استفاده از دسته‌بندی ترکیبی مبتنی بر جداسازی نمونه‌های متعارف و نامتعارف برای تشخیص سرطان پستان

امین رضایی‌پناه و حسام واقع‌بین

چکیده: سرطان پستان یکی از رایج‌ترین انواع سرطان‌ها در زنان می‌باشد و در سال‌های اخیر رشد قابل توجهی در تعداد افراد مبتلا به آن گزارش شده است. با گسترش روزافزون علم استفاده از داده‌کاوی در پزشکی به یکی از زمینه‌های پرکاربرد برای بهبود سیستم‌های درمانی تبدیل شده است. در این تحقیق فرایند تشخیص بیماری سرطان پستان در دو مرحله انجام می‌شود. در مرحله اول از یک الگوریتم ژنتیک بهبودیافته برای تشخیص ویژگی‌های مؤثر در پیش‌بینی این بیماری استفاده شده و در مرحله دوم نمونه‌های متعارف و نامتعارف به منظور افزایش دقت و ایجاد مدل دسته‌بندی نهایی شناسایی می‌شوند. برای کار دسته‌بندی مقایسه‌ای بین دو مدل درخت تصمیم و ماشین بردار پشتیبان انجام شده که نتایج، برتری مدل ماشین بردار پشتیبان را نشان می‌دهد. نتایج آزمایش‌های انجام‌شده دقت تشخیص سرطان پستان را روی مجموعه داده‌های WBCD، WDBC و WPBC به ترتیب ۹۹/۲۶٪، ۹۸/۵۵٪ و ۹۸/۴۵٪ گزارش می‌دهد.

کلیدواژه: **الگوریتم ژنتیک، دسته‌بندی، سرطان پستان، ویژگی‌های مؤثر، نمونه‌های متعارف و نامتعارف.**

۱- مقدمه

سرطان پستان یکی از شایع‌ترین انواع سرطان‌ها در زنان است و پس از سرطان ریه، دومین عامل مرگ‌ومیر در زنان شناخته می‌شود. از سال ۱۹۸۹ میلادی، پیشرفت‌های چشم‌گیری در زمینه درمان سرطان سینه به وجود آمده است. احتمال مرگ‌ومیر ناشی از این بیماری در زنان حدود ۷۲ درصد است [۱]. این بیماری در سنین حدود پنجاه سالگی بیشترین شیوع را دارد اما در هر سنی ممکن است مشاهده شود. نکته جالب توجه در مورد این بیماری این است که اگر زود تشخیص داده شود، در اکثر موارد به طور کامل درمان می‌شود [۱]. آزمایش آسپیراسیون سوزنی (FNA) روشی ارزان و غیر تهاجمی برای تشخیص دقیق سرطان پستان می‌باشد. در این روش با استخراج خصوصیات سیتولوژی از بافت پستان، می‌توان خوش‌خیم یا بدخیم‌بودن تومور را تشخیص داد [۲].

در سال‌های اخیر استفاده از روش‌های داده‌کاوی برای کمک به پزشکان و متخصصین در تشخیص زودهنگام این بیماری مورد توجه قرار گرفته است [۳]. پژوهش‌ها و روش‌های ارائه‌شده در این زمینه، موفقیت‌آمیزبودن سیستم‌های هوش مصنوعی را اثبات کرده است [۴]. در این تحقیق فرایند تشخیص سرطان پستان با استفاده از ویژگی‌های استخراج‌شده از FNA و با کمک الگوریتم‌های داده‌کاوی انجام می‌شود.

این مقاله در تاریخ ۱۰ آذر ماه ۱۳۹۷ دریافت و در تاریخ ۲۸ اردیبهشت ماه ۱۳۹۸ بازنگری شد.

امین رضایی‌پناه (نویسنده مسئول)، مؤسسه آموزش عالی رهجویان دانش برازجان، بوشهر، ایران، (email: amin.rezaeipanah@gmail.com).

حسام واقع‌بین، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی، واحد بوشهر، ایران، (email: hwhesam@gmail.com).

در این میان مطالعات انجام‌شده نشان می‌دهند که روش‌های ترکیب دسته‌بندی‌ها به یک ابزار مطلوب به منظور افزایش دقت و کارایی تبدیل شده است [۵]. در سال‌های اخیر روش‌های مختلفی برای ترکیب نتایج مدل‌های دسته‌بندی نظیر شبکه‌های عصبی و سیستم‌های فازی ارائه شده است [۶] تا [۸].

در این تحقیق نیز چندین مدل دسته‌بندی مختلف به منظور افزایش دقت در داده‌های سرطان پستان ترکیب می‌شوند. ما داده‌های آموزشی را به نحوی بین مدل‌های دسته‌بندی تقسیم می‌کنیم که حداکثر دقت حاصل شود. روش تقسیم‌بندی شناسایی نمونه‌های متعارف و نامتعارف است. در یک مدل دسته‌بندی مستقل، نمونه‌هایی را که به درستی دسته‌بندی شده‌اند نمونه‌های متعارف و نمونه‌هایی را که باعث ایجاد خطا شده‌اند نمونه‌های نامتعارف می‌نامیم.

