یک رویکرد نوین در تشخیص سرطان سینه با استفاده از پشته تعمیم

نویسندگان: ماهیار عشایری- m.ashayeri1988@gmail.com
امین رضایی پناه- amin.rezaeipanah@gmail.com

چکیده: سرطان سینه یکی از رایج‌ترین انواع سرطان‌ها است و رشد قابل‌ملاحظه‌ای از آن در سال‌های اخیر گزارش‌شده است. به‌منظور 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. در این تحقیق فرایند تشخیص این بیماری با استفاده از پشته تعمیم‌یافته در یک مدل ترکیبی مبتنی بر سه طبقه‌بندی شبکه عصبی MLP، درخت تصمیم ID3 و ماشین بردار پشتیبان ارائه می‌شود. برای بهبود عملکرد مدل طبقه‌بندی ترکیبی از یک روش جدید تحت عنوان بلاک جداکننده برای تشخیص نمونه‌هایی که باعث ایجاد خطا می‌شوند، استفاده می‌شود. به‌منظور ارزیابی دقت روش پیشنهادی از پایگاه داده ویسکانسین استفاده‌شده است. نتایج آزمایش‌ها برتری روش پیشنهادی را در مقابل سایر روش‌ها مشابه نشان می‌دهد. دقت مدل طبقه‌بندی ارائه‌شده روی مجموعه داده‌های سرطان سینه WBCD، WDBC و WPBC از پایگاه داده ویسکانسین به ترتیب ۹۹,۵۴٪، ۹۹,۵۸٪ و ۹۹,۸۴٪ می‌باشد.

کلیدواژه: پشته تعمیم‌یافته، طبقه‌بندی داده‌ها، پایگاه داده ویسکانسین، تشخیص سرطان سینه.

۱ - مقدمه

سرطان سینه یکی از سرطان‌های شایع زنان می‌باشد. احتمال ابتلا به سرطان سینه با بالارفتن سن افزایش می‌یابد، به‌طوری‌که ۸۰ درصد موارد بعد از سن ۵۰ سالگی اتفاق می‌افتد [۱]. علی‌رغم اینکه میزان شیوع سرطان سینه 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.

درصورتی‌که سرطان در مراحل اولیه و تومور کوچک‌تر از نه سانتیمتر و قبل از متاستاز به سایر نقاط شناسایی شود، گزینه‌های درمانی خیلی بیشتری وجود دارد و میزان مرگ‌ومیر به میزان قابل‌توجهی کاهش می‌یابد [۳]. بر این اساس و با توجه به ناشناخته بودن علت این بیماری تنها عامل مؤثر در روند درمان، تشخیص به‌موقع و شروع هرچه سریع‌تر درمان است. با توجه به زمان‌گیر بودن فرآیند تشخیص توسط افراد کارشناس، استفاده از سیستم‌های کامپیوتری می‌تواند به‌سرعت 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 به‌وسیله کامپیوتر بر اساس فاکتورهای مؤثر در این زمینه می‌باشد [۴].

این پژوهش بر اساس مجموعه داده‌های استاندارد از مخزن UCI که با استفاده از تست FNA ایجادشده است، جهت تشخیص سرطان سینه ارائه می‌گردد [۵]. درعین‌حال در این تحقیق به بررسی مهم‌ترین روش‌های ارائه‌شده در این زمینه، یعنی روش‌های شبکه عصبی [۶]، ماشین بردار پشتیبان [۷] و درخت تصمیم [۸] پرداخته و ضمن مقایسه این روش‌ها با یکدیگر، روش پشته تعمیم‌یافته را در قالب یک مدل طبقه‌بندی ترکیبی ارائه می‌دهیم.

در ادامه این تحقیق به بررسی برخی از جدیدترین کارهای انجام‌شده در بخش ۲ می‌پردازیم، در بخش ۳ مدل پیشنهادی مبتنی بر طبقه‌بندی ترکیبی و پشته تعمیم‌یافته به‌منظور تشخیص سرطان سینه به همراه یک بلاک جداکننده برای بهبود دقت طبقه‌بندی ارائه می‌گردد. نتایج حاصل از ارزیابی روش پیشنهادی در بخش ۴ آورده شده و درنهایت نتیجه‌گیری و پیشنهادها در بخش ۵ بیان‌شده است.

www.kahrobaonline.ir

۲- مروری بر کارهای انجام‌شده

در این بخش تلاش داریم برخی از روش‌های مختلف ارائه‌شده در زمینه تشخیص سرطان سینه را بررسی کنیم. به‌طورکلی روش‌هایی که در این حیطه ارائه‌شده‌اند از دو شیوه برای تشخیص سرطان سینه استفاده می‌کنند. روش اول مبتنی بر پردازش تصویر است که با استفاده از تصاویر ماموگرافی و یا ترموگرافی اقدام به تشخیص سرطان می‌شود [۹]. روش دوم استفاده از مدل‌های داده‌کاوی نظیر طبقه‌بندی برای تشخیص سرطان مبتنی بر پایگاه‌های داده‌ای است که با تست FNA تولید می‌شوند [۱۰، ۱۱].

