تشخیص سرطان پستان با استفاده از برآورد ناپارامتری چگالی احتمال مبتنی بر روش‌های هسته‌ای

* **رباب شیخ پور:** گروه تربیت بدنی، واحد تفت، دانشگاه آزاد اسلامی، تفت، ایران و مرکز تحقیقات خون و آنکولوژی، دانشگاه علوم پزشکی شهید صدوقی، یزد، ایران (* نویسنده مسئول). r.sheikhpour@yahoo.com

راضیه شیخ پور: گروه مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران. r_sheikhpour@stu.yazd.ac.ir

تاریخ دریافت: ۹۴/۸/۲۰ تاریخ پذیرش: ۹۴/۱۲/۱۷

چکیده

زمینه و هدف: سرطان پستان شایع‌ترین سرطان در میان زنان است و وجود یک سیستم دقیق و مطمئن برای تشخیص به موقع و خوش‌خیم یا بدخیم بودن توده سرطان ضروری به نظر می‌رسد. با استفاده از نتایج سیتولوژی آسپیراسیون سوزنی و تکنیک‌های داده‌کاوی و یادگیری ماشین می‌توان روش‌های نوینی را برای شناسایی و تشخیص زود هنگام سرطان پستان ارائه کرد که با دقت بالایی قادر به تشخیص سرطان پستان باشند. هدف از انجام این مطالعه تشخیص سرطان پستان با استفاده از برآورد ناپارامتری چگالی احتمال مبتنی بر روش‌های هسته‌ای است.

روش کار: در این مطالعه از داده‌های پایگاه داده WBCD شامل ۶۹۹ نمونه خوش‌خیم و بدخیم با ۹ ویژگی و WDCB شامل ۵۶۹ نمونه خوش‌خیم و بدخیم با ۳۰ ویژگی استفاده شد و سپس به ارائه مدلی برای طبقه‌بندی مجموعه داده‌های WBCD و WDBC با استفاده از روش‌های تخمین چگالی مبتنی بر هسته پرداخته شد.

یافته‌ها: نتایج بررسی روش‌های غیر پارامتری بر پایگاه داده WDBC نشان داد که روش برآورد چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی با دقت ۹۷/۹۳٪ بالاترین دقت را در میان سایر روش‌ها داشت و روش‌های برآورد چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی و k نزدیکترین همسایه با دقت ۹۸/۱۷٪ بالاترین دقت را در میان سایر روش‌ها برای تشخیص سرطان پستان بر روی پایگاه داده WBCD داشتند.

نتیجه‌گیری: نتایج این مطالعه نشان داد که روش‌های ناپارامتری چگالی احتمال مبتنی بر روش‌های هسته‌ای می‌تواند با دقت بالایی برای تشخیص سرطان پستان به کار رود.

کلیدواژه‌ها: سرطان پستان، روش ناپارامتری، برآورد چگالی هسته‌ای، یادگیری ماشین

مقدمه

سرطان پستان شایع‌ترین سرطان در میان زنان است (۱، ۲). این بیماری به شدت ناهمگن است و در اثر تأثیر متقابل عامل‌های خطر وراثتی و محیطی ایجاد می‌شود و به تجمع پیشرونده تغییرات ژنتیک و اپی ژنتیک در سلول‌های سرطان پستان منجر می‌شود (۳). تقریبا ۲۵ درصد از مرگ و میرهای ناشی از سرطان پستان بین سنین ۴۹-۴۰ سال مشاهده می‌شود (۴). اگرچه شیوع این بیماری در سنین قبل از ۲۵ تا ۳۰ سالگی نادر است اما بروز این سرطان در سنین کمتر حتی در سن جوانی نیز گزارش شده است (۸-۵). بر اساس آمار سازمان بهداشت جهانی، از هر ۸ تا ۱۰ نفر یک زن به سرطان پستان دچار می‌شود. بر اساس آمارهای موجود در ایران از هر ۱۰ تا ۱۵ زن، احتمال ابتلای یک زن به سرطان پستان وجود دارد (۹). پژوهشگران میزان بالای مرگ و میر زنان بر اثر سرطان پستان را ناشی از تشخیص دیرهنگام این بیماری می‌دانند و موفقیت کشورهای پیشرفته در کنترل مرگ و میر و سایر پیامدهای ناشی از بیماری را در گرو تشخیص به موقع و زودهنگام آن دانسته‌اند، زیرا بقای فرد به طور مستقیم در ارتباط با مرحله بیماری در زمان تشخیص می‌باشد. میزان بقای پنج ساله در خانم‌هایی که سرطان آن‌ها در مراحل اولیه تشخیص داده شده، ۹۰ درصد است، در حالی‌که این میزان در خانم‌هایی‌که سرطان آن‌ها پیشرفت کرده است، به ۶۰ درصد کاهش یافته است (۱۰، ۱۱). بنابراین وجود یک سیستم دقیق و مطمئن برای تشخیص به موقع و خوش‌خیم بودن یا

بدخیم بودن توده سرطان ضروری به نظر می‌رسد (۱۲). اگرچه شایع‌ترین و قطعی‌ترین روش تشخیص سرطان پستان، بیوپسی سینه و تشخیص ضایعه با روش‌های معمول آسیب شناسی بافتی است (۱۲) ولی آزمایش آسپیراسیون سوزنی (FNA) روشی سرپایی، کم هزینه، آسان و سریع است. در این روش مایع استخراج شده از بافت پستان برای بررسی خصوصیات سیتولوژی در زیر میکروسکوپ مورد بررسی قرار می‌گیرد (۱۳، ۱۴). بعد از استخراج خصوصیات سیتولوژی بیمار باید بتوان خوش‌خیم یا بدخیم بودن توده را تشخیص داد. در مواردی که پزشک با قاطعیت نمی‌تواند خوش‌خیم بودن یا بدخیم بون بیماری را گزارش نماید مدل‌های کامپیوتری وتکنیک‌های داده‌کاوی و یادگیری ماشین می‌توانند راهنمای خوبی برای پزشک باشند (۱۶، ۱۵). یادگیری ماشین شاخه ای از هوش مصنوعی (Artificial intelligence) است که با طرح و به‌کارگیری الگوریتم‌ها به کامپیوترها این امکان را می‌دهد که کارایی خود را براساس یادگیری، بهینه نمایند. امروزه تکنیک‌های داده‌کاوی و یادگیری ماشین که به عنوان تکنیک‌هایی برای شناسایی و تشخیص بیماری‌ها و دسته‌بندی بیماران در مدیریت بیماری به کار می‌روند، ضمن پیدا کردن الگوهایی برای تشخیص سریع‌تر بیماران و جلوگیری از بروز عوارض در آن‌ها، می‌توانند کمک بسیار بزرگی برای پزشکان باشند. در این‌جا، اهمیت سیستم‌های پشتیبان تصمیم‌گیری پزشکی (تصمیم یار پزشک) تبیین می‌شود. این سیستم‌ها به پزشکان و متخصصین این امر در تصمیم‌گیری دقیق‌تر یاری می‌رسانند و خطاهای احتمالی را کاهش می‌دهند. همچنین با استفاده از این سیستم‌ها، می‌توان اطلاعات پایگاه داده‌های پزشکی را در زمان بسیار کمتر و با جزییات بیشتری تحلیل نمود. کاهش هزینه‌ها وکاهش منابع انسانی به عنوان مزایای دیگر سیستم‌های پشتیبان تصمیم‌گیری پزشکی هستند (۲۴-۱۶).

