تشخیص سرطان سینه با استفاده از یک مدل طبقه‌بندی ترکیبی مبتنی بر جداسازی نمونه‌های متعارف و نامتعارف

نویسندگان: امین رضایی پناه- amin.rezaeipanah@gmail.com
حسام واقع‌بین- hwhesam@gmail.com

چکیده: پیش‌بینی و تشخیص بیماری‌های مختلف دو مؤلفه تأثیرگذار علوم پزشکی می‌باشند. اخیراً با گسترش روزافزون علم، استفاده از سیستم‌های پشتیبان تصمیم می‌تواند کمک زیادی در سیاست‌های درمانی پزشک داشته باشد. بدین منظور استفاده از سیستم‌های هوشمند مصنوعی در پیش‌بینی و تشخیص سرطان سینه که یکی از رایج‌ترین سرطان‌ها در بین زنان می‌باشد، موردتوجه است. در این تحقیق فرآیند تشخیص بیماری سرطان سینه در یک رویکردی دومرحله‌ای انجام می‌شود. در مرحله ویژگی‌های مؤثر در پیش‌بینی بیماری سرطان با یک الگوریتم ژنتیک بهبودیافته مبتنی بر طول رشته متغیر تشخیص داده می‌شوند. سپس تشخیص نمونه‌های متعارف و نامتعارف به‌منظور افزایش دقت و ایجاد مدل طبقه‌بندی نهایی در مرحله دوم انجام می‌شود. برای کار طبقه‌بندی از مدل‌های کلاسیک درخت تصمیم (*ID3*) و ماشین بردار پشتیبان (*SVM*) استفاده‌شده است. نتایج حاصل از آزمایش‌ها برتری روش پیشنهادی را با مدل طبقه‌بندی *SVM* در مقایسه با سایر روش‌های مشابه نشان می‌دهد. میانگین دقت تشخیص بیماری سرطان سینه درروش پیشنهادی روی مجموعه داده‌های *WBCD*، *WDBC* و *WPBC* با مدل *SVM* به ترتیب ۹۹,۲۶٪، ۹۸,۵۵٪ و ۹۸,۴۵٪ گزارش‌شده است.

کلیدواژه: الگوریتم ژنتیک، سرطان سینه، ویژگی‌های مؤثر، نمونه‌های متعارف، نمونه‌های نامتعارف.

۱ - مقدمه

سرطان سینه یکی از سرطان‌های شایع در زنان می‌باشد. احتمال ابتلا به سرطان سینه با بالا رفتن سن افزایش می‌یابد، بطوریکه ۸۰٪ موارد بعد از سن ۵۰ سالگی اتفاق می‌افتد. سرطان سینه در ابتدا به‌صورت یک غده ریز و کوچک می‌باشد و چندین سال باید طول بکشد تا یک غده سرطانی ریز تبدیل به یک غده سرطانی بزرگ شود و بافت‌های اطراف را درگیر کند. تشخیص زودرس غده‌های کوچک باعث می‌شود شانس موفقیت عمل جراحی از ۲۷٪ به ۹۸٪ افزایش یابد[۱]. آزمایش آسپیراسیون سوزنی (FNA) روشی ارزان و غیرتهاجمی برای تشخیص دقیق سرطان سینه می‌باشد. در این روش پس از استخراج خصوصیات سیتولوژی بیمار از بافت سینه، خوش‌خیم یا بدخیم بودن تومور تشخیص داده می‌شود[۲].

در سال‌های اخیر استفاده از روش‌های داده‌کاوی برای کمک به پزشکان و متخصصین در تشخیص زودهنگام این بیماری موردتوجه قرارگرفته است[۳]. پژوهش‌های انجام‌شده در این زمینه، موفق آمیز بودن سیستم‌های هوش مصنوعی را اثبات کرده است[۴]. فرایند تشخیص در این پژوهش‌ها با استفاده از استخراج ویژگی‌ها از آزمایش آسپیراسیون سوزنی و با کمک الگوریتم‌های داده‌کاوی انجام می‌شود. در این میان مطالعات انجام‌شده نشان می‌دهند که روش‌های ترکیب طبقه‌بندی‌ها به یک ابزار مؤثر به‌منظور افزایش کارایی و میزان دقت تبدیل‌شده است[۵]. در سال‌های اخیر روش‌ها و ایده‌های مختلفی برای ترکیب نتایج مدل‌های طبقه‌بندی نظیر شبکه‌های عصبی و سیستم‌های فازی ارائه‌شده است[۶-۸].