غلامی و میرعابدینی، استفاده هم‌زمان از همبستگی خطی پیرسون و ترکیب الگوریتم‌های داده‌کاوی به‌منظور بهبود پیش‌بینی نوع تومور در بیماران سرطانی را مطرح کردند [۱۲]. آن‌ها به‌منظور بهبود کارایی مدل تشخیصی از الگوریتم بیز ساده و آدابوست بهره گرفتند و به‌دقت ۹۶,۶۷٪ دست یافتند. شیخ‌پور و همکاران، از الگوریتم بهینه‌سازی ذرات به‌منظور تعیین پهنای باند و انتخاب ویژگی در تخمین چگالی کرنل مبتنی بر طبقه‌بندی، جهت تشخیص سرطان سینه استفاده کردند [۱۳]. مدل PSO-KDE معرفی‌شده دو الگوریتم PSO و KDE را به‌منظور تشخیص سرطان سینه به چالش می‌کشد.

بوداغی، یک روش داده‌کاوی از طریق سیستم‌های خبره را جهت پیش‌بینی و تشخیص سرطان سینه و مقایسه آن با دیگر روش‌های داده‌کاوی توسعه داد و با ۹۶۴ قانون به‌دقت ۹۶,۵٪ رسیدند [۱۴]. همچنین شمس‌الکتابی و همکاران، تشخیص و شناسایی الگوهای خوش‌خیم و بدخیم در سرطان سینه به کمک شبکه عصبی LVQ پیشنهاد دادند [۱۵]. در تحقیقی دیگر موسوی و رفیعی، تشخیص سرطان سینه را با استفاده از دسته‌بندی فازی مبتنی بر الگوریتم مورچه‌مزرح ساختند و به‌دقت ۹۸,۲۸٪ رسیدند [۱۶]. غلامی و فرضی، تشخیص سرطان سینه را با استفاده از یادگیری عمیق مطرح کردند [۱۷]. آن‌ها از شبکه عصبی عمیق کانولوشن برای پیکربندی و آموزش داده‌ها بر مبنای یک روش یادگیری سلسله مراتبی بهره گرفتند.

نیلاشی و همکاران، سیستمی مبتنی بر دانش برای تشخیص سرطان سینه با استفاده از روش منطق فازی توسعه دادند [۱۸]. 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. دیز و همکاران، یک رویکرد مبتنی بر داده‌کاوی برای غده شناسی در سرطان سینه ارائه دادند [۱۹]. این روش با استفاده از طبقه‌بندی چگالی سینه به شناسایی تمایزات در مجموعه داده‌ها کمک می‌کند از الگوریتم بیزین برای کار طبقه‌بندی بهره می‌گیرد. دوی و همکاران، یک الگوریتم سه مرحله‌ای را برای تشخیص زودهنگام سرطان سینه ارائه دادند [۲۰]. در مرحله اول داده‌ها با استفاده از الگوریتم خوشه‌بندی Farthest First گروه‌بندی می‌شوند. در مرحله دوم از الگوریتم ODA 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 WBCD نشان می‌دهد.

وایدهی و همکاران، تشخیص سرطان سینه را با استفاده از طبقه-بندی KNN ترکیبی ارائه دادند [۲۱]. آن‌ها از ترکیب سه ماتریس فاصله متفاوت یوکلیدان، کوزین و CITY-BLOCK برای طبقه-بندی استفاده کردند. اونان، یک مدل طبقه‌بندی نزدیکترین مجاور گنگ – سخت ترکیبی را با زیر مجموعه سازگار و گزینش موردی برای تشخیص اتوماتیک سرطان سینه موردمطالعه قراردادند [۲۲]. شن و همکاران، یک مدل هوشمند پیش‌بینی سرطان سینه با استفاده از تکنیک‌های داده‌کاوی ارائه دادند[۲۳]. آن‌ها برای انتخاب ویژگی‌های مؤثر در تشخیص سرطان سینه از روش خصیصه انتخابی و برای کار طبقه‌بندی از SVM استفاده کردند. آلایی و همکاران، انتخاب ویژگی را با استفاده از الگوریتم ژنتیک برای تشخیص سرطان سینه مطرح کردند و از مرتب‌سازی PS برای بهبود دقت بهره گرفتند [۲۴]. در تحقیقی مشابه رائو و همکاران، مطالعه‌ای در مورد تشخیص DNA در بیماران مبتلا به سرطان سینه با استفاده از الگوریتم ژنتیک ارائه دادند [۲۵].