امروزه روش‌های مختلف هوشمندانه‌ای در جهت تحقق یک مدل ریاضی برای طبقه بندی کردن الگوهای سرطانی انجام گرفته، ولی هیچ مدلی تاکنون نتوانسته است به طور دقیق تمام الگوهای سرطانی را طبقه بندی نماید (۲۵). Tan و همکاران، یک تکنیک طبقه‌بندی دو مرحله‌ای ترکیبی برای استخراج قوانین طبقه‌بندی ارائه دادند. روش پیشنهادی آن‌ها به دقت ۹۳/۰۴٪ بر روی پایگاه داده WDBC و ۹۷/۵۷٪ بر روی پایگاه WBCD دست یافت (۲۶). Kıyan و همکاران، به بررسی روش‌های RBF، GRNN و PNN بر روی مجموعه داده‌های WBCD پرداختند. نتایج آزمایشات آنها نشان داد که روش RBF دارای دقت ۹۶/۱۸٪، روش PNN دارای دقت ۹۸/۸٪ و روش MLP دارای دقت ۹۵/۷۴٪ است (۲۷). در (۲۸) کارایی طبقه‌بندی کننده درخت تصمیم CART با انتخاب ویژگی و بدون انتخاب ویژگی بر روی پایگاه‌های WBCD و WDBC بررسی شد. CART بدون استفاده از انتخاب ویژگی به دقت ۸۴ /۹۴٪ بر روی WBCD و دقت ۹۲/۹۷٪ بر روی WDBC و با استفاده از انتخاب ویژگی به دقت ۹۶/۹۹٪ بر روی WBCD و ۹۲/۰۹٪ بر روی WDBC دست یافت. با استفاده از خصوصیات استخراج شده از آزمایش آسپیراسیون سوزنی و با کمک تکنیک‌های یادگیری ماشین می‌توان سیستمی‌کارآمد را برای تشخیص سرطان پستان طراحی نمود که با دقت بالایی خوش‌خیم یا بدخیم بودن تومورهای پستان را تشخیص دهد. در این مقاله با استفاده از روش‌های برآورد ناپارامتری چگالی هسته‌ای و خصوصیات استخراج شده از آزمایش آسپیراسیون سوزنی به ارائه مدلی کارا برای تشخیص سرطان پستان می‌پردازیم.

روش کار

در این مطالعه، به ارائه مدلی برای تشخیص سرطان پستان با استفاده از مجموعه داده‌های WBCD و WDBC پرداخته شد. مدل ارائه شده مبتنی بر بهینه‌سازی پارامترها در روش‌های برآورد ناپارامتری چگالی بر اساس هسته است. برای این منظور با استفاده از تکنیک اعتبارسنجی متقاطع 4-fold بر روی مجموعه داده‌های آموزشی به تعیین مقدار بهینه پارامترهای مدل پرداخته شد.

جدول ۱- ویژگی‌های پایگاه داده WBCD

Attribute numbers	Attribute description	Values of attribute	Mean	Standard deviation
1	Clump thickness	1–10	4.44	2.83
2	size Uniformity of cell	1–10	3.15	3.07
3	shape Uniformity of cell	1–10	3.22	2.99
4	Marginal adhesion	1–10	2.83	2.86
5	cell size Single epithelial	1–10	2.23	2.22
6	Bare nuclei	1–10	3.54	3.64
7	Bland chromatin	1–10	3.45	2.45
8	Normal nucleoli	1–10	2.87	3.05
9	Mitoses	1–10	1.60	1.73

جدول ۲- ویژگی‌های پایگاه داده WDBC

Attribute numbers	description Attribute
1	Radius (mean of distances from center to points on the perimeter)
2	Texture (standard deviation of gray-scale values)
3	Perimeter
4	Area
5	Smoothness (local variation in radius lengths)
6	Compactness (perimeter^2 / area - 1.0)
7	Concavity (severity of concave portions of the contour
8	Concave points (number of concave portions of the contour)
9	Symmetry
10	Fractal dimension ("coastline approximation" - 1)

الف: توصیف مجموعه داده‌های WDBC و WBCD: این پژوهش توصیفی گذشته‌نگر است که مبتنی بر اطلاعات پرونده‌های بیمارستانی Wisconsin می‌باشد. مجموعه داده‌های پایگاه WBCD شامل اطلاعات ۶۹۹ بیمار با ۱۰ ویژگی است که ویژگی اول شماره شناسه پرونده بیمار و بقیه‌ی ویژگی ها نتایج کمی آزمایش آسپیراسیون سوزنی برای هر بیمار است. هر نمونه با یک برچسب خوش‌خیم یا بدخیم مشخص می‌گردد. ویژگی‌های پایگاه داده WBCD در جدول ۱ نشان داده شده است. مقادیر ویژگی‌ها، عددی صحیح بین ۱ تا ۱۰ است.

مجموعه داده‌های پایگاه WDBC شامل اطلاعات ۵۶۹ بیمار با ۳۱ ویژگی است که ویژگی اول شماره شناسه پرونده بیمار و ۳۰ ویژگی - باقیمانده از تصویر دیجیتالی آزمایش آسپیراسیون سوزنی توده پستان به‌دست آمده که این ویژگی‌ها خصوصیات هسته سلول در تصویر را بیان می‌کنند. ویژگی‌های پایگاه WDBC در جدول ۲ بیان شده‌اند. در واقع برای هر یک از این ویژگی‌ها میانگین، خطای استاندارد و بزرگترین مقدار (میانگین سه تا از بزرگ‌ترین مقادیر) این ویژگی‌ها محاسبه شده و به‌این ترتیب ۳۰ ویژگی با مقدار عددی حقیقی برای هر نمونه به‌دست می‌آید. هر یک از نمونه‌های پایگاه WDBC با یک برچسب خوش‌خیم یا بدخیم مشخص می‌گردند. از ۵۶۹ نمونه مذکور، ۳۵۷ نمونه دارای برچسب خوش‌خیم و ۲۱۲ نمونه دارای برچسب بدخیم هستند.