در ادامه این تحقیق در بخش دوم به بررسی برخی از جدیدترین کارهای انجام‌شده در تشخیص سرطان پستان می‌پردازیم. روش پیشنهادی در بخش سوم مطرح شده و نتایج حاصل از آن در بخش چهارم ارائه می‌شود. در نهایت در بخش پنجم نتیجه‌گیری و پیشنهادها مطرح می‌گردد.

۲- کارهای انجام‌شده

در زمینه پیش‌بینی و تشخیص سرطان پستان با استفاده از مدل‌های دسته‌بندی تحقیقات گسترده‌ای انجام شده است. در ادامه به برخی از این روش‌ها اشاره می‌کنیم. شیخ‌پور و همکاران برای انتخاب ویژگی‌های مؤثر در تشخیص سرطان پستان از یک مدل پارامتریک یادگیری ماشین استفاده کردند [۹]. صادقی‌پور و همکاران، تشخیص سرطان پستان را بر اساس یک سیستم هوشمند ترکیبی مبتنی بر الگوریتم تکاملی ارائه دادند [۱۰]. سیستم هوشمند پیشنهادشده ترکیبی از یک الگوریتم کرم شب‌تاب برای انتخاب ویژگی‌ها و مدل ماشین بردار پشتیبان برای کار دسته‌بندی است.

ماندال، الگوریتم‌های داده‌کاوی نائو بیز، رگرسیون لجستیک و درخت تصمیم را به منظور تشخیص سلول‌های سرطانی پستان مورد تجزیه و تحلیل قرار داد [۱۱]. هدف این تحقیق یافتن کوچک‌ترین زیرمجموعه از ویژگی‌ها است که می‌توانند دسته‌بندی خوش‌خیم و بدخیم‌بودن سرطان پستان را با دقت بالایی تضمین کند. نتایج برتری مدل رگرسیون لجستیک را با دقت ۹۷/۹٪ نشان می‌دهد. اون‌آن، یک مدل دسته‌بندی هوشمند ترکیبی را برای تشخیص سرطان پستان معرفی کرد [۱۲]. برای کار دسته‌بندی از الگوریتم نزدیک‌ترین همسایه Fuzzy-Rough استفاده شده که دقت ۹۹/۷۱٪ را روی مجموعه داده WBCD ارائه می‌دهد. در تحقیقی دیگری پاتیل و همکاران، دو روش قطعه‌بندی Thresholding و Watershed را به منظور تشخیص سلول‌های سرطانی پستان مورد مقایسه قرار دادند [۱۳].

شکل ۱: فرایند عملگر آمیزش پیشنهادی.

نیلاشی و همکاران، یک سیستم مبتنی بر دانش را برای دسته‌بندی سرطان پستان با استفاده از روش منطق فازی پیشنهاد دادند [۱۴]. در این تحقیق از روش EM (حداکثر انتظار) برای خوشه‌بندی داده‌ها و ایجاد گروه‌های مشابه استفاده شده و تولید قوانین فازی برای کار دسته‌بندی با استفاده از درخت رگرسیون انجام می‌شود. وانگ و همکاران، یک روش دسته‌بندی ترکیبی را برای تشخیص سرطان پستان ارائه دادند [۱۵]. الگوریتم‌های دسته‌بندی پیشنهادی ترکیبی از روش‌های SMOTE و بهینه‌سازی ازدحام ذرات (PSO) می‌باشند که در آنها برخی از مشهورترین روش‌های دسته‌بندی نظیر رگرسیون لجستیک، مدل درخت تصمیم C۵ و KNN نیز ادغام می‌شوند. دیز و همکاران، یک رویکرد مبتنی بر داده‌کاوی برای غده‌شناسی در سرطان پستان ارائه دادند [۱۷]. نتایج حاصل‌شده دقت تشخیص ۸۹/۳٪ برای داده‌های خوش‌خیم و ۶۴/۷٪ برای داده‌های بدخیم را نشان می‌دهد.

دیوی و همکاران، یک الگوریتم سه‌مرحله‌ای را برای تشخیص زودهنگام سرطان پستان ارائه دادند [۱۶]. در مرحله اول داده‌ها با الگوریتم Farthest First خوشه‌بندی شده و در مرحله دوم دسته‌بندی داده‌ها با الگوریتم ODA انجام می‌شود. در نهایت خوش‌خیم و بدخیم‌بودن داده‌ها با استفاده از الگوریتم دسته‌بندی J۴۸ در مرحله سوم تشخیص داده می‌شود. نتایج دقت ۹۹/۹٪ را برای مجموعه داده WBCD و ۹۹/۶٪ را برای مجموعه داده WDBC نشان می‌دهد.

در تحقیق دیگری گواش و همکاران، مقایسه‌ای بین مدل‌های کلاس‌بندی SVM و MLP BPN برای تشخیص سرطان پستان انجام دادند [۱۸]. وایدهی و همکاران، تشخیص سرطان پستان را با استفاده از دسته‌بندی KNN ترکیبی ارائه دادند [۱۹] که دقت این مدل ۹۱/۱۶٪ گزارش شده است. شیخ‌پور و همکاران، از الگوریتم بهینه‌سازی ذرات به منظور انتخاب ویژگی در تخمین چگالی کرنل مبتنی بر دسته‌بندی جهت تشخیص سرطان پستان استفاده کردند [۲۰]. احمد و همکاران، یک مدل تشخیص سرطان پستان مبتنی بر الگوریتم ژنتیک و شبکه عصبی ANN ارائه دادند [۲۱]. نتایج شبیه‌سازی دقت بهترین ۹۹/۲۴٪ و میانگین ۹۸/۲۹٪ را گزارش می‌دهد.