در این تحقیق از ترکیب چندین مدل طبقه‌بندی مختلف به‌منظور افزایش دقت در داده‌های سرطان سینه استفاده می‌شود. برای این منظور داده‌های آموزشی را به نحوی بین مدل‌های طبقه‌بندی تقسیم می‌کنیم که حداکثر دقت حاصل شود. بنابراین چالش سیستم پیش‌رو نحوه تقسیم‌بندی داده‌های آموزشی و شناسایی نمونه‌های نامتعارف است. نمونه‌های نامتعارف، نمونه‌هایی هستند که در یک مدل طبقه‌بندی مستقل باعث ایجاد خطا می‌شوند.

www.kahrobaonline.ir

ادامه این تحقیق بدین شرح است؛ در بخش ۲ به بررسی برخی از جدیدترین کارهای انجام‌شده در تشخیص سرطان سینه می‌پردازیم. روش پیشنهادی در بخش ۳ مطرح‌شده و نتایج حاصل از آن در بخش ۴ آورده شده است. درنهایت در بخش ۵ نتیجه‌گیری و پیشنهادها بیان می‌شود.

۲- پیشینه تحقیق

در زمینه پیش‌بینی و تشخیص سرطان سینه با مدل‌های طبقه‌بندی تحقیقات متعددی صورت گرفته است. شیخ‌پور و همکاران، انتخاب ویژگی‌های مؤثر در تشخیص سرطان سینه را با استفاده از مدل‌های پارامتریک یادگیری ماشین معرفی کردند[۹]. صادقی‌پور و همکاران، تشخیص سرطان سینه را بر اساس رویکرد ترکیبی مبتنی بر الگوریتم کرم شب‌تاب و یک سیستم هوشمند ترکیبی ارائه دادند[۱۰]. سیستم هوشمند ارائه‌شده ترکیبی از الگوریتم تکاملی و مدل SVM به‌منظور انتخاب بهترین ویژگی‌ها و طبقه‌بندی است.

نیلاشی و همکاران، تشخیص سرطان سینه را با استفاده از روش منطق فازی توسعه دادند[۱۱]. قوانین فازی با استفاده از خوشه‌بندی حداکثر انتظار، حذف نویز و طبقه‌بندی CART تولید می‌شوند. دوی و همکاران، به‌منظور افزایش کارایی طبقه‌بندی J48 در تشخیص سرطان سینه، ویژگی‌ها را با استفاده از الگوریتم Farthest First گروه‌بندی کرده و سپس از هر گروه مؤثرترین ویژگی‌ها را انتخاب می‌کند[۱۲]. این روش به‌دقت حدود ۹۹٪ روی مجموعه داده WBCD رسیده است.

وانگ و همکاران، ترکیبی از روش های SMOTE (پیش نمونه جزئی) و PSO (بهینه‌سازی ازدحام ذرات) را برای تشخیص بیماران مبتلا به سرطان سینه بکار بردند[۷]. در این روش برخی از مشهورترین روش‌های طبقه‌بندی مانند رگرسیون لجستیک، درخت تصمیم‌گیری C5 و نزدیک‌ترین همسایه ادغام می شوند. نتایج بهترین عملکرد را از ترکیب SMOTE+PSO+C5 نشان می‌دهد.

دیز و همکاران، یک رویکرد مبتنی بر داده کاوی برای غده شناسی در سرطان سینه ارائه دادند[۱۳]. این روش با استفاده از طبقه‌بندی چگالی سینه به شناسایی تمایزات در مجموعه داده‌ها کمک می‌کند. وایدهی و همکاران، تشخیص سرطان سینه را با استفاده از طبقه‌بندی KNN ترکیبی ارائه دادند[۱۴]. KNN با سه ماتریس فاصله متفاوت یوکلیدان، کوزین، CITY-BLOCK و ترکیبات آن‌ها برای طبقه‌بندی مورداستفاده قرار می‌گیرد. آونان، یک طبقه‌بندی نزدیک‌ترین مجاور گنگ-سخت ترکیبی را با زیرمجموعه سازگار و گزینش موردی برای تشخیص اتوماتیک سرطان سینه مورد ارزیابی قرار می‌دهد[۱۵]. شن و همکاران، مدل هوشمند پیش‌بینی سرطان سینه با استفاده از تکنیک‌های داده‌کاوی را ارائه دادند[۳]. در این تحقیق از یک روش خصیصه انتخابی(Interact) به‌منظور انتخاب ویژگی های مربوطه برای تشخیص سرطان سینه و از ماشین بردار پشتیبان برای ساخت مدل طبقه‌بندی استفاده می‌شود.