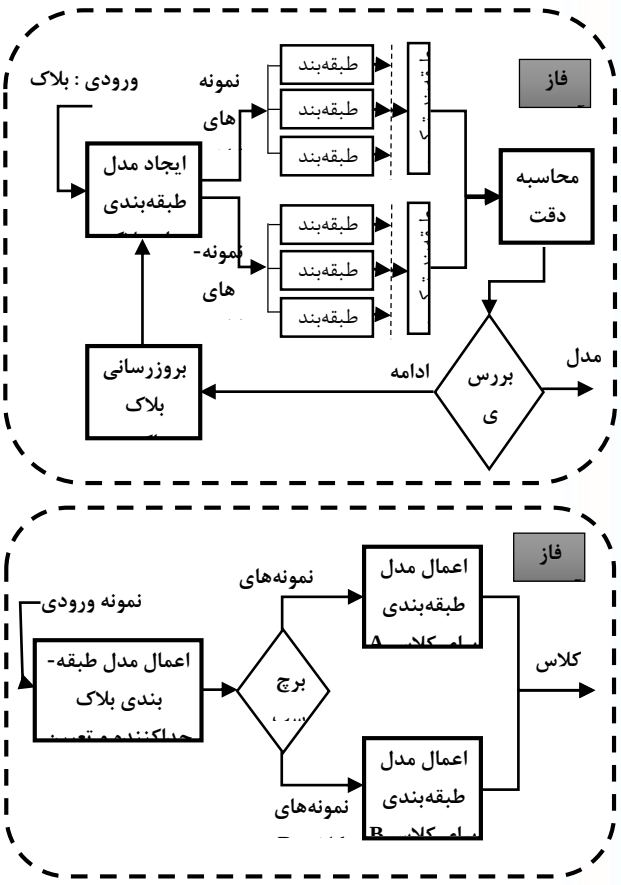
احمد و همکاران، یک مدل تشخیص سرطان سینه مبتنی بر الگوریتم ژنتیک و شبکه عصبی ANN ارائه دادند [۲۶]. در این از الگوریتم ژنتیک به‌طور هم‌زمان به‌منظور انتخاب ویژگی و بهینه-سازی پارامترهای شبکه عصبی ANN استفاده‌شده است. نتایج دقت بهترین ۹۹,۲۴٪ و میانگین ۹۸,۲۹٪ را در مجموعه داده WBCD نشان می‌دهد. کبیر و همکاران، مدل CAFS را بر مبنای شبکه عصبی ANN ارائه دادند [۲۷]. CAFS به‌طور خودکار به تعیین تعداد توده‌های لایه مخفی در طول فرآیند انتخاب ویژگی می‌پردازد.

۳- روش پیشنهادی

در این تحقیق با استفاده از پشته تعمیم‌یافته [۲۸] و ترکیب سه الگوریتم شبکه عصبی(NN)، ماشین بردار پشتیبان(SVM) و درخت تصمیم(DT) به مسئله تشخیص و پیش‌بینی سرطان سینه

www.kahrobaonline.ir

می‌پردازیم. مفهوم پشته تعمیم‌یافته در حیطه داده‌کاوی، پیش‌بینی دقیق‌تر با استفاده از ترکیب چند مدل پیش‌بینی می‌باشد. این روش در مواردی که نوع مدل‌های استفاده‌شده، بسیار متفاوت باشند، عملکرد بهتری دارد. در اینجا برای ایجاد مدل طبقه‌بندی از داده‌های جمع‌آوری‌شده سرطان سینه با روش FNA بر مبنای تشخیص دو حالت خوش‌خیم و بدخیم استفاده می‌کنیم. روش پیشنهادی ترکیبی بوده و از چندین مرحله تشکیل می‌شود. شکل ۱ روند جریان روش پیشنهادی را نشان می‌دهد.



شکل ۱ : روند جریان روش پیشنهادی

در اولین مرحله داده‌های استخراج‌شده از FNA پیش‌پردازش می‌شوند. این مرحله جهت آماده‌سازی داده‌ها برای تجزیه‌وتحلیل و انجام آزمایش‌ها است. در ادامه برای بهبود عملکرد طبقه‌بندی‌ها، سه طبقه‌بندی شبکه عصبی، ماشین بردار پشتیبان و درخت تصمیم با بهره‌گیری از پشته تعمیم‌یافته در یک مدل ابتکاری در فاز آموزش ترکیب می‌شوند. سپس یک بلاک جداکننده به‌منظور تقسیم‌بندی نمونه‌ها به k کلاس جدید در مرحله دوم طراحی می‌شود. در این تحقیق $k = 2$ در نظر گرفته‌شده است، بنابراین به هر نمونه یکی از دو کلاس A یا B تخصیص داده‌شده است. در مرحله سوم یک مدل طبقه‌بندی برای بلاک جداکننده ایجاد می‌شود. خروجی مدل بلاک جداکننده باعث تعمیم مجموعه داده اصلی به دو مجموعه داده کوچک‌تر از طریق دو کلاس A و B می‌شود. در چهارمین مرحله از روش پیشنهادی هر مجموعه داده با توجه به کلاس A و B بر مبنای استراتژی پشته تعمیم‌یافته طبقه‌بندی‌شده و یک مدل بر مبنای آن ایجاد می‌شود. مرحله پنجم ارزیابی خروجی مدل پیشنهادی و بررسی معیار توقف است. درنهایت با به‌روزرسانی بلاک جداکننده با توجه به خروجی فعلی از مدل، تکرار بهینه‌سازی تا بهبود پارامترهای بلاک جداکننده در مرحله ششم انجام می‌شود.

برای تعیین کلاس نمونه ورودی در فاز آزمایش، ابتدا کلاس A یا B نمونه ورودی با استفاده از مدل طبقه‌بندی بلاک جداکننده تشخیص داده می‌شود. تعیین کلاس A یا B برای نمونه ورودی، جهت‌دهی نمونه را به سمت یکی از طبقه‌بندی‌ها برای تعیین کلاس واقعی فراهم می‌کند. خروجی طبقه‌بندی مربوطه، کلاس خروجی نمونه ورودی را مشخص می‌کند.

۳-۱- پیش پردازی داده‌ها

پیش‌پردازش داده‌ها به‌منظور آماده‌سازی و بهبود کیفیت داده‌های اولیه انجام می‌شود. در این تحقیق پیش‌پردازش داده‌ها شامل دو بخش پاک‌سازی و نرمال‌سازی می‌باشد. در پاک‌سازی فیلدهایی از مجموعه داده نظیر شناسه بیمار که به دانش نهفته در اطلاعات مرتبط نمی‌باشد، حذف می‌گردند. همچنین نمونه‌هایی که دارای مقادیر گم‌شده [۲۹] هستند (با نماد «؟» مشخص‌شده‌اند) نیز از مجموعه داده حذف می‌شوند.