ب) مدل ارائه شده: در روش‌های برآورد چگالی هسته‌ای، انتخاب پارامتر هموارسازی مناسب از اهمیت خاصی در برآورد چگالی برخوردار است. مدل ارائه شده در این مقاله برای طبقه‌بندی داده‌های پایگاه‌های WBCD و WDBC مبتنی بر بهینه‌سازی پارامترها در روش‌های تخمین چگالی بر اساس هسته است. در طبقه‌بندی داده‌های پایگاه‌های WBCD و WDBC، ابتدا به صورت تصادفی داده‌ها در دوگروه آموزش (Train) و آزمون (Test) قرار می‌گیرند. برای بهینه‌سازی پارامترهای مدل، با استفاده از تکنیک اعتبارسنجی

متقـاطع 4-fold بـر روی مجموعـه داده‌هـای آموزشی به تعیین مقدار بهینه پـارامترهـای مـدل می‌پردازیم. پـس از آن بـا اسـتفاده از پارامترهـای بهینه تعیین شده ، فرآیند یـادگیری بـر روی داده های گروه آموزش انجام می شود و مدل مورد نظر با استفاده از داده های گروه آزمون سنجیده خواهد شـد. مـدل ارائـه شـده دارای سـه مرحلـه پـیش پردازش داده‌ها، اعتبارسنجی و طبقه‌بندی است که در ادامـه بـه توضـیح هریـک از مراحـل پرداختـه می‌شود.

مرحله اول: پیش پردازش داده‌ها

اولـین مرحلـه در ایجـاد هـر مـدلی بـر اسـاس تکنیـک‌هـای داده‌کـاوی، مرحلـه پـیش‌پـردازش (Preprocessing)می‌باشد که جهت بهبود کیفیت داده‌های واقعـی بـرای داده‌کـاوی لازم اسـت. ایـن مرحلـه شـامل برخـورد بـا داده‌هـای گمشـده و نرمال‌سازی داده‌ها است.

در پایگـاه داده WBCD، ۱۶ نمونـه بـا مقـادیر گمشده (Missing) وجود دارد. در مرحلـه پـیش پردازش داده‌های این پایگاه، ابتـدا شـماره شناسـه بیمار و نمونه‌های دارای مقـادیر گمشـده را حـذف نمـوده و آزمایشـات را بـا ۶۸۳ نمونـه و ۹ ویژگـی ادامه می‌دهیم. از ۶۸۳ نمونه مـذکور، ۴۴۴ نمونـه دارای برچسـب خـوش‌خـیم (۶۵٪) و ۲۳۹ نمونـه دارای برچسب بدخیم (۳۵٪) هستند.

در مرحله پـیش پـردازش داده‌هـای پایگـاه داده WDBC، ابتدا شماره شناسه بیمار را حذف نموده و سـپس داده‌هـا را نرمـال‌سـازی مـی‌نمـاییم. نرمال‌سازی تغییر مقیاس داده‌ها بـه‌گونـه‌ای اسـت که آنها را به یک دامنـه کوچـک و معـین نگاشـت کند. نرمال سازی داده‌ها معمولاً منجـر بـه کسـب نتایج بهتر می‌شود. به علت متفاوت بـودن مقیـاس داده‌هـای پایگـاه WDBC، بـا اسـتفاده از روش نرمـال‌سـازی مـاکزیمم- مـی‌نـیمم (Min-max) داده‌ها را نرمال‌سازی می‌نمائیم.

مرحله دوم: اعتبارسنجی

در این مرحله، با استفاده از تکنیک اعتبارسنجی متقـاطع 4-fold بـر روی مجموعـه داده‌هـای آموزشی به تعیین مقدار بهینه پـارامترهـای مـدل می‌پردازیم و سپس آزمایشات را بـا مقـادیر بهینـه بدست آمده بر روی مجموعه داده‌های آزمون انجام مـی‌دهـیم. روش اعتبارسـنجی متقـاطع 4-Fold نمونه‌های آموزشی را به صورت تصـادفی بـه چهـار گروه نسبتاً مساوی تقسـیم مـی‌کنـد و چهـار بـار عملیات آموزش را برای سه گروه و عملیات آزمـون را برای گروه چهارم تکرار می‌کند. دقت کـل مـدل به وسیله میانگین چهار دقت مجزا در پیش‌بینی‌ها محاسبه می‌شـود. در مـورد تعیـین مقـادیر بهینـه پارامترهای مدل با استفاده از تکنیک اعتبارسنجی متقاطع 4-fold بر روی داده‌های آموزشی، مقـادیر مختلفـی را بـه ازای پارامترهـای مـدل آزمـایش می‌نماییم و مقداری که بـالاترین دقـت را دارد بـه عنوان مقدار بهینه در نظر می‌گیریم.

مرحله سوم: طبقه‌بندی

در این مرحله با استفاده از روش‌های طبقه‌بنـدی ناپارامتری مبتنی بر بـرآورد چگـالی هسـته‌ای بـه طبقه‌بندی مجموعه داده‌های آزمون با اسـتفاده از مقادیر بهینه پارامترهای مدل می‌پردازیم. در ادامه روش‌های ناپارامتری استفاده شـده در ایـن مقالـه شرح داده می‌شوند.

برآورد چگالی ناپـارامتری: تـابع توزیـع چگـالی، مفهومی بنیادی در آمار است. متغیر تصادفی X را در نظر بگیرید که تابع توزیع چگـالی آن P اسـت. اگر تابعP را داشته باشیم مـی‌تـوانیم تخمینـی از توزیع X داشته باشیم. فرض کنیـد مجموعـه‌ای از داده‌های مشاهده شده از نمونه‌ها وجود دارنـد کـه تـابع توزیـع چگـالی آن ناشـناخته اسـت. بـرآورد چگـالی (Density estimation) بـه فرآینـد تخمین تابع چگالی احتمال یک متغیر تصـادفی بـا استفاده از نمونـه‌هـای مشـاهده‌شـده از آن متغیـر گفته می‌شود. معمـولاً بـرای بـرآورد تـابع چگـالی احتمال از روش‌های پارامتری و ناپارمتری استفاده می‌شود. در روش‌های پارامتری فرض می‌شـود کـه داده‌ها از یک توزیع مشخص پیروی می‌کنند و تنها باید مقادیر پارامترهـای آن تخمـین‌زده شـوند. در برآورد ناپارامتری، فرضیات کمتری درباره‌ی توزیع داده‌های مشاهده شده صورت می‌گیرد. روش‌هـای طبقه‌بندی پارامتری و ناپارامتری بر اسـاس قـانون بیز عمل می‌کنند.