۳- روش پیشنهادی

روش پیشنهادی ترکیبی بوده و شامل چندین مرحله است. در مرحله اول پیش‌پردازش داده‌ها انجام می‌شود. انتخاب ویژگی‌های مؤثر با استفاده از بهبود الگوریتم ژنتیک در مرحله دوم انجام می‌شود. درنهایت مدل طبقه‌بندی در مرحله سوم ارائه می‌شود. برای طبقه‌بندی داده‌ها از سیستم طبقه‌بندی ترکیبی جدیدی استفاده می‌شود که مبتنی بر شناسایی نمونه‌های متعارف و نامتعارف است.

۳-۱- پیش‌پردازش داده‌ها

پیش‌پردازش داده‌ها به‌منظور آماده‌سازی داده‌ها جهت پردازش و همچنین بهبود کیفیت داده‌های واقعی انجام می‌شود. هنگامی‌که مقادیر ویژگی در دامنه متفاوتی قرار داشته باشند آن‌ها را در دامنه مشابهی قرار می‌دهد. این امر معمولاً به کاهش هزینه محاسباتی و افزایش عملکرد طبقه‌بندی کمک می‌کند. در این تحقیق از روش مینیمم – ماکزیمم برای نرمال‌سازی استفاده می‌شود.

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

در این رابطه $x_{i,j}$ مقدار j-مین ویژگی از i-مین نمونه مجموعه داده است. $\min(x_j)$ و $\max(x_j)$ به ترتیب کوچک‌ترین و بزرگ‌ترین مقدار j-مین ویژگی از تمتم نمونه‌ها است.

۳-۲- انتخاب ویژگی‌های مؤثر با الگوریتم ژنتیک بهبودیافته

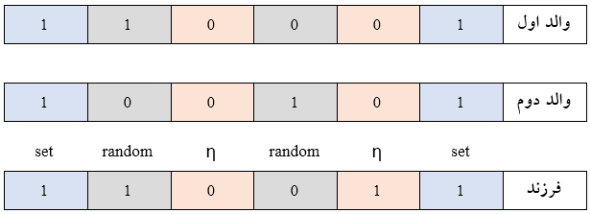
یکی از اهداف اصلی در مسئله انتخاب ویژگی حذف ویژگی‌های نامرتبط و همچنین دارای افزونگی است[۱۶]. روش استفاده‌شده

www.kahrobaonline.ir

در این تحقیق از تکنیک طول رشته متغیر در الگوریتم ژنتیک استفاده می‌کند که علاوه بر انتخاب ویژگی‌های مؤثر، تعداد آن‌ها را نیز به‌صورت خودکار مشخص می‌کند.

ساختار کروموزوم پیشنهادی به‌صورت رشته‌ای باینری به طول تعداد کل ویژگی‌ها تعیین می‌شود. ژن ۱ به مفهوم انتخاب و ژن ۰ به مفهوم عدم انتخاب ویژگی است. اگر از نماد DNF برای نمایش تعداد ویژگی‌های مطلوب استفاده شود، مقدار این پارامتر با توجه به تعداد ژن‌های ۱ کروموزوم مشخص می‌شود. در تولید جمعیت اولیه از مقادیر مختلف DNF استفاده می‌شود که درنهایت این امر منجر به تشخیص تعداد ویژگی‌های مطلوب خواهد شد. برای کار طبقه‌بندی از دو الگوریتم ID3 و SVM استفاده می‌شود. هدف مقایسه عملکرد این دو مدل روی‌داده‌های تشخیص سرطان است.