در اغلب مواقع اگر مقادیر ویژگی‌ها در محدوده‌های مختلفی وجود داشته باشد، نرمال کردن آن‌ها در دامنه‌های مشابهی منجر به کسب نتایج بهتری در طبقه‌بندی می‌شود. در این تحقیق از روش Z-score برای نرمال‌سازی استفاده می‌شود [۳۰]. 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.

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

جائیکه $e_{i,j}$ مقدار ویژگی i را برای نمونه j نشان می‌دهد. μ_j و σ_j به ترتیب میانگین و انحراف معیار مقادیر ویژگی i برای تمام نمونه‌ها می‌باشد.

۳-۲- آموزش مدل طبقه‌بندی با بهره‌گیری از پشته تعمیم‌یافته

www.kahrobaonline.ir

در پشته تعمیم‌یافته خروجی هر یک از طبقه‌بندها، یک ورودی برای یادگیر سطح بالاتر در دیگر طبقه‌بندی‌ها می‌باشد که مشخص می‌کند به چه خوبی آن‌ها را ترکیب نموده است. تجربه نشان داده است که ترکیب پیشگویی‌های انجام‌شده توسط چند روش معمولاً پیشگویی‌های دقیق‌تری را نسبت به مدل‌های انفرادی حاصل می‌کند. پیشگویی‌های به‌دست‌آمده از طبقه‌بندی‌های مختلف می‌توانند به‌عنوان ورودی طبقه‌بند دیگر مورداستفاده قرار گیرد و پشته تعمیم‌یافته پیشگویی‌های انجام‌شده را جهت ایجاد یک طبقه‌بندی پیشگویانه بهینه نهایی ترکیب کند. در این تحقیق یک روش ابتکاری بر مبنای پشته تعمیم‌یافته برای ترکیب سه طبقه‌بند شبکه عصبی، ماشین بردار پشتیبان و درخت تصمیم ارائه می‌شود.

در این تحقیق روش پشته تعمیم‌یافته برای ترکیب طبقه‌بندها روی جداسازی نمونه‌ها از مجموعه داده تمرکز دارد. 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 برای برخی از نمونه‌ها نشده است. دلایل ایجاد خطا در طبقه‌بندی‌ها را می‌تواند به‌صورت زیر توصیف کرد.

- مقادیر برخی از ویژگی‌ها در بعضی نمونه‌ها ممکن است در حالات بسیار نادر و خاصی رخ داده‌شده 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 درواقعیت نیز مشاهده نمی‌شوند.

تکنیک بلاک جداکننده پیشنهادی به‌منظور آموزش دو مدل طبقه‌بندی به‌جای یک مدل ارائه می‌شود. در این تحقیق به نمونه‌های ورودی، در داده‌های آموزشی دو برچسب A و B به‌صورت مجازی اختصاص داده می‌شود. این برچسب‌ها بدون توجه به برچسب‌های واقعی مجموعه داده هستند. وظیفه بلاک جداکننده تعمیم نمونه‌های آموزشی به دو برچسب A و B است بطوریکه هر برچسب با یک مدل مجزا طبقه‌بندی شود. تشخیص و نحوه جداسازی نمونه‌ها به دو برچسب A و B چالش این تحقیق است.

به‌منظور جداسازی اولیه و تخصیص برچسب مجازی به نمونه‌ها از الگوریتم خوشه‌بندی k-means [۳۱] با سه خوشه استفاده می‌کنیم. به نمونه‌های موجود در خوشه‌هایی با تعداد اعضاء بیشتر برچسب A و B و به نمونه‌های موجود در خوشه با کمترین اعضاء هر دو برچسب A و B تخصیص داده می‌شود. 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.

به‌منظور مدل‌سازی جداسازی انجام‌شده از یک مدل طبقه‌بندی استفاده می‌کنیم. این مدل با توجه به برچسب‌های A و B ایجاد می‌شود و مشخص می‌کند نمونه‌های ورودی از چه مدل طبقه‌بندی استفاده می‌کنند. بهینگی این مدل نقش مؤثری در افزایش کارایی مدل طبقه‌بندی نهایی دارد، لذا نحوه جداسازی و تخصیص برچسب‌های مجازی باید بهینه‌سازی شود. در اینجا از طبقه‌بندی شبکه عصبی برای ایجاد مدل بلاک جداکننده بهره می‌گیریم.

در مرحله بعد برای هر بخش از نمونه‌ها (نمونه‌ها با برچسب A و نمونه‌ها با برچسب B) به‌صورت مجزا یک مدل طبقه‌بندی ایجاد می‌شود. مدل ارائه‌شده در این قسمت بر مبنای پشته تعمیم‌یافته یک طبقه‌بندی ترکیبی ایجاد می‌کند. ساختار الگوریتم پشته تعمیم‌یافته برای ترکیب مدل‌های طبقه‌بندی از دولایه تشکیل‌شده 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 جمع‌آوری‌شده و در مجموعه‌ای جدید قرار می‌گیرند. این مجموعه به‌عنوان ورودی به طبقه‌بند لایه دوم داده می‌شود که برای آموزش این ورودی‌ها همان بردار آموزش داده‌های مجموعه یادگیری اصلی به‌صورت نظیر به نظیر خواهد بود. درواقع در این روش ابتدا اجماعی از طبقه‌بندها ساخته می‌شود که خروجی آن‌ها به‌عنوان ورودی فراطبقه‌بند لایه دوم به شمار می‌آید، سپس فراطبقه‌بند نگاشت میان خروجی‌های هر یک از طبقه‌بندهای معمولی لایه اول را با کلاس‌های خروجی واقعی یاد می‌گیرد.