روش‌های مختلفی برای برآورد چگالی ناپارامتری

وجود دارد که در ادامه به توضیح آن‌ها پرداخته می‌شود:

فرض کنید:

N داده‌ی آموزشی $X = \{x^t\}_{t=1}^N$ وجود دارد که هر یک دارای d ویژگی است. همچنین فرض می‌شود که داده‌ها به صورت مستقل توزیع شده و دارای چگالی احتمال ناشناخته $P(.)$ هستند. $\hat{P}(.)$ برآوردگر $P(.)$ است. برآورد ناپارامتری به صورت رابطه زیر تعریف می‌شود:

$$(۱) \quad \hat{P}(x) = \frac{1}{h}\left[\frac{\#\{x^t \le x+h\} - \#\{x^t \le x\}}{N}\right]$$

در این رابطه، h به عنوان طول بازه تعریف می‌شود و x^t نمونه‌هایی است که در این بازه قرار می‌گیرند.

برآوردگر هیستوگرام: هیستوگرام قدیمی‌ترین برآوردگر ناپارامتری چگالی است که فضای ورودی را به بازه‌هایی با طول مساوی به نام *bin* تقسیم می‌کند.

برای داشتن هیستوگرام باید یک نقطه ($x0$) و طول بازه (h) را انتخاب کرد. این دو انتخاب بر روی برآورد تاثیر می‌گذارند. انتخاب طول بازه، تعیین‌کننده‌ی مقدار هموارسازی است و با داشتن hهای مختلف، هموارسازی‌های مختلفی وجود خواهد داشت. هیستوگرام ارائه شده به نقطه ی $x0$ انتخابی نیز بستگی دارد و انتخاب‌های مختلف $x0$ می تواند نتایج مختلفی داشته باشد. از معایب هیستوگرام می‌توان به گسسته بودن و وابستگی شکل ارائه شده برای توزیع به $x0$ اشاره کرد.

برآوردگر ساده: یکی دیگر از روش‌های برآورد چگالی، برآوردگر ساده (Naïve estimator) است. با استفاده از این برآوردگر نیاز به انتخاب نقطه‌ی $x0$ نمی‌باشد. در واقع این برآوردگر یکی از روش‌های موجود برای برطرف کردن مشکل وابستگی هیستوگرام به نقطه‌ی $x0$ است.

تخمینی که با روش برآوردگر ساده ارائه می‌شود، ناهموار است و برای رفع این مشکل از برآوردگرهای هسته‌ای (Kernel Estimators) استفاده می‌شود (۲۹).

برآوردگر هسته‌ای گوسین: برآوردگر هسته‌ای تعمیمی از برآوردگر ساده می‌باشد. و به صورت رابطه زیر تعریف می‌شود(۲۹):

$$(۲) \quad \hat{P}(x) = \frac{1}{Nh}\sum_{t=1}^{N} K\left(\frac{x-x^t}{h}\right)$$

در این رابطه $K(.)$ تابع هسته، N تعداد نمونه‌های آموزشی و h اندازه‌ی پنجره است که به آن پارامتر هموارساز یا پهنای باند نیز می‌گویند. انتخاب پارامتر هموارسازی مناسب مهمترین مسئله در برآورد هسته‌ای است.

یکی از معروفترین توابع هسته، هسته گوسین به صورت زیر است:

$$(۳) K(u) = \left(\frac{1}{\sqrt{2\pi}}\right)^d exp\left[-\frac{\|u\|^2}{2}\right]$$

که منظور از $\|u\|$ فاصله اقلیدسی است. می‌توان در تخمین چگالی هسته‌ای به جای فاصله اقلیدسی از فاصله‌ی ماهالانوبیس *(Mahalonobis)* استفاده کرد تا همبستگی را نیز در نظر بگیرد(۲۹). برآوردگرد هسته‌ای مبتنی بر فاصله ماهالانوبیس به صورت رابطه‌ی زیر تعریف می‌شود:

$$(۴) \quad K(u) = \frac{1}{(2\pi)^{d/2}|S|^{1/2}} exp\left[-\frac{1}{2}u^T S^{-1} u\right]$$

در این رابطه، S ماتریس کوواریانس نمونه‌ها است.

هنگامی‌که نمونه‌های ورودی دارای مقادیر گسسته باشند، می‌توان به جای فاصله اقلیدسی و ماهالانوبیس، از فاصله همینگ مطابق رابطه زیراستفاده کرد که تعداد ویژگی‌های غیر منطبق را می‌شمارد.

$$(۵) \quad \sum_{j=1}^{d} 1\left(x_j \neq x_j^t\right) = Hamina\ distance\ (x, x^t)$$

برآوردگر هسته‌ای k نزدیکترین همسایه: در برآوردگرهای نزدیک ترین همسایه، هموارسازی با

توجه به چگالی محلی داده‌ها انجام می‌گیرد. این روش، ناحیه‌ی اطراف x را تا یافتن k امین همسایه ی نزدیک به x گسترش می‌دهد و شدت هموارسازی با توجه به مقدار k تعیین می‌شود(۲۹).

برآوردگر هسته‌ای k نزدیکترین، مطابق فرمول زیر تعریف می‌شود:

$$(۶)\ \hat{p}(x) = \frac{1}{Nd_k(x)} \sum_{t=1}^{N} K\left(\frac{x-x^t}{d_k(x)}\right)$$

در این رابطه، معمولاً $K(\cdot)$ یک تابع هسته گوسین در نظر گرفته می‌شود و $d_k(x)$ فاصله نمونه x تا k امین نزدیکترین همسایه است.