۳- روش پیشنهادی

به طور کلی در برخی موارد ممکن است اطلاعات یک مجموعه داده با نویز همراه باشد که این امر باعث کاهش دقت مدل‌های دسته‌بندی می‌شود. همچنین ممکن است داده‌های برخی نمونه‌ها بسیار نادر بوده و در نتیجه فراوانی ویژگی‌های مربوط به آنها در پایگاه داده بسیار کم باشد. کم‌بودن این نمونه‌ها اغلب باعث کاهش کیفیت مدل‌های دسته‌بندی و بروز خطا در آنها می‌شود. هدف ما در این تحقیق یافتن نمونه‌هایی با این خصوصیات در پایگاه داده سرطان پستان به منظور افزایش دقت دسته‌بندی می‌باشد.

روش پیشنهادی شامل چندین مرحله است. در مرحله اول پیش‌پردازش داده‌ها انجام می‌شود. انتخاب ویژگی‌های مؤثر با استفاده از الگوریتم ژنتیک بهبودیافته در مرحله دوم انجام می‌شود. در نهایت مدل دسته‌بندی در مرحله سوم ارائه می‌شود. مدل دسته‌بندی پیشنهادی از یک سیستم ترکیبی جدید مبتنی بر جداسازی نمونه‌های متعارف و نامتعارف استفاده می‌کند. در ادامه جزئیات روش پیشنهادی تشریح می‌گردد.

۳-۱ پیش‌پردازش

این مرحله برای آماده‌سازی داده‌ها جهت پردازش و همچنین بهبود کیفیت داده‌های واقعی انجام می‌شود. هنگامی که مقادیر ویژگی در دامنه متفاوتی قرار داشته باشند آنها را در دامنه مشابهی قرار می‌دهد که این امر اغلب باعث کاهش پیچیدگی محاسباتی و افزایش عملکرد مدل‌های دسته‌بندی می‌شود. برای نرمال‌سازی از روش مینیمم- ماکسیمم به صورت (۱) استفاده می‌شود

$$x_{i,j}^{new} = \frac{x_{i,j} - \min(x_j)}{\max(x_j) - \min(x_j)} \qquad (۱)$$

جایی که $x_{i,j}$ به ویژگی j ام از نمونه i ام اشاره دارد و $\min(x_j)$ و $\max(x_j)$ نیز به ترتیب کوچک‌ترین و بزرگ‌ترین مقدار از ویژگی j ام در همه نمونه‌ها می‌باشد.

۳-۲ انتخاب ویژگی با الگوریتم ژنتیک

مهم‌ترین هدف در مسئله انتخاب ویژگی حذف ویژگی‌های نامرتبط و همچنین دارای افزونگی است [۲۲]. در این تحقیق برای انتخاب ویژگی‌ها از یک الگوریتم ژنتیک بهبودیافته با تکنیک طول رشته متغیر استفاده می‌کنیم که علاوه بر ویژگی‌های مؤثر، تعداد آنها را نیز به صورت خودکار تشخیص می‌دهد.

کروموزوم‌ها با رشته باینری به طول تعداد کل ویژگی‌ها نمایش داده می‌شود که در آن ژن با مقدار ۱ به مفهوم انتخاب و ۰ به مفهوم عدم انتخاب ویژگی است. در این نوشتار از نماد DNF برای نمایش تعداد ویژگی‌های انتخاب‌شده با توجه به تعداد ژن‌های ۱ کروموزوم استفاده می‌شود. برای اعمال تکنیک طول رشته متغیر از مقادیر مختلف DNF به صورت تصادفی در تولید جمعیت اولیه استفاده می‌شود که در نهایت این امر باعث تشخیص تعداد ویژگی‌های مؤثر خواهد شد.

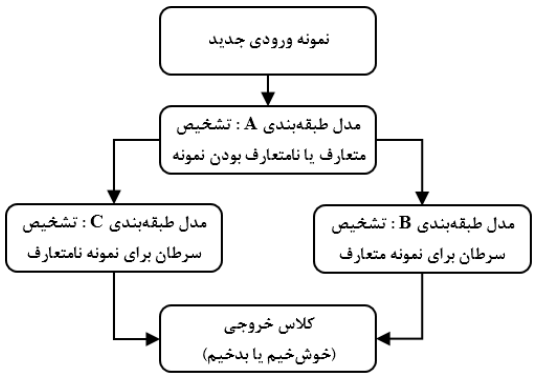
برای محاسبه برازندگی و کار دسته‌بندی از دو الگوریتم ID۳ و SVM به صورت مقایسه‌ای استفاده می‌شود. در الگوریتم ژنتیک پیشنهادی برای عملگر انتخاب از روش تورنومنت استفاده گردیده است [۲۳].

کروموزوم‌های انتخاب‌شده برای تولیدمثل به احتمال C_R با عملگر آمیزش مبادله می‌شوند. در عملگر آمیزش پیشنهادی، ژن‌های مشترک انتخابی (ژن با مقدار ۱) از هر دو والد در کروموزوم فرزند کپی شده و سایر ژن‌ها با احتمال η از ویژگی‌های مشاهده‌نشده مقداردهی می‌شوند. ویژگی‌های مشاهده‌نشده، ویژگی‌هایی هستند که در والدین وجود ندارند. شکل ۱ مثالی از عملگر آمیزش پیشنهادی را نشان می‌دهد.