در این بخش برچسب مجازی از نمونه‌ها حذف‌شده و کلاس واقعی هر نمونه برای ایجاد مدل استفاده می‌شود. در هر دو بخش از مدل ارائه‌شده سه طبقه‌بندی درخت تصمیم، ماشین بردار پشتیبان و شبکه عصبی در لایه اول وجود دارد. طبقه‌بندی درخت تصمیم و ماشین بردار پشتیبان به ترتیب برای بخش A و B در لایه دوم استفاده می‌شود. با توجه به هدف الگوریتم در بهینه‌سازی بلاک جداکننده، نیاز به تکرار و به‌روزرسانی بلاک جداکننده می‌باشد. در این تحقیق از یک الگوریتم تپه نوردی برای جستجوی بهینه‌ترین بلاک جداکننده استفاده می‌شود. در فرایند تکرار دقت کل مدل طبقه‌بندی پیشنهادی به‌صورت رابطه (۲) محاسبه می‌شود.

www.kahrobaonline.ir

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

در این رابطه C_1 دقت طبقه‌بندی بلاک جداکننده، C_2 دقت خروجی مدل برای نمونه‌های برچسب A و C_3 دقت خروجی مدل برای نمونه‌های برچسب B است. اگر دقت محاسبه‌شده در طی Γ تکرار متوالی تغییر نکند، فرایند اجرای الگوریتم متوقف می‌شود. در این تحقیق $\Gamma = 35$ در نظر گرفته‌شده است.

در هر تکرار بلاک جداکننده به‌منظور بهبود دقت، به‌روزرسانی می‌شود. اگر نمونه‌های آموزشی با برچسب A و B را در دو مجموعه داده مجزا قرار دهیم، جائیکه DS_A نمونه‌های برچسب A و DS_B نمونه‌های برچسب B را نشان می‌دهد (اشتراک DS_A و DS_B تهی نمی‌باشد). درواقع نمونه‌هایی که با یک طبقه‌بندی مستقل به‌راحتی مدل می‌شوند و باعث افزایش کارایی دسته‌بندها می‌شوند را در هر دو مجموعه داده قرار می‌دهیم. از طرفی نمونه‌هایی که بنا به دلایل مختلف تشخیص صحیح کلاس آن‌ها با استفاده از یک طبقه‌بندی مستقل ممکن نمی‌باشد را جداسازی کرده و بین دو مجموعه داده DS_A و DS_B تقسیم می‌کنیم. به‌روزرسانی بلاک جداکننده به‌صورت زیر انجام می‌شود.

- هر نمونه از مجموعه داده DS_A که به‌اشتباه طبقه‌بندی‌شده است، به‌احتمال α به مجموعه داده DS_B انتقال پیدا می‌کند و به‌احتمال $1-\alpha$ با یک نمونه از مجموعه داده DS_B که به‌اشتباه طبقه‌بندی‌شده است، جابجا می‌شود.
- هر نمونه از مجموعه داده DS_B که به‌اشتباه طبقه‌بندی‌شده است، به‌احتمال α به مجموعه داده DS_A انتقال پیدا می‌کند و به‌احتمال $1-\alpha$ با یک نمونه از مجموعه داده DS_A که به‌اشتباه طبقه‌بندی‌شده است، جابجا می‌شود.
- اگر در یکی از دو مجموعه داده DS_A یا DS_B نمونه‌ای با کلاس‌بندی اشتباه یافت نشد، همواره احتمال α برقرار است و نمونه‌ها انتقال پیدا می‌کنند.

در این تحقیق مقدار $\alpha = 0.5$ در نظر گرفته‌شده است. به‌طورکلی بلاک جداکننده داده‌های نمونه‌های آموزشی را به دو قسمت تقسیم نمی‌کند، بلکه از داده‌های آموزشی دو مجموعه داده جدید می‌سازد به‌نحوی‌که اشتراک دو مجموعه داده تهی نباشد. درواقع هر مجموعه داده نسخه مشابه مجموعه داده اصلی است که فقط نمونه‌هایی که باعث کاهش دقت طبقه‌بند می‌شوند از آن حذف‌شده است.

۳-۳- آزمایش مدل طبقه‌بندی

برای پیش‌بینی کلاس نمونه ورودی جدید در فاز آزمایش، ابتدا نمونه ورودی با استفاده از مدل طبقه‌بندی بلاک جداکننده ارزیابی‌شده و برچسب A یا B آن مشخص می‌شود. برچسب نمونه نوع طبقه‌بندی جهت پیش‌بینی کلاس واقعی را تعیین می‌کند. نمونه دارای برچسب A از مدل فراطبقه‌بند درخت تصمیم و نمونه‌های برچسب B از مدل فراطبقه‌بند ماشین بردار پشتیبان جهت پیش‌بینی سرطان استفاده می‌کنند. دقت نهایی داده‌های آموزشی از رابطه زیر حاصل می‌شود.