عملگرهای ژنتیکی استفاده‌شده در این تحقیق انتخاب، ترکیب و جهش است. برای عملگر انتخاب از روش تورنمنت استفاده‌شده است[۱۷]. کروموزوم‌هایی که از قسمت تورنمنت به‌عنوان والد انتخاب شدند، با استفاده از عملگر ترکیب ژن‌هایشان را به‌احتمال C_R مبادله می‌کنند. در عملگر ترکیب ارائه‌شده، ژن‌های مشترک از هر دو والد در کروموزوم فرزند کپی شده و سایر ژن‌ها با احتمال η از ویژگی‌های مشاهده نشده مقداردهی می‌شوند. ویژگی‌های مشاهده نشده، ویژگی‌هایی هستند که در هر دو کروموزوم والد وجود نداشته باشند. شکل ۱ نمونه‌ای از عملگر ترکیب پیشنهادی را نشان می‌دهد.



شکل ۱- فرایند عملگر ترکیب پیشنهادی

بعد از اعمال عملگر ترکیب، عملگر جهش با احتمال M_R روی هر ژن انجام می‌پذیرد و باعث تغییر محتوای هر ژن می‌شود. بعد از اتمام مراحل الگوریتم ژنتیک، ویژگی‌های انتخاب‌شده در بهترین عضو جمعیت (بالاترین دقت) به‌عنوان ویژگی‌های مطلوب و تعداد این ویژگی‌ها پارامتر DNF بهینه را مشخص می‌کند.

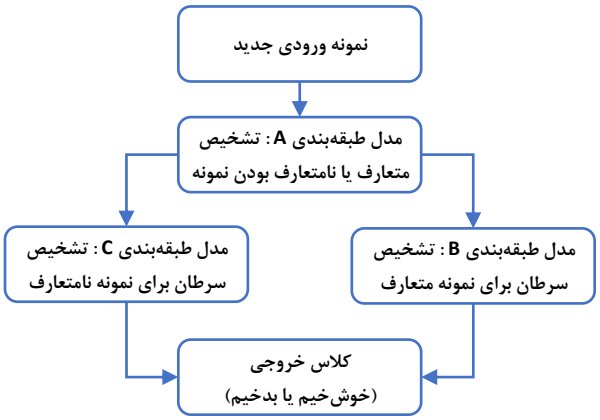
۳-۳- مدل طبقه‌بندی ترکیبی بر مبنای تقسیم‌بندی نمونه‌ها

روش‌های طبقه‌بندی با بررسی نمونه‌های یک مجموعه داده، یک الگوی پیش‌بینی را به‌صورت استقرایی مدل می‌کنند[۱۸]. در این تحقیق جهت تشخیص سرطان سینه یک مدل طبقه‌بندی ترکیبی بر مبنای تشخیص نمونه‌های متعارف و نامتعارف ارائه می‌شود.

در برخی مجموعه داده‌ها ممکن است مقادیر ویژگی‌های برخی از نمونه‌ها دارای حالات خاصی باشند[۱۹]. این نمونه‌ها با سایر نمونه‌های کلاس مشابه فاصله زیادی دارند و باعث اشتباه مدل‌های طبقه‌بندی می‌شوند. این نمونه‌ها، نامتعارف نامیده می‌شود و با شناسایی آن‌ها می‌تواند عملکرد یک مدل طبقه‌بندی را بهبود داد.

در این تحقیق برای تشخیص نمونه‌های نامتعارف از الگوریتم ژنتیک استفاده می‌شوند. الگوریتم ژنتیک نمونه‌ها را به دو بخش متعارف و نامتعارف تقسیم می‌کند. سپس یک مدل طبقه‌بندی کلاسیک بر اساس این تقسیم‌بندی ایجاد می‌شود. بنابراین نوع نمونه‌ها (متعارف یا نامتعارف) بر اساس این مدل تشخیص داده می‌شود. درنهایت تشخیص سرطان سینه بر مبنای نوع نمونه با یک روش طبقه‌بندی کلاسیک متمایز مدل می‌شود.

بنابراین از سه مدل طبقه‌بندی کلاسیک در پیشنهادی استفاده می‌شود؛ مدل طبقه‌بندی اول (A) برای تشخیص نوع نمونه (متعارف یا نامتعارف) است. مدل طبقه‌بندی دوم (B) و سوم (C) نیز به ترتیب برای تشخیص سرطان در نمونه‌های متعارف و نامتعارف می‌باشد. در فاز آزمایش، تعیین کلاس نمونه ورودی مطابق شکل ۲ می‌باشد.