ج) شاخص های ارزیابی

برای بررسی عملکرد سیستم‌های تشخیص کامپیوتری بصورت عمده از شاخص‌های گوناگونی استفاده می‌شود. در این مقاله، برای ارزیابی مدل‌ها از شاخص‌های دقت (Accuracy)، حساسیت (Sensitivity) و شفافیت (Specificity) استفاده می‌شود. میزان دقت یک روش دسته‌بندی بر روی مجموعه داده‌های آزمون، درصد مشاهداتی از مجموعه آزمون است که به درستی توسط مدل مورد استفاده دسته‌بندی شده است. حساسیت عبارت است از میزانی برای مشخص‌کردن توانایی سیستم در تشخیص و دسته‌بندی موارد واقعاً بیمار (سرطانی) که سیستم آنها را صحیح و سرطانی تشخیص می‌دهد. شفافیت عبارت است از میزانی برای مشخص‌کردن توانایی سیستم در تشخیص و دسته‌بندی موارد واقعاً سالم که سیستم آنها را به صورت صحیح و سالم تشخیص می‌دهد. شاخص‌های دقت، حساسیت و شفافیت توسط فرمول‌های زیر بیان می‌شوند.

$$(۷)\ \text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$$

$$(۸)\ \text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}}$$

$$(۹)\ \text{Specificity} = \frac{\text{TN}}{\text{FP+TN}}$$

TP (True Positive): برابر است با تعداد بیماران مبتلا به سرطان پستان که سیستم تشخیص کامپیوتری آنها را بصورت صحیح سرطانی تشخیص داده است.

TN (True Negative): برابراست با تعداد بیماران مبتلا به تومور خوش‌خیم یا سالم که سیستم تشخیص کامپیوتری آنها را بصورت صحیح سالم تشخیص داده است.

FP (False Positive): برابر است با تعداد بیمارانی که سیستم تشخیص کامپیوتری آنها را بصورت اشتباه سرطانی تشخیص داده است.

FN (False Negative): برابر است با تعداد بیمارانی که سیستم تشخیص کامپیوتری آنها را بصورت اشتباه سالم تشخیص داده است.

یافته‌ها

در طبقه‌بندی داده‌های پایگاه‌های WBDC و WDBC، ابتدا در مرحله اعتبارسنجی مقادیر بهینه پارامتر h در روش برآورد چگالی هسته‌ای گوسین و پارامترk در روش k نزدیکترین همسایه را با استفاده از روش اعتبارسنجی متقاطع 4-fold بر روی داده‌های مجموعه آموزش به‌دست آورده و سپس آزمایشات را با مقادیر بهینه بدست آمده بر روی مجموعه داده‌های آزمون انجام می‌دهیم. برای تعیین مقدار بهینه پارامتر h از بازه‌ی {۰/۰۱، ۰/۰۵، ۰/۰۷، ۰/۰۸، ۰/۰۹، ۰/۱، ۰/۱۱، ۰/۱۲، ۰/۱۳، ۰/۱۴، ۰/۱۵، ۰/۲، ۰/۳، ۰/۴، ۰/۵، ۰/۶، ۰/۷، ۰/۸، ۰/۹، ۱، ۱/۱، ۱/۲، ۱/۳، ۱/۴، ۱/۵، ۱/۶، ۱/۷، ۱/۸، ۱/۹، ۲، ۳، ۴} و برای تعیین مقدار بهینه‌ی پارامتر k از بازه‌ی {۱، ۳، ۵، ۷، ۹، ۱۱، ۱۳، ۱۵، ۱۷، ۱۹، ۲۱، ۳۱، ۴۱، ۵۱، ۵۵، ۶۱، ۶۵، ۷۱، ۷۳، ۷۵، ۷۷، ۷۹، ۸۱، ۸۳، ۸۵، ۸۷، ۸۹، ۹۱، ۹۳، ۹۵، ۹۷، ۹۹}استفاده شده است.

الف: نتایج آزمایشات بر روی مجموعه داده WBDC: در روش برآورد چگالی هسته‌ای گوسین بر روی داده‌های WBDC، از فاصله‌های مختلفی نظیر فاصله اقلیدسی و فاصله همینگ استفاده

جدول ۳- نتایج بررسی روش‌های غیرپارامتری مبتنی بر برآورد هسته‌ای بر روی پایگاه WBCD

نام روش	دقت (Accuracy)	حساسیت (Sensitivity)	شفافیت (Specificity)
روش هسته‌ای گوسین (فاصله اقلیدسی) (h=1.4)	۹۸/۱۷	۹۴/۷۴	۱۰۰
روش هسته‌ای گوسین (فاصله همینگ) (h=0.05)	۹۷/۴۴	۹۵/۷۹	۹۸/۳۱
k نزدیکترین همسایه (k=83)	۹۸/۱۷	۹۴/۷۴	۱۰۰
k نزدیکترین همسایه (k= 85)	۹۸/۱۷	۹۴/۷۴	۱۰۰

جدول ۴- نتایج بررسی روش‌های غیر پارامتری مبتنی بر برآورد هسته‌ای بر روی داده‌های پایگاه WDBC

نام روش	دقت (Accuracy)	حساسیت (Sensitivity)	شفافیت (Specificity)
روش هسته‌ای گوسین (فاصله اقلیدسی) (h=0.2)	۹۷/۹۳	۹۳/۱۸	۹۹/۳۳
روش هسته‌ای گوسین (فاصله ماهالانوبیس) (h=0.2)	۵۲/۸۵	۱۰۰	۳۸/۹۳
k نزدیکترین همسایه (k=11)	۹۴/۳۰	۹۴/۶۳	۹۳/۱۸
k نزدیکترین همسایه (k=13)	۹۴/۸۲	۹۵/۳۰	۹۳/۱۸
k نزدیکترین همسایه (k=15)	۹۴/۸۲	۹۴/۶۳	۹۵/۴۵

می‌نماییم. دلیل استفاده از فاصله همینگ بر روی داده‌های WBCD گسسته بودن (discrete) مقادیر داده‌های این پایگاه است. در مورد تعیین مقادیر بهینه پارامترهای h و k، ابتدا با استفاده از تکنیکی مجموعه داده‌های آموزشی را به زیرمجموعه‌هایی جداگانه برای ایجاد و تست مدل‌ها با مقادیر مختلف این پارامترها تفکیک کرده و مقادیر مختلف h را آزمایش نموده و مقداری که بالاترین دقت را دارد به عنوان مقدار بهینه h در نظر می‌گیریم.