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

در این رابطه N تعداد نمونه‌های آزمایش، c_i کلاس پیش‌بینی‌شده نمونه ورودی و t_i کلاس واقعی نمونه ورودی می‌باشد. برای بررسی نمونه‌های واقعی که در بانک دیجیتال نیستند، ابتدا باید ویژگی‌های موردنیاز از بیمار استخراج شود. ویژگی‌های موردنیاز با توجه به مدل طبقه‌بندی ارائه‌شده روی هر مجموعه داده از پایگاه ویسکانسین مشخص می‌شود. پس از استخراج ویژگی‌ها، نمونه را مطابق با شکل ۱ به‌عنوان یک داده آزمایشی وارد سیستم کرده و درنهایت برچسب نمونه توسط مدل پیش‌بینی می‌شود.

۴- نتایج و آزمایش‌ها

برای انجام شبیه‌سازی و تجزیه‌وتحلیل روش پیشنهادشده از نرم‌افزار MATLAB ورژن 2017a 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 [۳۲] استفاده می‌شود. جدول ۱ چند نمونه از رکوردهای مجموعه داده WBCD را با ۹ ویژگی نشان می‌دهد.

جدول ۱ : چند نمونه از رکوردهای مجموعه داده WBCD

برچسب	تقسیم میتوز	هستک	کروماتین	هسته خالی	اندازه سلول	چسبندگی	یکنواختی	یکنواختی	ضخامت توده	شناسه بیمار
۲	۱	۱	۳	۱	۲	۱	۱	۱	۵	۱۰۰۰۰۲۵
۲	۱	۲	۳	۱۰	۷	۵	۴	۴	۵	۱۰۰۲۹۴۵
۲	۱	۱	۳	۲	۲	۱	۱	۱	۳	۱۰۱۵۴۲۵

در این تحقیق از تکنیک اعتبارسنجی 5-Fold برای ارزیابی مدل استفاده می‌شود [۳۳]. در هر مرحله از اعتبارسنجی 10-Fold، مجموعه داده اصلی به دو بخش E^T (آموزش) و E^P (آزمایش)

www.kahrobaonline.ir

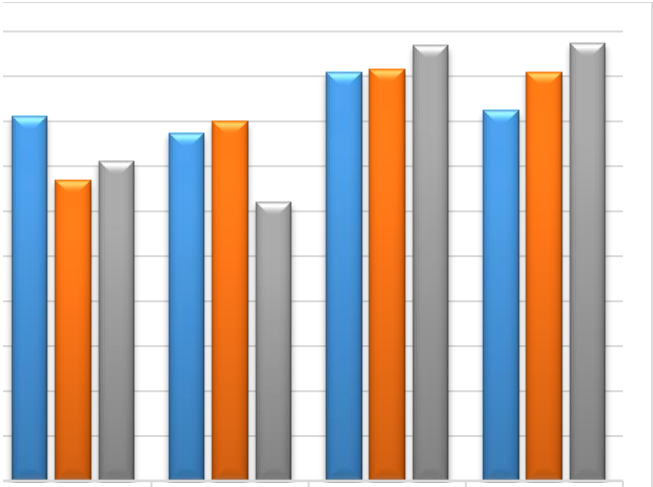
تقسیم می‌شود جائیکه ۹۰٪ نمونه‌ها برای E^T و ۱۰٪ دیگر برای E^P استفاده می‌شود. به‌منظور حصول اطمینان از نتایج ارائه‌شده میانگین ۲۰ بار اجرای الگوریتم در تمام آزمایش‌ها محاسبه‌شده است. جدول ۲ نتایج روش پیشنهادی را بر روی سه مجموعه داده WBCD، WDBC و WPBC نشان می‌دهد.

در روش پیشنهادی برای ساخت مدل بلاک جداکننده از طبقه-بندی شبکه عصبی، برای مدل کردن نمونه‌های برچسب A از طبقه‌بندی درخت تصمیم و برای مدل کردن نمونه‌های برچسب B از طبقه‌بندی ماشین بردار پشتیبان بهره گرفته‌شده است. دلیل انتخاب این طبقه‌بندها برای این سه وظیفه، عملکرد بهتر آن‌ها بوده است. برای بررسی صحت این موضوع ما ترکیب‌های مختلف این سه طبقه‌بند را برای سه قسمت مختلف روش پیشنهادی بکار گرفته و نتایج حاصل‌شده را مقایسه کرده‌ایم. شکل ۲ نتایج این آزمایش را نشان می‌دهد.

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

مجموعه داده	دقت مدل بلاک جداکننده (٪)	دقت مدل برچسب A (٪)	دقت مدل برچسب B (٪)	دقت مدل 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

نتایج نشان می‌دهد که ترکیب NN; DT; SVM بهترین عملکرد را دارد. جائیکه از NN برای مدل بلاک جداکننده، از DT برای مدل کردن نمونه‌های برچسب A و از SVM برای مدل کردن نمونه‌های برچسب B استفاده‌شده است. در ادامه نتایج روش پیشنهادی را با تعدادی از روش‌های طبقه‌بندی کلاسیک مقایسه می‌کنیم. روش-های طبقه‌بندی مورد مقایسه شبکه عصبی MLP، درخت تصمیم ID3 و ماشین بردار پشتیبان هستند. نتایج این مقایسه برای سه مجموعه داده WBCD، WDBC و WPBC در جدول ۳ گزارش‌شده است.



