‌نمایند که هیچ‌گونه تضاد منافعی در خصوص پژوهش حاضر وجود ندارد.

**References**

**1.** King MC, Marks JH, Mandell JB, New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2Science 2003;302(5645):643-6. doi: 10.1126/science.1088759

**2.** Siegel RL, Miller KD, Jemal A. Cancer statistics, 2015. CA Cancer J Clin 2015;65(1):5-29. doi: 10.3322/caac.21254

**3.** Asghar U, Witkiewicz AK, Turner NC, Knudsen ES. The history and future of targeting cyclin-dependent

kinases in cancer therapy. Nat Rev Drug Discov 2015;14(2):130-46. doi: 10.1038/nrd4504
**4.** Sadeghi B. Prediction of MicroRNAs (miRNAs) Targets in Breast Cancer Using Bioinformatics Methods. Journal of Health and Biomedical Informatics 2016;3(1):18-28. [In Persian]
**5.** Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 2003;72(5):1117-30. doi: 10.1086/375033
**6.** Orooji A, Langarizadeh M. Evaluation of the Effect of Feature Selection and Different kernel Functions on SVM Performance for Breast Cancer Diagnosis. Journal of Health and Biomedical Informatics 2018;5(2):244-51. [In Persian]
**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(4):266-78. [In Persian]
**8.** Nilashi M, Ibrahim O, Ahmadi H, Shahmoradi L. A knowledge-based system for breast cancer classification using fuzzy logic method. Telematics and Informatics 2017;34(4):133-44. https://doi.org/10.1016/j.tele.2017.01.007
**9.** Diz J, Marreiros G, Freitas A. Applying data mining techniques to improve breast cancer diagnosis. J Med Syst 2016;40(9):203. doi: 10.1007/s10916-016-0561-y
**10.** Devi RD, Devi MI. Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. Int J Adv Engg Tech 2016; 7(2):93:93-8.
**11.** Vaidehi K, Subashini TS. Breast tissue characterization using combined KNN classifier. Indian Journal of Science and Technology 2015;8(1):23-6. doi: 10.17485/ijst/2015/v8i1/52818
**12.** Onan A. A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer. Expert Systems with Applications 2015; 42(20):6844-52.
https://doi.org/10.1016/j.eswa.2015.05.006
**13.** Shen R, Yang Y, Shao F. Intelligent breast cancer prediction model using data mining techniques. Sixth International Conference on Intelligent Human-Machine Systems and Cybernetics; 2014 Aug 26-27; Hangzhou, China: IEEE; doi: 10.1109/IHMSC.2014.100
**14.** Aalaei S, Shahraki H, Rowhanimanesh A, Eslami S. Feature selection using genetic algorithm for breast cancer diagnosis: experiment on three different datasets. Iranian Journal of Basic Medical Sciences. 2016;19(5):476-82. doi: 10.22038/IJBMS.2016.6931
**15.** Rao R, Rivers A, Rahimi A, Wooldridge R, Rao M, Leitch M, Euhus D, Haley BB. Genetic Ancestry using Mitochondrial DNA in patients with Triple-negative breast cancer (GAMiT study). Cancer 2017;123(1):107-13. doi: 10.1002/cncr.30267
**16.** Ahmad F, Isa NA, Hussain Z, Osman MK, Sulaiman SN. A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. Pattern Analysis and Applications 2015;18(4):861-70. doi: 10.1007/s10044-014-0375-9
**17.** Kabir MM, Islam MM, Murase K. A new wrapper feature selection approach using neural network. Neurocomputing 2010;73(16-18):3273-83. https://doi.org/10.1016/j.neucom.2010.04.003
**18.** Asuncion A, Newman DJ. UCI machine learning repository, 2007.
**19.** Robert C, Guilpin C, Limoge A. Review of neural network applications in sleep research. J Neurosci Methods 1998;79(2):187-93. doi: 10.1016/s01650270(97)00178-7
**20.** Vapnik V, Izmailov R. Knowledge transfer in SVM and neural networks. Annals of Mathematics and Artificial Intelligence 2017;81(1-2):3-19. doi: 10.1007/s10472-017-9538-x
**21.** Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics 1991;21(3):660-74. doi: 10.1109/21.97458
**22.** Wolpert DH. Stacked generalization. Neural Networks 1992;5(2):241-59. https://doi.org/10.1016/S0893-6080(05)80023-1
**23.** Chaurasia V, Pal S. Data mining techniques: to predict and resolve breast cancer survivability. International Journal of Computer Science and Mobile Computing 2014;3(1):10-22.
**24.** Cheadle C, Vawter MP, Freed WJ, Becker KG. Analysis of microarray data using Z score transformation. J Mol Diagn 2003;5(2):73-81. doi: 10.1016/S1525-1578(10)60455-2
**25.** Wagstaff K, Cardie C, Rogers S, Schrödl S. Constrained k-means clustering with background knowledge. Proceedings of the Eighteenth International Conference on Machine Learning; 2001 Jun- Jul 28 -1; Williams College, Williamstown, MA, USA: p. 577-84.
**26.** Center for Machine Learning and Intelligent Systems. Breast Cancer Wisconsin (Original) Data Set; UCI: 1992.
**27.** Bengio Y, Grandvalet Y. No unbiased estimator of the variance of k-fold cross-validation. Journal of Machine Learning Research 2004;5:1089-105.
**28.** Nahato KB, Harichandran KN, Arputharaj K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Comput Math Methods Med 2015;2015:460189. doi: 10.1155/2015/460189
**29.** Salama GI, Abdelhalim MB, Zeid MA. Experimental comparison of classifiers for breast cancer diagnosis. In2012 Seventh International Conference on Computer Engineering & Systems; 2012 Nov 27-29; Cairo, Egypt: IEEE; 2012. p. 180-5. doi: 10.1109/ICCES.2012.6408508