بررسی مقادیر مختلف پارامتر h با استفاده از روش اعتبارسنجی متقاطع 4-fold بر روی داده‌های آموزشی نشان داد که در روش هسته‌ای گوسین با استفاده از فاصله اقلیدسی مقدار بهینه این پارامتر ۱/۴ و دارای دقت ۳۶ /۹۵٪ است و با استفاده از فاصله همینگ مقدار بهینه این پارامتر۰/۰۵ و دارای دقت ۱۰/ ۹۶٪ است. بنابراین در ادامه برای بررسی ارزیابی روش هسته‌ای گوسین بر روی دادهای آزمون از این مقادیر استفاده می‌شود. بررسی مقادیر مختلف k با استفاده از روش اعتبارسنجی متقاطع 4-fold بر روی داده‌های آموزشی نشان داد که در روش k نزدیکترین همسایه مقدار بهینه پارامتر k مربوط به مقادیر ۸۳ و ۸۵ با دقت ۵۸ /۹۶٪ است. در جدول ۳ نتایج بررسی روش‌های غیرپارامتری مبتنی بر برآورد هسته‌ای بر روی مجموعه داده‌های آزمون پایگاه WBCD نشان داده شده است.

همانگونه در جدول ۳ مشاهده می‌شود، روش‌های تخمین چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی و k نزدیکترین همسایه با دقت ۹۸/۱۷٪ دارای بالاترین دقت هستند.

ب: نتایج آزمایشات بر روی مجموعه داده WDBC: در روش برآورد چگالی هسته‌ای گوسین بر روی داده‌های WDBC، از فاصله‌های مختلفی نظیر فاصله اقلیدسی و فاصله مهالانوبیس (Mahalanobis) استفاده می‌نماییم. بررسی مقادیر مختلف پارامتر h با استفاده از روش اعتبارسنجی متقاطع 4-fold بر روی داده‌های آموزشی نشان داد که در روش هسته‌ای گوسین با استفاده از فاصله اقلیدسی، مقدار بهینه پارامتر$h=0.2$ دارای دقت ۲۷ /۹۶٪ است و با استفاده از فاصله مهالانوبیس مقدار بهینه پارامتر $h=0.2$ دارای دقت ۳۴/ ۷۲٪ است. در بررسی مقادیر مختلف k در روش k نزدیکترین همسایه، با استفاده از روش اعتبارسنجی متقاطع 4-fold بر روی داده‌های آموزشی WDBC مشخص گردید که مقدار بهینه پارامتر k مربوط به مقادیر ۱۱،۱۳و ۱۵ با دقت ۹۳/۸۸٪ است. در جدول ۴ نتایج بررسی روش‌های غیرپارامتری مبتنی بر برآورد هسته‌ای بر

روی مجموعه داده‌های آزمون پایگاه WDBC نشان داده شده است.

همانگونه در جدول ۴ مشاهده می‌شود، در میان انواع روش‌های غیرپارامتری بر روی پایگاه WDBC، روش برآورد چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی بالاترین دقت را در مقایسه با دیگر روش‌ها دارا است.

بحث و نتیجه‌گیری

در این مطالعه مدلی برای تشخیص سرطان پستان با استفاده از مجموعه داه‌های WBCD و WDBC ارائه شد که مبتنی بر بهینه‌سازی پارامترهای روش‌های تخمین چگالی هسته‌ای با استفاده از تکنیک اعتبارسنجی متقاطع 4-fold بر روی مجموعه داده‌های آموزشی است. نتایج آزمایشات نشان داد که روش برآورد چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی با دقت ۹۷/۹۳٪ بالاترین دقت را بر روی مجموعه داده‌های WDBC برای تشخیص سرطان پستان دارد و روش‌های تخمین چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی و k نزدیکترین همسایه با دقت ۹۸/۱۷٪ بالاترین دقت را بر روی مجموعه داده‌های WBCD برای تشخیص سرطان پستان دارند.

استفاده از روش‌های هوشمند داده‌کاوی و یادگیری ماشین می‌تواند کمک بزرگی برای پزشکان در تشخیص بیماری‌ها باشد (۱۷). با استفاده از این روش‌ها، پزشک می‌تواند متغیرهای بیشتر و متنوع تری را در زمان تشخیص بیماری یا انتخاب درمان در نظر بگیرد(۲۰). این روش‌ها می توانند قابلیت تصمیم گیری پزشکان را ارتقا بخشند و خطا های احتمالی ناشی از خستگی یا بی‌تجربگی متخصصین بالینی را کاهش دهند (۲۱، ۲۲) . تحقیقات متعددی توسط محققان برای پیش بینی انواع سرطان با تکنیک‌های داده‌کاوی و یادگیری ماشین انجام شده است. Tan و همکاران یک تکنیک طبقه‌بندی دو مرحله‌ای ترکیبی برای استخراج قوانین طبقه‌بندی ارائه دادند. روش پیشنهادی آن‌ها به دقت ۹۳/۴٪ بر روی پایگاه داده WDBC و ۹۷/۵۷٪ بر روری پایگاه WBCD دست یافت (۱۹). Kiyan و همکاران، به بررسی تکنیک RBF بر روی پایگاه WBCD پرداختند. نتایج آزمایشات آن‌ها نشان داد که روش RBF دارای دقت ۹۶/۱۸٪ در تشخیص سرطان پستان است (۲۰). Raad و همکاران در سال ۲۰۱۲، برای تشخیص سرطان پستان با تکنیک شبکه عصبی RBF به بررسی پایگاه داده WBCD پرداختند و با دقت ۹۷ درصد سرطان پستان را تشخیص دادند(۱۴). Chaurasia و همکاران با استفاده از روش SVM و انتخاب ویژگی‌های مناسب بر روی پایگاه WBCD به دقت ۹۶/۴٪ در تشخیص سرطان پستان دست یافتند (۳۰).

Aruna و همکاران به بررسی و مقایسه طبقه کننده‌های یادگیری با نظارت نظیر روش بیزین ساده، SVM-RBF، شبکه عصبیRBF، درخت تصمیم J48 و CART برروی مجموعه داده‌های پایگاه WBCD پرداختند تا بهترین طبقه‌بندی کننده را بر روی این داده‌ها پیدا نمایند. نتایج آزمایشات آنها نشان داد که SVM-RBF با دقت ۹۶/۸۴٪ نسبت به طبقه‌بندی‌کننده‌های دیگر دقت بیشتری دارد (۳۱). فلاحتی و جعفری در سال ۲۰۱۱ سیستم خبره‌ای برای تشخیص سرطان پستان با استفاده از داده‌های پیش پردازش و شبکه بیزین طراحی نمودند که توانست با دقت ۹۸/۱٪ داده‌های پایگاه WBCD را به درستی تشخیص دهد (۳۲). در مطالعه Lavanyaو همکاران در سال ۲۰۱۱، کارایی طبقه بندی‌کننده درخت تصمیم CART بدون انتخاب ویژگی بر روی پایگاه‌های WBCD بررسی شد. این محققان با تکنیک طبقه‌بندی دو مرحله‌ای ترکیبی به دقت ۹۴/۸۴ ٪ رسیدند (۳۳).