SVM; NN; DT	SVM; DT; NN	NN; DT; SVM	NN; SVM; DT
99.05	98.87	99.54	99.12
98.35	99	99.58	99.55
98.56	98.1	99.84	99.87

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

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

مجموعه داده	شبکه عصبی MLP	درخت تصمیم ID3	ماشین بردار پشتیبان	پشته تعمیم‌یافته
WBCD	۸۹,۳۳	۹۲,۸۵	۹۳,۵۸	۹۹,۵۴
WDBC	۸۵,۶۰	۹۳,۱۷	۹۰,۲۳	۹۹,۵۸
WPBC	۸۸,۱۰	۹۰,۷۰	۹۱,۴۷	۹۹,۸۴

نتایج مقایسه، عملکرد بهتر روش پشته تعمیم‌یافته پیشنهادی را نشان می‌دهد. روش ارائه‌شده با توجه به تخصیص برچسب جدید و هدایت مسیر نمونه‌ها به سمت طبقه‌بندی با ارزیابی دقیق‌تر توانسته است دقت بهتری نسبت به روش‌های طبقه‌بندی کلاسیک گزارش دهد. در ادامه برای بررسی عملکرد و ارزیابی روش پشته تعمیم‌یافته، نتایج این تحقیق را با تعدادی از جدیدترین روش‌های تشخیص سرطان سینه که از پایگاه داده ویسکانسین در آزمایش-های خود بهره گرفته‌اند، مورد مقایسه قرار می‌دهیم. جدول ۴ نتایج این مقایسه را نشان می‌دهد.

www.kahrobaonline.ir

جدول ۴ : مقایسه دقت تشخیص سرطان سینه مدل پیشنهادی با سایر روش‌ها مشابه روی پایگاه داده ویسکانسین

مجموعه داده	EM-Fuzzy, 2017 [۱۸]	ODA, 2016 [۲۰]	RS-BPNN, 2015 [۳۴]	SMO-IBK, 2012 [۳۵]	پشته تعمیم‌یافته
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

روش تشخیص سرطان پیشنهادی در مقایسه با سایر روش‌های موردبررسی و به ازای برخی از پایگاه‌های داده دقت بیشتری داشته و در بقیه موارد نیز دقت مناسبی را ارائه می‌دهد.

۵- نتیجه‌گیری و پیشنهادات

با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوین و باصرفه‌تری در نظام سلامت و درمان ارائه کرد که با دقت بالایی قادر به تشخیص سرطان سینه باشند. استفاده از این سیستم‌ها می‌تواند موجب کاهش عمل‌های غیرضروری شود و دقت تشخیص سرطان سینه را بهبود بخشد. فرایند تشخیص بیماری سرطان سینه با استفاده از پشته تعمیم‌یافته در یک مدل ترکیبی دولایه ارائه شد که نتایج بهتری را نسبت به روش‌های مشابه نشان می‌دهد. 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. به‌طورکلی روش پشته تعمیم‌یافته ارائه‌شده با توجه به بلاک جداکننده و شناسایی نمونه-هایی که باعث ایجاد خطا می‌شوند، در پایگاه‌های داده‌ای که احتمال وجود نویز و نمونه‌های با مقادیر خاص در آن‌ها بیشتر است عملکرد بهتری خواهد داشت، هرچند به‌طور خاص وجود نویز و پالایش آن‌ها در مدل ارائه‌شده در نظر گرفته نشده است. برای کارهای آتی پیشنهادی می‌شود قبل از اعمال مدل طبقه‌بندی از روش‌هایی نظیر الگوریتم ژنتیک جهت کاهش ابعاد مجموعه داده در قالب یک مسئله انتخاب ویژگی استفاده شود.

مراجع

[1] King, M. C., Marks, J. H., & Mandell, J. B. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. Science. 2003; 302(5645): 643-646.

[2] Siegel, R. L., Miller, K. D., & Jemal, A. Cancer statistics, 2015. CA: a cancer journal for clinicians. 2015; 65(1): 5-29.

[3] Asghar, U., Witkiewicz, A. K., Turner, N. C., & Knudsen, E. S. The history and future of targeting cyclin-dependent kinases in cancer therapy. Nature reviews Drug discovery. 2015; 14(2): 130-146.

[۴] کیانی, بهزاد و علیرضا آتشی، ایجاد یک مدل پیش‌آگهی مبتنی بر داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان، مجله انفورماتیک سلامت و زیست پزشکی، ۱۳۹۳, ۱(۱), ۲۶-۳۱.

[5] Asuncion, A., & Newman, D. UCI machine learning repository. 2007.

[6] Robert, C., Guilpin, C., & Limoge, A. Review of neural network applications in sleep research. Journal of Neuroscience methods. 1998; 79(2): 187-193.

[7] Vapnik, V., & Izmailov, R. Knowledge transfer in SVM and neural networks. Annals of Mathematics and Artificial Intelligence. 2017: 1-17.

[8] Safavian, S. R., & Landgrebe, D. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics. 1991; 21(3): 660-674.

[9] Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., ... & Pasini, B. 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. The American Journal of Human Genetics. 2003; 72(5): 1117-1130.

[۱۰] دهقان, پروانه، مائده مقربی، ایمان ذباح، کامران لایقی و علی ماروسی، مدل‌سازی بیماری سرطان پستان با استفاده از روش‌های مبتنی بر داده‌کاوی، مجله انفورماتیک سلامت و زیست پزشکی، ۱۳۹۶, ۴(۴), ۲۶۶-۲۷۸.