شکل ۲- تعیین کلاس نمونه جدید

بررسی‌های انجام‌شده نشان‌دهنده کاهش کارایی مدل طبقه‌بندی ترکیبی ارائه‌شده در حالت آزمایش است. دلیل این امر کاهش نمونه‌ها در دو مدل طبقه‌بندی B و C به دلیل جداسازی می‌باشد. برای حل این مشکل نمونه‌های خنثی از مجموعه داده در هر مدل

www.kahrobaonline.ir

طبقه‌بندی B و C در نظر گرفته می‌شود. نمونه‌های خنثی، نمونه-هایی هستند که تقسیم‌بندی هم‌زمان آن‌ها به‌صورت متعارف و نامتعارف تأثیری در عملکرد نهایی مدل ندارد. در ادامه پروسه تقسیم‌بندی نمونه‌ها به دو بخش متعارف و نامتعارف با استفاده از الگوریتم ژنتیک تشریح می‌شود.

ساختار راه‌حل‌ها در الگوریتم ژنتیک پیشنهادی به‌صورت یک آرایه دوبعدی باینری به طول کل نمونه‌ها می‌باشد. سطر اول این آرایه، نمونه-های متعارفی هستند که با طبقه‌بندی B مدل می‌شوند. مقدار ۱ هر عنصر برای یک نمونه نشان‌دهنده تخصیص آن به نوع متعارف است. به‌طور مشابه سطر دوم نمونه‌های نامتعارف استفاده‌شده در مدل طبقه‌بندی C را مشخص می‌کند. محدودیت هر راه‌حل، انتخاب یک نمونه برای حداقل یک نوع نمونه (متعارف یا نامتعارف) است، یعنی جمع هر ستون در راه‌حل پیشنهادی حداقل باید ۱ باشد.
با توجه به استفاده از مدل طبقه‌بندی A برای شناسایی نوع نمونه ورودی، بنابراین عملکرد نهایی روش پیشنهادی به نحوه تقسیم-بندی نمونه‌ها و دقت این مدل وابستگی بالایی دارد. بدین منظور از رابطه (۲) برای محاسبه برازندگی هر راه‌حل و اعمال تأثیر بیشتر برای مدل طبقه‌بندی A استفاده می‌شود.

$$Fitness = \frac{2.A_{acc} + B_{acc} + C_{acc}}{4} \qquad (۲)$$

در این رابطه تأثیر دقت مدل طبقه‌بندی A دو برابر مدل‌های طبقه‌بندی B و C در نظر گرفته‌شده است.
در مرحله بعد از معیار تورنمنت به‌منظور انتخاب راه‌حل‌ها برای تکثیر استفاده می‌شود. عملگر ترکیب استفاده‌شده تک-نقطه‌ای می‌باشد که به‌احتمال C_R از دو والد انتخاب‌شده، دو فرزند تولید می‌کند. عملگر جهش پیشنهادی این تحقیق، محتوای ژن‌هایی (نمونه‌هایی) که باعث ایجاد خطا طبقه‌بندی (در هر سه مدل A، B و C) شده‌اند را به‌احتمال M_R تغییر می‌دهد. عملگر جهش باعث تغییر وضعیت ژن‌ها از ۰ به ۱ و بلعکس با توجه به محدودیت تعیین‌شده می‌شود.

جدول ۱- نتایج روش پیشنهادی با مقایسه مدل‌های طبقه‌بندی ID3 و SVM

مجموعه داده	تعداد کل ویژگی‌ها	مدل انتخاب ویژگی	مدل طبقه-بندی ترکیبی	دقت بدون انتخاب	تعداد ویژگی-های انتخاب‌شده	دقت مدل در انتخاب	دقت نهایی مدل(٪)
WBCD	۹	ID3	ID3	۹۳،۸۵	۷،۹	۹۴،۱۳	۹۶،۵۵
			SVM	۹۷،۲۳	۷،۱	۹۵،۶۶	۹۹،۰۶
		SVM	ID3	۹۳،۸۵	۵،۰	۹۵،۶۶	۹۸،۹۰
			SVM	۹۷،۲۳	۵،۳	۹۶،۶۴	**۹۹،۲۶**
WDBC	۳۰	ID3	ID3	۹۱،۱۳	۱۴،۳	۹۳،۲۳	۹۶،۲۳
			SVM	۹۳،۴۵	۱۵،۵	۹۵،۱۴	۹۸،۴۷
		SVM	ID3	۹۲،۱۳	۱۰،۶	۹۵،۴۵	۹۷،۹۷
			SVM	۹۵،۴۵	۱۲،۲	۹۷،۲۴	**۹۸،۵۵**
WPBC	۳۲	ID3	ID3	۸۸،۵۶	۲۳،۲	۷۵،۱۳	۹۷،۰۱
			SVM	۹۲،۱۴	۲۶،۵	۸۱،۳۷	۹۸،۲۹
		SVM	ID3	۸۸،۵۶	۲۶،۱	۸۰،۴۳	۹۸،۱۳
			SVM	۹۲،۱۴	۲۱،۸	۸۳،۳۹	**۹۸،۴۵**