پس از عملگر آمیزش، عملگر جهش با احتمال M_R روی هر ژن از کروموزوم فرزند اعمال شده و محتوای آن را تغییر می‌دهد. در نهایت فرایند بهینه‌سازی الگوریتم ژنتیک با انتخاب ویژگی‌های مؤثر در بهترین کروموزوم جمعیت خاتمه می‌یابد.

۳-۳ مدل دسته‌بندی ترکیبی بر مبنای جداسازی نمونه‌های متعارف و نامتعارف

به طور کلی وظیفه یک مدل دسته‌بندی پیش‌بینی الگوهای یک مجموعه داده با بررسی نمونه‌ها به صورت استقرایی می‌باشد [۲۴]. در این



شکل ۲: فرایند تعیین کلاس نمونه جدید در فاز آزمایش.

تحقیق جهت تشخیص سرطان پستان یک مدل دسته‌بندی ترکیبی جدید بر مبنای تشخیص نمونه‌های متعارف و نامتعارف ارائه می‌گردد. در برخی پایگاه‌های داده مقادیر ویژگی‌های برخی از نمونه‌ها دارای حالات بسیار خاصی هستند که با سایر نمونه‌های کلاس مشابه فاصله زیادی دارند [۲۵]. این نمونه‌ها اغلب باعث ایجاد خطا در مدل‌های دسته‌بندی می‌شوند. در این نوشتار این نمونه‌ها با عنوان نمونه‌های نامتعارف شناخته شده و با شناسایی آنها عملکرد مدل‌های دسته‌بندی به طور قابل توجهی بهبود می‌یابد.

در این تحقیق برای تشخیص نمونه‌های نامتعارف از الگوریتم ژنتیک استفاده می‌شود. الگوریتم ژنتیک نمونه‌ها را به دو بخش متعارف و نامتعارف تقسیم می‌کند و سپس یک مدل دسته‌بندی کلاسیک بر اساس این جداسازی ایجاد می‌شود. بنابراین نوع نمونه‌ها (متعارف یا نامتعارف) بر اساس این مدل تشخیص داده شده و در نهایت تشخیص سرطان بر مبنای نوع نمونه با یک روش دسته‌بندی کلاسیک متمایز مدل می‌شود. به طور خاص در این تکنیک از سه مدل دسته‌بندی کلاسیک استفاده می‌شود: مدل دسته‌بندی اول (A) برای تشخیص نوع نمونه (متعارف یا نامتعارف) و مدل دسته‌بندی دوم (B) و سوم (C) به ترتیب برای تشخیص سرطان در نمونه‌های متعارف و نامتعارف است. در فاز آزمایش، تعیین کلاس نمونه ورودی مطابق شکل ۲ انجام می‌شود.

نتایج بررسی‌ها نشان‌دهنده عملکرد ضعیف مدل پیشنهادی در فاز آزمایش است. دلیل این امر تعداد کم نمونه‌ها در دو مدل دسته‌بندی B و C است. برای رفع این مشکل نمونه‌های خنثی را برای هر دو دسته‌بندی B و C در نظر می‌گیریم. نمونه‌های خنثی، نمونه‌هایی هستند که انتخاب آنها به طور هم‌زمان برای نمونه‌های متعارف و نامتعارف تأثیری در عملکرد نهایی مدل ندارد. در ادامه فرایند جداسازی نمونه‌های متعارف و نامتعارف را به صورت ابتکاری با یک الگوریتم ژنتیک انجام می‌دهیم.

ساختار کروموزوم‌ها در الگوریتم ژنتیک پیشنهادی به صورت یک آرایه دوبعدی باینری به طول کل نمونه‌های مجموعه داده است. مقدار ۱ در سطر اول این آرایه، نمونه‌های متعارف و به طور مشابه مقدار ۱ در سطر دوم نمونه‌های نامتعارف را نشان می‌دهد. محدودیت راه‌حل برای هر کروکوزوم، انتخاب یک نمونه برای حداقل یک نوع متعارف یا نامتعارف است.

برای محاسبه برازندگی (دقت نهایی مدل دسته‌بندی) از (۲) استفاده می‌کنیم

$$Fitness = \frac{۲.acc(A) + acc(B) + acc(C)}{۴}. \qquad (۲)$$

جایی که acc دقت دسته‌بندی نمونه‌های مرتبط را نشان می‌دهد. با توجه به وابستگی بالای کارایی مدل پیشنهادی به شناسایی نوع نمونه ورودی با مدل دسته‌بندی A، تأثیر محاسبه دقت نهایی در این مدل دو برابر مدل‌های B و C در نظر گرفته شده است.

عملگرهای ژنتیکی استفاده‌شده برای انتخاب، آمیزش و جهش به ترتیب تورنومنت، تک‌نقطه‌ای و تغییر بیت هستند. عملگر جهش تغییر بیت، محتوای نمونه‌هایی که باعث ایجاد خطا در مدل‌های دسته‌بندی A، B و C شده‌اند را با احتمال M_R تغییر می‌دهد.