[۱۱] اروجی, اعظم و مصطفی لنگری‌زاده، ارزیابی تأثیر انتخاب ویژگی و توابع کرنل مختلف در عملکرد SVM برای تشخیص سرطان پستان، مجله انفورماتیک سلامت و زیست پزشکی، ۱۳۹۷, ۵(۲), ۲۴۴-۲۵۱.

[۱۲] غلامی, محسن و سید جواد میر عابدینی، استفاده هم‌زمان از همبستگی خطی پیرسون و ترکیب الگوریتم های داده‌کاوی به‌منظور بهبود پیش بینی نوع تومور در بیماران سرطانی. مجله مهندسی مخابرات، ۱۳۹۶, ۷(۲۵), ۲۹-۱۲.

[۱۳] شیخ پور, راضیه و مهدی آقاصرام، انتخاب ویژگی‌های مؤثر در تشخیص سرطان سینه با استفاده از مدل‌های پارامتریک یادگیری ماشین. فصلنامه علمی-پژوهشی بیماری های سینه، ۱۳۹۴, ۸(۲), ۱۶-۲۳.

[۱۴] بوداغی, محمد، داده‌کاوی از طریق سیستم های خبره جهت

پیش بینی و تشخیص سرطان سینه و مقایسه آن با دیگر روش های داده‌کاوی. دومین کنفرانس ملی ریاضی: مهندسی پیشرفته با تکنیک های ریاضی، ارومیه، دانشگاه آزاد اسلامی واحد ارومیه، ۱۳۹۶.

[۱۵] شمس الکتابی, شهرام؛ حمیدرضا امامی و سمیرا زارع چاوشی، تشخیص و شناسایی الگوهای خوش‌خیم و بدخیم در سرطان سینه به کمک شبکه عصبی. سومین همایش ملی برق و کامپیوتر امین، اصفهان، موسسه آموزش عالی امین فولادشهر، ۱۳۹۶.

[۱۶] موسوی, سیدمحمدمهدی و مهناز رفیعی، تشخیص سرطان سینه با استفاده از دسته بندی فازی مبتنی بر الگوریتم ACO. سومین کنفرانس ملی رویکردهای نوین در مهندسی کامپیوتر و برق، رودسر، دانشگاه آزاد اسلامی واحد رودسر و املش- باشگاه پژوهشگران جوان و نخبگان واحد رودسر و املش، ۱۳۹۶.

[۱۷] غلامی, صبا و سعید فرضی، تشخیص سرطان سینه با استفاده از یادگیری عمیق. چهارمین کنفرانس ملی فناوری اطلاعات، کامپیوتر و مخابرات، مشهد، دانشگاه تربت حیدریه، ۱۳۹۶.

[18] 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): 133-144.

[19] Diz, J., Marreiros, G., & Freitas, A. Applying data mining techniques to improve breast cancer diagnosis. Journal of medical systems. 2016; 40(9): 203.

[20] Devi, R. D. H., & Devi, M. I. Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer. International Journal of Advanced Engineering Technology. 2016; 7(2): 93-98.

[21] Vaidehi, K., & Subashini, T. S. Breast tissue characterization using combined K-NN classifier. Indian Journal of Science and Technology. 2015; 8(1): 23-26.

[22] 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): 6844-6852.

[23] Shen, R., Yang, Y., & Shao, F. Intelligent breast cancer prediction model using data mining techniques. In Intelligent Human-Machine Systems and Cybernetics (IHMSC), IEEE. 2014; p. 384-387.

[24] 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.

[25] Rao, R., Rivers, A., Rahimi, A., Wooldridge, R., Rao, M., Leitch, M., ... & Haley, B. B. Genetic Ancestry using Mitochondrial DNA in patients with Triple-negative breast cancer (GAMiT study). Cancer. 2017; 123(1): 107-113.

[26] Ahmad, F., Isa, N. A. M., Hussain, Z., Osman, M. K., & Sulaiman, S. N. A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer. Pattern Analysis and Applications. 2015; 18(4): 861-870.

[27] Kabir MM, Islam MM, Murase K. A new wrapper feature selection approach using neural network. Neurocomputing. 2010; 73(16–18): 3273–3283.

[28] Wolpert, D. H. Stacked generalization. Neural networks. 1992; 5(2): 241-259.

[29] 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.

[30] Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. Analysis of microarray data using Z score transformation. The Journal of molecular diagnostics. 2003; 5(2): 73-81.

[31] Wagstaff, K., Cardie, C., Rogers, S., & Schrödl, S. Constrained k-means clustering with background knowledge. 2001; In ICML, 1: 577-584.

[32] Breast Cancer Wisconsin (Original) dataset, UCI machine language repository. 1992.

[33] Bengio, Y., & Grandvalet, Y. No unbiased estimator of the variance of k-fold cross-validation. Journal of machine learning research. 2004; 5(Sep): 1089-1105.

[34] Nahato, K. B., Harichandran, K. N., & Arputharaj, K. Knowledge mining from clinical datasets using rough sets and backpropagation neural network. Computational and mathematical methods in medicine. 2015; 2015(460189): 13.

[35] Salama, G. I., Abdelhalim, M. B., & Zeid, M. A. E. Experimental comparison of classifiers for breast cancer diagnosis. In Computer Engineering & Systems (ICCES), IEEE. 2012; p. 180-185.