Setiono و همکاران با روش درخت تصمیم C4.5 به بررسی مجموعه داده‌های پایگاه داده WDBC پرداختند و به دقت ۹۲/۶۱٪ در تشخیص سرطان پستان دست یافتند (۳۴). Bamakan و همکاران در سال ۲۰۱۴ روشی برای انتخاب ویژگی براساس تحلیل پوششی داده های مجتمع و مدل آنتروپی ارائه دادند که با استفاده از روش‌های دسته‌بندی SVM ، درخت تصمیم C5.0 و رگرسیون لجستیک بر روی پایگاه

WDBC به ترتیب بـه دقـت ۸۹/۸۶٪، ۹۳/۹۲٪ و ۹۵/۹۵٪ دست یافتند. آنها همچنین با اسـتفاده از روش رگرسیون لجستیک و تکنیک انتخاب ویژگی CfsSubsetEval به دقت ۹۵/۹۵٪ و با اسـتفاده از روش رگرسیون لجستیک و تکنیک انتخاب ویژگی فیلتر به دقت ۹۶/۶۲٪ دسـت یافتنـد. در مطالعـه آنها روش SVM با تکنیـک هـای انتخـاب ویژگـی فیلتر و CfsSubsetEval با دقت ۸۷/۸۴٪ سـرطان پستان را به درستی پیش‌بینی نمود (۳۵).

Salama و همکاران در سال ۲۰۱۲ با استفاده از روش بیزین ساده و درخت تصمیم J48 به ترتیـب به دقـت ۹۲/۹۷٪ و ۹۳/۱۵٪ بـر روی پایگـاه داده WDBC دست یافتند (۳۶). Yao و همکـاران در سال ۲۰۱۳ با اسـتفاده از روش‌هـای RF، MARS و RF&MARS بـر روی داده‌هـای WDBC بـه ترتیب به دقت ۹۶/۲۶٪، ۹۶/۷٪ و ۹۶/۲۹٪ دسـت یافتند. آنها همچنین نشان دادندکـه روش درخـت تصمیم C4.5 دارای دقـت ۹۳/۱۶٪ و روش SVM دارای دقـت ۹۵/۸۵٪ اسـت (۳۷). Maldonado و همکـاران در سـال ۲۰۱۱ بـا اسـتفاده از تکنیـک انتخاب ویژگی Fisher و طبقه‌بندی کننـده SVM به دقت ۹۴/۷٪ و با تکنیک انتخاب ویژگی RFE و طبقه‌بندی‌کننده SVM بـه دقـت ۹۵/۲۵٪ دسـت یافتند (۳۸).

مدل ارائه شده در این مطالعه، با استفاده از روش برآورد چگالی هسته‌ای گوسین مبتنـی بـر فاصـله اقلیدسی بر روی مجموعـه داده‌هـای WDBC بـه دقـت ۹۷/۹۳٪ دسـت یافـت و بـا اسـتفاده از روش‌های تخمین چگالی هسته‌ای گوسین مبتنـی بر فاصله اقلیدسی و k نزدیکترین همسایه بر روی مجموعـه داده‌هـای WBCD بـه دقـت ۹۸/۱۷٪ دست یافت. بدین ترتیب با استفاده از مـدل ارائـه-شـده، دقـت شناسـایی سیسـتم هـای تشـخیص سرطانن پستان افزایش یافت. نتایج مطالعات ذکـر شده محققان بر روی داده‌های پایگاه‌هـای WBCD و WDBC در مقایسه با مطالعه حاضـر، حـاکی از برتری مدل ارائه شده در این مطالعه اسـت. نتـایج این مطالعـه نشـان داد کـه روش‌هـای ناپـارامتری برآورد چگالی هسته‌ای می‌تواننـد بـا دقـت بـالایی برای تشخیص سرطان پستان به کار روند. بنابراین اسـتفاده از تکنیـک‌هـای یـادگیری ماشـین و داده‌کاوی می‌توانند به عنوان یک سیسـتمی بـرای کمک به پزشکان باشند تا خطای احتمالی ناشی از تشخیص پزشک تا حد امکان کاهش یابد.

منابع

1. Sedehi M, Amani F, Momeni Dehaghi F. Analysis of survival data of patient with breast cancer using artificial neural network and cox regression models. J Zabol Univ Medi Sci; 2013. 5(4): 1-8.
2. Wang YA, Johnson SK, Brown BL, Carragher LM, Sakkaf KL, Royds JA, et al. Enhanced anticancer effect of a phosphatidylinositol-3 kinase inhibitor and doxorubicin on human breast epithelial cell lines with different p53 and oestrogen receptor status. IJC; 2008. 123(7): 1536-44.
3. Wang YA, Johnson SK, Brown BL, Carragher LM, Sakkaf KL, Royds JA et al. Enhanced anticancer effect of a phosphatidylinositol-3 kinase inhibitor and doxorubicin on human breast epithelial cell lines with different p53 and oestrogen receptor status.IJC ; 2008. 123(7):1536–44.
4. RichieJohn RC, Swanson O. Breast Cancer: A Review of the Literature.J Insur Med; 2003. 35: 85 -101.
5. Abbasi Layegh M, Ghobadi Ch, J Nourinia J, Muhammadi B. 3-D breast cancer detection using support vector machines and finite element methods. J Urmia Univ Med Sci; 2014. 25(6): 539-548.
6. McSherry EA, Brennan K, Hudson L, DK Hill A, Hopkins AM. Breast cancer cell migration is regulated through junctional adhesion molecule-A-mediated activation of Rap1 GTPase. Breast Canc Res; 2011. 13:R31.
7. Sheikhpour R, Ghasemi N, Yaghmaei P, Mohiti J. Immunohistochemical assessment of p53 protein and its correlation with clinicopathological parameters in breast cancer patients. Indian J Sci Technol; 2014. 7(4): 472-479.
8. IAf RoC L. World Cancer Report.IARC; 2003:188-193.
9. Jain R, Abraham A. A Comparative Study of Fuzzy Classification Methods on Breast Cancer Data. 7th International Work Conference on Artificial and Natural Neural Networks, IWANN'03; 2003:1-6.
10. Sheikhpour R, Mohiti Ardekani J. The effect of progesterone on p53 protein in T47D cell line. J Urmia Univ Med Sci; 2014. 25(7): 18-28.
11. Fentiman IS. Fixed and modifiable risk factors for breast cancer. IJCP; 2001. 55(8):527-

based on thermal images using a combination of network SOM and MLP. Iranian J breast Canc; 2012. 5(2,3): 71-86.