نتایج نشان می‌دهد که روش پیشنهادی با در نظر گرفتن انتخاب ویژگی در هر سه مجموعه داده WBCD، WDBC و WPBC و با استفاده از هر دو مدل طبقه‌بندی ID3 و SVM عملکرد بهتری نسبت به حالت بدون انتخاب ویژگی دارد. بهترین نتایج زمانی است که در هر دو بخش انتخاب ویژگی و طبقه‌بندی ترکیبی از مدل SVM استفاده‌شده است. در این حالت دقت میانگین روش پیشنهادی برای مجموعه داده‌های WBCD، WDBC و WPBC به ترتیب ۹۹،۲۶٪، ۹۸،۵۵٪ و ۹۸،۴۵٪ گزارش‌شده است. درعین‌حال حساسیت، ویژگی و دقت در جدول ۲ به ازای استفاده از مدل طبقه‌بندی SVM در دو بخش انتخاب ویژگی و طبقه‌بندی ترکیبی ارائه‌شده است.

جدول ۲- عملکرد الگوریتم پیشنهادی در معیارهای مختلف

مجموعه داده	**حساسیت** (٪)	**ویژگی** (٪)	**دقت** (٪)
WBCD	۰۱.۹۸	۵۹.۹۹	۹۹،۲۶
WDBC	۲۱.۹۷	۹۹،۰	۹۸،۵۵
WPBC	۸۸.۹۷	۹۴.۹۸	۹۸،۴۵

بررسی‌های انجام‌شده نشان می‌دهد که در مدل طبقه‌بندی ترکیبی ارائه‌شده، مدل طبقه‌بندی A عملکرد ضعیف‌تری نسبت به مدل-های طبقه‌بندی B و C دارند. مدل طبقه‌بندی A به‌منظور

www.kahrobaonline.ir

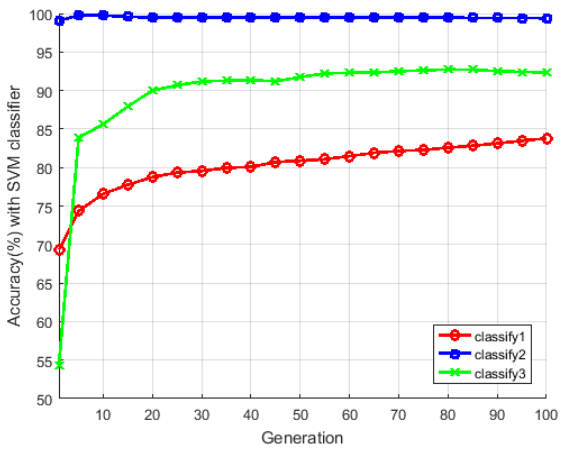
تشخیص جداسازی نمونه‌ها استفاده می‌شود. با جداسازی نمونه‌ها دقت تشخیص دو بخش متعارف و نامتعارف بر اساس دو مدل طبقه‌بندی B و C مناسب است. ازاین‌رو به نظر می‌رسد ضعف روش تشخیص سرطان پیشنهادی در مقایسه با سایر روش‌های موردبررسی و به ازای برخی از مجموعه داده‌ها دقت بیشتری داشته و در بقیه موارد نیز دقت مناسبی را ارائه می‌دهد. به‌طورکلی به نظر می‌رسد که روش ارائه‌شده با توجه به تشخیص نمونه‌های نامتعارف روی مجموعه داده‌هایی که احتمال وجود نویز در آن‌ها بیشتر است عملکرد بهتری خواهد داشت، هرچند این موضوع به‌طور کامل در این تحقیق بررسی نشده است.