٤- بحث و نتایج

شبیه‌سازی با نرم‌افزار Matlab ورژن ۲۰۱۶a روی پایگاه داده ویسکانسین انجام شده است. پایگاه داده ویسکانسین متشکل از سه مجموعه داده WBCD، WDBC و WPBC می‌باشد [۲۶]. به منظور حصول اطمینان از نتایج، میانگین ۱۵ بار اجرای متمایز در تمام آزمایش‌ها لحاظ شده است. برای کار دسته‌بندی در روش پیشنهادی الگوریتم‌های ID۳ و SVM مقایسه شده‌اند. همچنین معیارهای دقت[1]، حساسیت[2] و ویژگی[3] برای ارزیابی روش پیشنهادی بر مبنای اعتبارسنجی ۱۰-fold استفاده شده است. روابط زیر روش محاسبه این معیارها را نشان می‌دهد

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (۳)$$

$$Sensitivity = \frac{TP}{TP + FN} \qquad (۴)$$

$$Specificity = \frac{FP}{TN + FP} \qquad (۵)$$

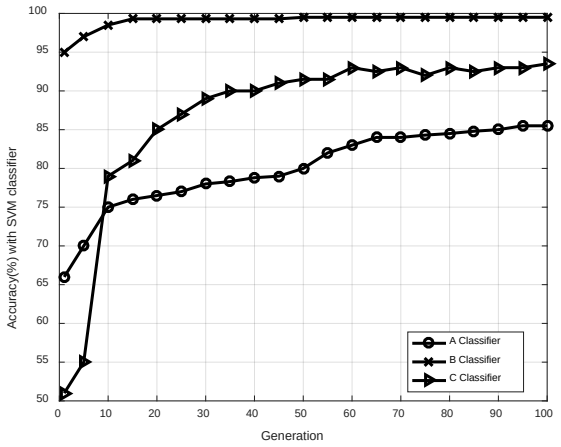
که نمادهای TP مثبت حقیقی، TN مثبت کاذب، FP منفی حقیقی و FN منفی کاذب را با توجه به ماتریس بی‌نظمی نشان می‌دهند.

نتایج مقایسه دو مدل ID۳ و SVM برای کار دسته‌بندی روش پیشنهادی در جدول ۱ نشان داده شده است. معیارهای مورد بررسی تعداد ویژگی‌های انتخابی، دقت بدون/ با انتخاب ویژگی و دقت نهایی مدل پیشنهادی می‌باشد. تعداد ویژگی‌های انتخابی با توجه به میانگین ۱۵ اجرای مستقل گزارش شده و بنابراین مقادیر این معیار به صورت اعشاری گزارش می‌شود. در این مقایسه حالات مختلف استفاده از مدل ID۳ و SVM در بخش انتخاب ویژگی بررسی شده و برای مدل‌های A، B و C دسته‌بندی یکسانی به کار رفته است.

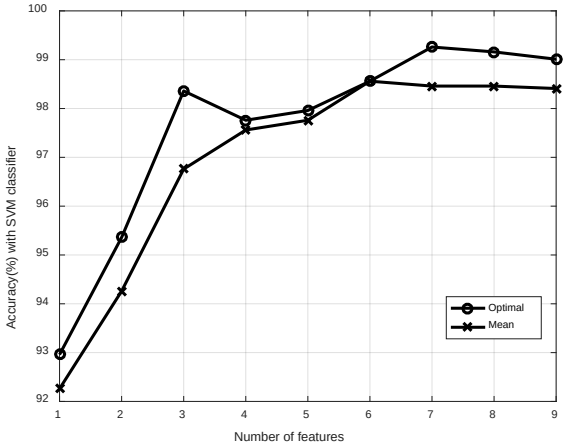
نتایج این آزمایش نشان می‌دهد که روش پیشنهادی با در نظر گرفتن انتخاب ویژگی در هر سه مجموعه داده و با استفاده از هر دو مدل دسته‌بندی عملکرد بهتری نسبت به حالت بدون انتخاب ویژگی دارد. بهترین نتایج زمانی است که در هر دو بخش انتخاب ویژگی و دسته‌بندی ترکیبی از SVM استفاده شده است. در این حالت دقت میانگین روش پیشنهادی برای مجموعه داده‌های WBCD، WDBC و WPBC به ترتیب ۹۹/۲۶٪، ۹۸/۵۵٪ و ۹۸/۴۵٪ است.

با توجه به برتری مدل SVM، در ادامه آزمایش‌ها از این روش برای کار دسته‌بندی استفاده می‌شود. در عین حال معیارهای دقت، حساسیت و ویژگی در جدول ۲ به ازای استفاده از مدل دسته‌بندی SVM در دو بخش انتخاب ویژگی و دسته‌بندی ترکیبی ارائه شده است.

1. Accuracy
2. Sensitivity
3. Specificity



شکل ۳: عملکرد سه مدل دسته‌بندی A، B و C در روش پیشنهادی.



شکل ۴: دقت با تعداد ویژگی‌های مختلف در روش پیشنهادی.

جدول ۱: نتایج روش پیشنهادی با مقایسه مدل‌های دسته‌بندی ID۳ و SVM.