27. Tan KC, Yu Q, Heng CM, Lee TH. Evolutionary computing for knowledge discovery in medical diagnosis. Artif Intell Med; 2003. 27: 129–154.

28. Kıyan T, Yıldırım T. Breast cancer diagnosis using statistical neural network, International XII. TAINN; 2003. 1-6.

29. Alpaydin E. Introduction to machine learning. (2th ed.). London: MIT press; 2010.

30. Chaurasia S, Chakrabarti P. An approach with Support Vector Machine using Variable Features Selection on Breast Cancer Prognosis IJARAI; 2013. 2(9): 38-42.

31. Aruna S, Rajagopalan DS, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. Comput sci Inform Technol; 2011. 2: 37-45.

32. Fallahi A, Jafari S. An expert system for detection of breast cancer using data preprocessing and Bayesian network. IJAST; 2011. 34: 65-70.

33. Lavanya D, Rani KU. Ensemble decision tree classifier for breast cancer data. IJITCS; 2012. 2(1): 17-24.

34. Setiono R. Generating concise and accurate classification rules for breast cancer diagnosis. Artif Intell Med; 2000.18(3):205-19.

35. Bamakan SMH, Gholami PA. Novel Feature Selection Method based on an Integrated Data Envelopment Analysis and Entropy Model. Procedia Comput Sci; 2014. 31: 632-8.

36. Salama GI, Abdelhalim MB, Zeid MAE. Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers. Int J Comput Sci Inform Tech; 2012. 1(1): 2277- 0764.

37. Yao D, Yang J, Zhan X. A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines. J Compu; 2013. 8(1): 170-7.

38. Maldonado S, Weber R, Basak J. Simultaneous feature selection and classification using kernel-penalized support vector machines. Inform Sci; 2011. 181(1): 115-28.

30.

12. Berdi Ghourchaei A, Charkazi A, RazzaqNejad A. Knowledge, Practice and Perceived Threat toward Breast Cancer in the Women living in Gorgan, Iran. JRNM; 2013. 10(1): 32.

13. Litigate J. Predictive models for breast cancer susceptibility from multiple single nucleotide polymorphisms. J clin Canc Res; 2004. 10: 2725-37.

14. Madani S, Izad B, Kanani M, Khazaei S, Hamzeloie K, Molaei Tavana P. Comparison of FNA and surgical biopsy result of palpable breast masses. J Urmia Univ Med Sci; 2012. 23(4): 422-426 [Persian].

15. Raad A, Kalakech A, Ayache M. Breast cancer classification using neural network approach: MLP and RBF. The ACIT; 2012:10-13.

16. Sheikhpour R, Sarram MA, Zare Mirakabad M, Sheikhpour R. Breast Cancer Detection Using Two-Step Reduction of Features Extracted From Fine Needle Aspirate and Data Mining Algorithms. Iranian J Breast diseas; 2015. 7(4):43-51 [Persian].

17. Nahar J, Imam T, Tickle KS, Shawkat Ali ABM, Chen PComputational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer Expert Syst Appl; 2012. 39:12371–12377.

18. Ameri H, Alizade S, Barzegari A. Knowledge Extraction of Diabetics Data by Decision Tree Method. JHAE; 2013. 6(53); 58-72.

19. Lavanya D, Usha Rani K. Analysis of feature selection with classification: Breast cancer datasets. IJCSE; 2011. 2(5):726- 734.

20. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med; 2005. 34(2):113–127.

21. Sadoughi F, Sheikhtaheri A. Applications of Artificial Intelligence in Clinical Decision Making: Opportunities and Challenges. HIMJ; 2011. 8(3): 440-446. [Persian]

22. Milovic B. Prediction and decision making in Health Care using Data Mining. IJPHS; 2012. 1(2): 69-78.

23. Marcano-Cedeño A, Quintanilla-Domínguez J, Andina D. WBCD breast cancer database classification applying artificial metaplasticity neural network. Expert Syst Appl; 2011. 38: 9573-9579.

24. Hariz M, Adnan M, Husain W, Rashid NA. Data Mining for Medical Systems: A Review. ACSTY; 2012:17-22.

25. Alavi Majd H, Vahedi M, Mehrabi Y, Naghavi B. Clustering approach in DNA microarray. Pagoohesh in pezeshki; 2007.31(1): 19-25.

26. Ghayomi zadeh H, Droud gar moghadam A, Hadad nia J. Clustering and breast cancer screening

Breast cancer diagnosis using non-parametric kernel density estimation

****Robab Sheikhpour**, Department of Physical Education & Sport Science, Taft Branch, Islamic Azad University, Taft, Iran And Hematology and Oncology Research Center, Shahid Sadoughi University of Medical Sciences, Yazd, Iran (*Corresponding author). r.sheikhpour@yahoo.com
Razieh Sheikhpour, Department of Computer Engineering, Yazd University, Yazd, Iran. r_sheikhpour@stu.yazd.ac.ir

Abstract

Background: Breast cancer is the most common cancer in women. An accurate and reliable system for early diagnosis of benign or malignant tumors seems necessary. We can design new methods using the results of FNA and data mining and machine learning techniques for early diagnosis of breast cancer which able detection of breast cancer with high accuracy. The aim of this study was to diagnosis of breast cancer using non-parametric kernel density estimation.

Methods: In this study, 699 samples of benign and malignancy with 9 characteristics from WBCD and 569 samples of benign and malignancy with 30 characteristics from WDBC were used. Then, a model based on non-parametric kernel density estimation was proposed for classification of WBCD and WDBC data.

Results: The results of non-parametric methods showed that Gaussian kernel method based on Euclidean distance with accuracy ٪97.93 has the highest accuracy on WDBC data and Gaussian kernel based on Euclidean distance and k-nearest neighbor methods with accuracy ٪98.17 has the highest accuracy compared with other methods on WBCD data for breast cancer disease.

Conclusion: The result of this study showed that non-parametric kernel density estimation based classification can be used for breast cancer diagnosis with high accuracy.

Keywords: Breast cancer, Non-parametric method, Kernel based density estimation, Machine learning brush