جدول ۳- مقایسه دقت تشخیص سرطان سینه روش پیشنهادی با سایر روش‌ها

مجموعه داده	ANN [۲۱]	SMO, IBK, NB, MLP [۲۲]	SMO, IBK, NB, J48 [۲۲]	EM-PCA, Fuzzy [۱۱]	Fuzzy, KNN [۱۵]	SVM [۴]	روش پیشنهادی
WBCD	-	۹۷,۱۳	۹۷,۲۸	۹۳,۲۰	۹۹,۷۱	-	۹۹,۲۶
WDBC	۹۴,۱۵	۹۷,۵۳	۹۷,۰۱	۹۴,۱۰	-	۹۶,۸۵	۹۸,۵۵
WPBC	۸۳,۰۵	۹۷,۳۱	۹۴,۲۲	-	-	-	۹۸,۴۵

روش حاضر ایجاد یک مدل مناسب برای جداسازی و یا تشخیص نمونه‌های نامتعارف است.

شکل ۳ نشان می‌دهد که در فرایند بهینه‌سازی الگوریتم ژنتیک مدل طبقه‌بندی A نسبت به مدل‌های طبقه‌بندی 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. در اینجا classify1 نقش دقت مدل طبقه‌بندی A را در طول تکرار الگوریتم ژنتیک دارد.



شکل ۳- عملکرد سه مدل طبقه‌بندی A، B و C

به‌منظور ارزیابی عملکرد روش پیشنهادی، نتایج حاصل از آزمایش‌ها را در بهترین حالت (طبقه‌بندی SVM) با تعدادی از جدیدترین روش‌های ارائه‌شده در زمینه تشخیص سرطان سینه مقایسه می‌کنیم. جدول ۳ نتایج این مقایسه را نشان می‌دهد.

۵- نتیجه‌گیری

امروزه با توجه به حجم وسیع و پیچیدگی داده‌های موجود، بیش‌ازپیش نیاز به ابزاری کارا، مؤثر و مطمئن به‌منظور کشف دانش سودمند و موردنیاز حس می‌شود. داده‌کاوی ابزاری است که برای حصول به چنین دانشی ما را یاری می‌کند. یکی از زمینه‌های پرکاربرد داده‌کاوی در علم پزشکی است. در برخی مجموعه داده‌ها وجود نمونه‌های نامتعارف ممکن است باعث اشتباه مدل‌های طبقه‌بندی شوند. در این تحقیق برای تشخیص نمونه‌های نامتعارف از الگوریتم ژنتیک استفاده شد. الگوریتم ژنتیک نمونه‌ها را به دو بخش متعارف و نامتعارف تقسیم می‌کند. سپس یک مدل طبقه‌بندی کلاسیک بر اساس این تقسیم‌بندی ایجاد می‌شود. درنهایت تشخیص سرطان سینه بر مبنای نوع نمونه با یک روش طبقه‌بندی کلاسیک متمایز مدل می‌شود.

استفاده از روش‌های ابتکاری برای ایجاد جمعیت اولیه در تشخیص نمونه‌های متعارف از نامتعارف در ادامه راه این تحقیق پیشنهادی می‌شود. برای مثال می‌توان خروجی مدل طبقه‌بندی را برای ایجاد جمعیت اولیه استفاده کرد.

مراجع

[1] A. G. Freifeld, E. J. Bow, K. A. Sepkowitz, M. J. Boeckh, J. I. Ito, C. A. Mullen, ... and J. R. Wingard, "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, 2011, pp. 56-93.

[2] K. J. Cios, and G. W. Moore, "Uniqueness of medical data mining", Artificial intelligence in medicine, Vol. 26, NO. 1, 2002, pp. 1-24.

[3] R. Shen, Y. Yang, and F. Shao, "Intelligent breast cancer prediction model using data mining techniques", In Intelligent Human-Machine Systems and Cybernetics (IHMSC), Vol. 1, August 2014, pp. 384-387, IEEE.

[4] S. Ghosh, S. Mondal, and B. Ghosh, "A comparative study of breast cancer detection based on SVM and MLP BPN classifier", In Automation, Control, Energy and

Systems (ACES), 2014 First International Conference on (pp. 1-4), 2014, IEEE.