مجموعه داده	تعداد کل ویژگی‌ها	مدل انتخاب ویژگی	مدل دسته‌بندی ترکیبی	دقت بدون انتخاب ویژگی (٪)	تعداد ویژگی‌های انتخابی	دقت مدل در انتخاب ویژگی (٪)	دقت نهایی مدل (٪)
WBCD	۹	ID۳	ID۳	۹۳٫۸۵	۷٫۹	۹۴٫۱۳	۹۶٫۵۵
			SVM	۹۷٫۲۳	۷٫۱	۹۵٫۶۶	۹۹٫۰۶
		SVM	ID۳	۹۳٫۸۵	۵٫۰	۹۵٫۶۶	۹۸٫۹۰
			SVM	۹۷٫۲۳	۵٫۳	۹۶٫۶۴	۹۹٫۲۶
WDBC	۳۰	ID۳	ID۳	۹۱٫۱۳	۱۴٫۳	۹۳٫۲۳	۹۶٫۲۳
			SVM	۹۳٫۴۵	۱۵٫۵	۹۵٫۱۴	۹۸٫۴۷
		SVM	ID۳	۹۲٫۱۳	۱۰٫۶	۹۵٫۴۵	۹۷٫۹۷
			SVM	۹۵٫۴۵	۱۲٫۲	۹۷٫۲۴	۹۸٫۵۵
WPBC	۳۲	ID۳	ID۳	۸۸٫۵۶	۲۳٫۲	۷۵٫۱۳	۹۷٫۰۱
			SVM	۹۲٫۱۴	۲۶٫۵	۸۱٫۳۷	۹۸٫۲۹
		SVM	ID۳	۸۸٫۵۶	۲۶٫۱	۸۰٫۴۳	۹۸٫۱۳
			SVM	۹۲٫۱۴	۲۱٫۸	۸۳٫۳۹	۹۸٫۴۵

جدول ۲: عملکرد الگوریتم پیشنهادی در معیارهای مختلف.

مجموعه داده	دقت (٪)	حساسیت (٪)	ویژگی (٪)
WBCD	۹۹٫۲۶	۹۸٫۰۱	۹۹٫۵۹
WDBC	۹۸٫۵۵	۹۷٫۲۱	۹۹٫۰۰
WPBC	۹۸٫۴۵	۹۷٫۸۸	۹۸٫۹۴

جدول ۳: مقایسه دقت تشخیص سرطان پستان در روش‌های مختلف.

روش‌ها	WBCD	WDBC	WPBC
[۲۷] ANN	–	۹۴٫۱۵	۸۳٫۰۵
[۲۸] SMO, IBK, NB, MLP	۹۷٫۱۳	۹۷٫۵۳	۹۷٫۳۱
[۲۸] SMO, IBK, NB, J۴۸	۹۷٫۲۸	۹۷٫۰۱	۹۴٫۲۲
[۱۷] EM-PCA, Fuzzy	۹۳٫۲۰	۹۴٫۱۰	–
[۲۱] Fuzzy, KNN	۹۹٫۷۱	–	–
[۴] SVM	–	۹۶٫۸۵	–
[۲۱] GAANN_RP	۹۸٫۲۹	–	–
[۲۰] PSO-KDE	۹۸٫۵۳	۹۸٫۴۵	–
روش پیشنهادی	۹۹٫۲۶	۹۸٫۵۵	۹۸٫۴۵

شکل ۳ نشان می‌دهد که مدل دسته‌بندی A در فرایند بهینه‌سازی الگوریتم ژنتیک نسبت به مدل‌های B و C عملکرد ضعیف‌تری دارد. دلیل این امر چالش تخصیص کلاس جدید به نمونه‌هایی با مفهوم متعارف و نامتعارف است که بدون در نظر گرفتن کلاس واقعی آنها انجام می‌شود. با اعمال جداسازی نمونه‌ها، دقت تشخیص دو بخش متعارف و نامتعارف بر اساس مدل‌های B و C عملکرد نسبتاً مناسبی دارد. از این رو به نظر می‌رسد ضعف روش حاضر ایجاد یک مدل مناسب برای جداسازی و یا به طور خاص تشخیص نمونه‌های نامتعارف است.

استفاده از تکنیک طول رشته متغیر در الگوریتم ژنتیک علاوه بر انتخاب ویژگی‌های مؤثر، تعداد بهینه این ویژگی‌ها را نیز مشخص می‌کند. شکل ۴ دقت الگوریتم پیشنهادی را با تعداد مختلف ویژگی‌ها گزارش می‌دهد. محاسبه دقت برای تعداد ویژگی‌های مختلف در ۲ حالت میانگین و بهترین روی مجموعه داده WBCD نشان می‌دهد که بهترین دقت دسته‌بندی با ۷ ویژگی ۹۹٫۲۶٪ می‌باشد.

به منظور بررسی بهتر عملکرد روش پیشنهادی، نتایج حاصل‌شده را با جدیدترین روش‌های ارائه‌شده در این زمینه مقایسه می‌کنیم. جدول ۳ نتایج این مقایسه را برای سه مجموعه داده WBCD، WDBC و WPBC نشان می‌دهد.