Original Article

*Journal of Health and Biomedical Informatics*
*Medical Informatics Research Center*
*2020; 7(2): 102-112*

# An Ensemble Classification Model for the Diagnosis of Breast Cancer Using Stacked Generalization

**Ashayeri Mahyar[1]*, Rezaeipanah Amin[2]**

• **Received:** 11 Apr 2019 • **Accepted:** 04 Aug 2019

**Introduction:** Breast cancer is one of the most common types of cancer whose incidence has increased dramatically in recent years. In order to diagnose this disease, many parameters must be taken into consideration and mistakes are possible due to human errors or environmental factors. For this reason, in recent decades, Artificial Intelligence has been used by medical practitioners to diagnose this disease.

**Method:** In this applied-descriptive study, the diagnosis of breast cancer using stacked generalization was presented in the form of an ensemble model based on MLP neural network, ID3 decision tree, and support vector machine methods. To improve the performance of the ensemble classification model, a new approach called separator block was used. This block is responsible for identifying instances that cause errors in the classification model.

**Results:** In order to evaluate the accuracy of the proposed method, the Wisconsin database for breast cancer was used. The experimental results showed the superiority of the proposed method over other similar methods. The accuracy of the classification model presented on the WBCD, WDBC, and WPBC datasets from the Wisconsin database was 99.54%, 99.58% and 99.84%, respectively.

**Conclusion**: Data mining algorithms can provide new and more cost-effective systems in the field of health and treatment that can diagnose breast cancer with high accuracy. In this study, modeling based on the stacked generalization technique was of high accuracy in the diagnosis of breast cancer.

**Keywords:** Stacked Generalization, Data Classification, Wisconsin Database, Data Mining, Breast Cancer

• **Citation:** Ashayeri M, Rezaeipanah A. An Ensemble Classification Model for the Diagnosis of Breast Cancer Using Stacked Generalization. Journal of Health and Biomedical Informatics 2020; 7(2): 102-12. [In Persian]

1. M.Sc. in Computer Engineering, Computer Engineering Dept., Bushehr Branch, Islamic Azad University, Bushehr, Iran
2. M.Sc. in Computer Engineering, Computer Engineering, Dept., University of Rahjuyan Danesh Borazjan, Bushehr, Iran

***Corresponding Author:** Mahyar Ashayeri
**Address:** Islamic Azad University, Bushehr Branch, Alishahr, Bushehr, Iran
• **Tel:** 09173775648 • **Email:** m.ashayeri1988@gmail.com