[5] A. K. Sampath, and N. Gomathi, “Probabilistic Model Based Hybrid Classifier for Character Recognition”, International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, Vol. 25, NO. 4, 2017, pp. 621-647.

[۶] مجید عبدالرزاق نژاد، "طبقه‌بندی و شناسایی وب سایت‌های فیشینگ به کمک مجموعه قوانین فازی و الگوریتم اصلاح‌شده بهینه‌سازی صفحات شیب‌دار"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، دوره ۱۴، شماره ۳، پاییز ۱۳۹۵، صفحه ۳۱۱-۳۲۱.

[۷] هادی صدوقی یزدی، علی محی الدینی شاهم آبادی پور و مرتضی خادمی، "طبقه‌بند خودسازمانده هندسی مبتنی بر یادگیری فعال برای نهان‌کاوی در محیط ویدئو با صرف حداقل برچسب"، نشریه مهندسی برق و مهندسی کامپیوتر ایران، دوره ۱۶، شماره ۱، بهار ۱۳۹۷، صفحه ۲۸-۴۰.

[۸] زهرا مروج و جواد آذرخش، "شبیه سازی و طبقه‌بندی وقایع کیفیت توان با استفاده از شبکه عصبی"، فصلنامه مدل سازی در مهندسی، دوره ۱۳، شماره ۴۱، تابستان ۱۳۹۴، صفحه ۱۳۷-۱۴۶.

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

[۱۰] احسان صادقی پور، نصرالله صحراگرد، محمودرضا سایبانی و زینب بهمن زاده، "تشخیص سرطان سینه بر اساس رویکرد ترکیبی مبتنی بر الگوریتم کرم شب تاب وترکیب سیستم های هوشمند"، کنفرانس بین المللی مهندسی(ICOAC)، هنر و محیط زیست، کشور لهستان، ۱۳۹۳.

[11] 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, 2017, pp. 133-144.

[12] R. D. H. Devi, and M. I. Devi, “Outlier detection algorithm combined with decision tree classifier for early diagnosis of breast cancer”, Int J Adv Engg Tech/Vol. VII/Issue II/April-June, Vol. 93, 2016, p. 98.

[13] J. Diz, G. Marreiros, and A. Freitas, “Applying Data Mining Techniques to Improve Breast Cancer Diagnosis”, Journal of medical systems, Vol. 40, NO. 9, 2016. p. 203.

[14] K. Vaidehi, and T. S. Subashini, “Breast tissue characterization using combined K-NN classifier”, Indian Journal of Science and Technology, Vol. 8, NO. 1, 2015, pp. 23-26.

[15] 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, 2015, pp. 6844-6852.

[16] M. A. Hall, “Correlation-based feature selection of discrete and numeric class machine learning”, ICML '00 Proceedings of the Seventeenth International Conference on Machine Learning, 2000, pp. 359-366.

[17] J. L. J. Laredo, S. S. Nielsen, G. Danoy, P. Bouvry, and C. M. Fernandes, “Cooperative selection: Improving tournament selection via altruism”, In European Conference on Evolutionary Computation in Combinatorial Optimization (pp. 85-96), 2014, Springer, Berlin, Heidelberg.

[18] M. Bulmer, “The effect of selection on genetic variability”, The American Naturalist, Vol. 105, NO. 943, 1971, pp. 201-211.

[19] M. Tavallaee, N. Stakhanova, and A. A. Ghorbani, “Towards credible evaluation of anomaly based intrusion detection methods”, IEEE Transaction on System, Man and Cybernetics, Part-c, Applications and Reviews, Vol. 40, NO. 5, 2010, pp. 516-524.

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

[۲۱] سکینه زنگنه، رضا جوانمرد، علی تپه و محمدمهدی عبادزاده، "رویکرد ترکیبی برای کاهش ابعاد ویژگیهای مجموعه های داده ای با استفاده ازالگوریتم ترکیبی شبکه عصبی و الگوریتم ژنتیک-در تشخیص پزشکی"، سومین کنفرانس داده‌کاوی، تهران، ۱۳۸۸.

[22] G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Experimental comparison of classifiers for breast cancer diagnosis." Computer Engineering & Systems (ICCES), 2012 Seventh International Conference on, 2012, IEEE.