روش تشخیص سرطان پستان پیشنهادی در مقایسه با سایر روش‌های مورد بررسی و به ازای برخی از مجموعه داده‌ها دقت بیشتری داشته و در سایر موارد نیز دقت مناسبی را ارائه می‌دهد. به طور کلی به نظر می‌رسد که این روش با توجه به تشخیص نمونه‌های نامتعارف، روی مجموعه داده‌هایی که احتمال وجود نویز در آنها بیشتر است کارایی بهتری خواهد

داشت، هر چند این موضوع به طور کامل در این تحقیق بررسی نشده است. روش پیشنهادی از نظر حجم پردازش با اغلب روش‌های مورد مقایسه در جدول ۳ قابل بررسی نیست به این دلیل که در اغلب این روش‌ها از فرایند بهینه‌سازی در بخش انتخاب ویژگی و دسته‌بندی استفاده نمی‌شود.

۵- نتیجه‌گیری و پیشنهادها

در علم پزشکی به علت وجود تعداد پارامترهای متعدد در تشخیص بیماری، تشخیص برای یک متخصص خبره نیز به سختی امکان‌پذیر است. به این علت در چند دهه اخیر نیاز به تجزیه و تحلیل روابط بین داده‌های این حوزه، سبب استفاده از فناوری‌های نوینی از جمله داده‌کاوی در پیشگیری، تشخیص و درمان افراد شده است. در پژوهش حاضر با بررسی مطالعات انجام‌گرفته درباره پیش‌بینی بیماری سرطان پستان، یک مدل دسته‌بندی ترکیبی بر مبنای جداسازی نمونه‌های متعارف و نامتعارف ارائه شده است. با شناسایی نمونه‌های نامتعارف که باعث ایجاد خطا می‌شوند می‌توان عملکرد کلی مدل‌های دسته‌بندی را بهبود داد. برای کارهای آینده استفاده از روش‌های ابتکاری برای ایجاد جمعیت اولیه در تشخیص نمونه‌های متعارف و نامتعارف پیشنهادی می‌شود.

مراجع

[1] A. G. Freifeld, *et al*., "Clinical practice guideline for the use of antimicrobial agents in neutropenic patients with cancer: 2010 update by the Infectious Diseases Society of America," *Clinical Infectious Diseases*, vol. 52, no. 4, pp. 56-93, Feb. 2011.

[2] K. J. Cios and G. W. Moore, "Uniqueness of medical data mining," *Artificial Intelligence in Medicine*, vol. 26, no. 1, pp. 1-24, Sept. 2002.

[3] R. Shen, Y. Yang, and F. Shao, "Intelligent breast cancer prediction model using data mining techniques," in *Proc.6th Int. Conf. on Intelligent Human-Machine Systems and Cybernetics, IHMSC'14*, pp. 384-387, Hangzhou, China, 26-27 Aug. 2014.

[4] S A. R. M. Al-shamasneh, ans U. H. B. Obaidellah, "Artificial intelligence techniques for cancer detection and classification: review study," *European Scientific Journal*, vol. 13, no. 3, pp. 342-370, Jan. 2017.

[5] A. K. Sampath and N. Gomathi, "Probabilistic model based hybrid classifier for character recognition," *International J. of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 25, no. 4, pp. 621-647, Aug. 2017.

[۶] م. عبدالرزاق‌نژاد، "طبقه‌بندی و شناسایی وب‌سایت‌های فیشینگ به کمک مجموعه قوانین فازی و الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار،" *نشریه مهندسی برق و مهندسی کامپیوتر ایران، ب- مهندسی کامپیوتر*، جلد ۱۴، شماره ۳، صص. ۳۲۱-۳۱۱، پاییز ۱۳۹۵.

[۷] ه. صدوقی یزدی، ع. محی‌الدینی شاهم‌آبادی‌پور و م. خادمی، "طبقه‌بند خودسازمانده هندسی مبتنی بر یادگیری فعال برای نهان‌کاوی در محیط ویدئو با صرف حداقل برچسب،" *نشریه مهندسی برق و مهندسی کامپیوتر ایران، ب- مهندسی کامپیوتر*، جلد ۱۶، شماره ۱، صص. ۴۰-۲۸، بهار ۱۳۹۷.

[۸] ز. مروج و ج. آذرخش، "شبیه‌سازی و طبقه‌بندی وقایع کیفیت توان با استفاده از شبکه عصبی،" *فصلنامه مدل‌سازی در مهندسی*، جلد ۱۳، شماره ۴۱، صص. ۱۳۷-۱۴۶، تابستان ۱۳۹۴.

[۹] ر. شیخ‌پور و م. آقاصرام، "انتخاب ویژگی‌های مؤثر در تشخیص سرطان سینه با استفاده از مدل‌های پارامتریک یادگیری ماشین،" *فصلنامه علمی- پژوهشی بیماری‌های سینه*، جلد ۸، شماره ۲، صص. ۱۶-۲۳، تابستان ۱۳۹۴.

[۱۰] ا. صادقی‌پور، ن. ا. صحراگرد، م. ر. سایبانی و ز. بهمن‌زاده، "تشخیص سرطان سینه بر اساس رویکرد ترکیبی مبتنی بر الگوریتم کرم شب‌تاب و ترکیب سیستم‌های هوشمند،" *مجموعه مقالات کنفرانس بین‌المللی مهندسی، ICOAC، هنر و محیط زیست، کشور لهستان*، صص. ۳۱-۲۴، پاییز ۱۳۹۳.

[11] S. K. Mandal, "Performance analysis of data mining algorithms for breast cancer cell detection using Naive Bayes, logistic regression and decision tree," *International J. of Engineering and Computer Science*, vol. 6, no. 2, pp. 20388-20391, Feb. 2017.

[12] A. Onan, "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*, vol. 42, no. 20, pp. 6844-6852, Nov. 2015.

[13] B. G. Patil and S. N. Jain, "Cancer cells detection using digital image processing methods," *International J. of Latest Trends in Engineering and Technology*, vol. 3, no. 4, pp. pp. 45-49, Mar.2014.

[14] M. Nilashi, O. Ibrahim, H. Ahmadi, and L. Shahmoradi, "A knowledge-based system for breast cancer classification using fuzzy logic method," *Telematics and Informatics*, vol. 34, no. 4, pp. 133-144, Jul. 2017.

[15] K. J. Wang, B. Makond, K. H. Chen, and K. M. Wang, "A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients," *Applied Soft Computing*, vol. 20, pp. 15-24, Jul. 2016.

[16] R. D. H. Devi and M. I. Devi, "Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer," *International Journal of Advanced Engineering Technology*, vol. 93, no. 2, pp. 93-98, Apr. 2016.

[17] J. Diz, G. Marreiros, and A. Freitas, "Applying data mining techniques to improve breast cancer diagnosis," *J. of Medical Systems*, vol. 40, no. 9, pp. 203-209, Aug. 2016.

[18] S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier," in *1st IEEE Int. Conf. on Automation, Control, Energy and Systems, ACES'14*, 4 pp., Hooghy, India, 1-2 Feb. 2014.

[19] K. Vaidehi and T. S. Subashini, "Breast tissue characterization using combined K-NN classifier," *Indian J. of Science and Technology*, vol. 8, no. 1, pp. 23-26, Jan. 2015.

[20] R. Sheikhpour, M. A. Sarram, and R. Sheikhpour, "Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer," *Applied Soft Computing*, vol. 40, no. C, pp. 113-131, Mar. 2016.

[21] F. Ahmad, N. A. M. Isa, Z. Hussain, M. K. Osman, and S. N. Sulaiman, "A GA-based feature selection and parameter optimization of an ANN in diagnosing breast cancer," *Pattern Analysis and Applications*, vol. 18, no. 4, pp. 861-870, Nov. 2015.

[22] M. A. Hall, "Correlation-based feature selection of discrete and numeric class machine learning," in *Proc. of the 17th Int.Conf. on Machine Learning, ICML'00*, pp. 359-366, 29 Jun.-2 Jul. 2000.

[23] J. L. J. Laredo, S. S. Nielsen, G. Danoy, P. Bouvry, and C. M. Fernandes, "Cooperative selection: improving tournament selection via altruism," in *Proc. European Conf. on Evolutionary Computation in Combinatorial Optimization*, pp. 85-96, Apr. 2014.

[24] M. Bulmer, "The effect of selection on genetic variability," *Tthe American Naturalist*, vol. 105, no. 943, pp. 201-211, May 1971.

[25] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, "Towards credible evaluation of anomaly based intrusion detection methods," *IEEE Trans. on System, Man and Cybernetics, Part C, Applications and Reviews*, vol. 40, no. 5, pp. 516-524, Sept. 2010.

[26] Breast Cancer Wisconsin (Original) dataset, *UCI Machine Language Repository*, 1992.

[۲۷] س. زنگنه، ر. جوانمرد، ع. تپه و م. م. عبادزاده، "رویکرد ترکیبی برای کاهش ابعاد ویژگی‌های مجموعه‌های داده‌ای با استفاده از الگوریتم ترکیبی شبکه عصبی و الگوریتم ژنتیک در تشخیص پزشکی،" *مجموعه مقالات سومین کنفرانس داده‌کاوی*، صص. ۴۶-۳۶، تهران، پاییز ۱۳۸۸.

[28] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis," in *Proc. 7th IEEE Int. Conf. on Computer Engineering & Systems, ICCES'12*, pp. 180-185, Cairo, Egypt, 27-29 Nov.2012..

امین رضایی‌پناه تحصیلات خود را در مقاطع کارشناسی مهندسی نرم‌افزار کامپیوتر در سال ۱۳۸۹ از موسسه آموزش عالی فردوس مشهد و کارشناسی ارشد مهندسی هوش مصنوعی کامپیوتر خود را در سال ۱۳۹۲ از دانشگاه شیراز به پایان رسانده است و هم‌اکنون استاد دانشگاه فنی و حرفه‌ای و موسسه آموزش عالی رهجویان دانش برازجان می‌باشد. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از: رباتیک، الگوریتم‌های بهینه‌سازی، داده‌کاوی، روش‌های زمانبندی و کاربردهای آن، شبکه‌های اجتماعی و محاسبات ابری.

حسام واقع‌بین تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی نرم‌افزار کامپیوتر به‌ترتیب در سال‌های ۱۳۹۴ و ۱۳۹۷ از دانشگاه آزاد اسلامی واحد بوشهر دریافت نمود. از سال ۱۳۸۹ الی ۱۳۹۸ نام‌برده به عنوان کارشناس سیستم‌های کامپیوتری و شبکه در اداره کل اسناد و املاک استان بوشهر ایران به کار مشغول بود. زمینه‌های علمی مورد علاقه نام‌برده متنوع بوده و شامل موضوعاتی مانند اینترنت اشیاء، خلاصه‌سازی متن، پردازش تصویر، داده‌کاوی و پردازش موازی می‌باشد.