

## 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 پذیرش: ۹۴/۱۲/۱۷

تاریخ دریافت: ۹۴/۸/۲۰

### 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 پستان 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 WDCB با استفاده از روش‌های تخمین چگالی مبتنی بر هسته پرداخته شد.

**یافته‌ها:** نتایج بررسی روش‌های غیر پارامتری بر پایگاه داده WDCB 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

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 حالی که این میزان در خانم‌هایی که 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

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). تقریباً 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 (۵-۸). 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،

مدلی تاکنون نتوانسته است به طور دقیق تمام الگوهای سرطانی را طبقه بندی نماید (۲۵). Tan و همکاران، یک تکنیک طبقه بندی دو مرحله ای ترکیبی برای استخراج قوانین طبقه بندی ارائه دادند. روش پیشنهادی آن‌ها به دقت  $93/04\%$  بر روی پایگاه داده WDBC و  $97/57\%$  بر روی پایگاه WBCD دست یافت (۲۶). Kiyani و همکاران، به بررسی روش‌های GRNN, RBF و PNN بر روی مجموعه داده‌های WBCD پرداختند. نتایج آزمایشات آنها نشان داد که روش RBF دارای دقت  $96/18\%$ ، روش PNN دارای دقت  $98/8\%$  و روش MLP دارای دقت  $95/74\%$  است (۲۷). در (۲۸) کارایی طبقه بندی 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  $94/84\%$  بر روی WBCD و دقت  $92/97\%$  بر روی WDBC و با استفاده از انتخاب ویژگی به دقت  $96/99\%$  بر روی WBCD و  $92/09\%$  بر روی WDBC دست یافت. با استفاده از خصوصیات استخراج شده از آزمایش اسپیراسیون سوزنی و با کمک تکنیک‌های یادگیری ماشین می توان سیستمی کارآمد را برای تشخیص سرطان پستان طراحی نمود که با دقت بالایی خوش خیم 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 اسپیراسیون سوزنی به ارائه مدلی کارا برای تشخیص سرطان پستان می پردازیم.

### 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 بهینه سازی 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.

بدخیم بودن توده سرطان ضروری به نظر می‌رسد (۱۲). اگرچه شایع ترین و قطعی ترین روش تشخیص سرطان پستان، بیوپسی سینه و تشخیص ضایعه با روش‌های معمول آسیب شناسی بافتی است (۱۲) ولی آزمایش اسپیراسیون 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 خوش خیم یا بدخیم بودن توده را تشخیص داد. در مواردی که پزشک با قطعیت نمی‌تواند خوش خیم بودن یا بدخیم بون بیماری را گزارش نماید مدل‌های کامپیوتری و تکنیک‌های داده کاوی 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 به کارگیری الگوریتم‌ها به کامپیوترها این امکان را می‌دهد که کارایی خود را براساس یادگیری، بهینه نمایند. امروزه تکنیک‌های داده کاوی و یادگیری ماشین که به عنوان تکنیک‌هایی برای شناسایی و تشخیص بیماری‌ها و دسته بندی بیماران در مدیریت بیماری به کار می‌روند، ضمن پیدا کردن الگوهای برای تشخیص سریع تر بیماران و جلوگیری از بروز عوارض در آن‌ها، می‌توانند کمک بسیار بزرگی برای پزشکان باشند. در این جا، اهمیت سیستم‌های پشتیبان تصمیم گیری پزشکی (تصمیم یار پزشک) تبیین می‌شود. این سیستم‌ها به پزشکان و متخصصین این امر در تصمیم گیری دقیق تر 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 و کاهش منابع انسانی به عنوان مزایای دیگر سیستم‌های پشتیبان تصمیم گیری پزشکی هستند (۲۴-۱۶).

امروزه روش‌های مختلف هوشمندانه‌ای در جهت تحقق یک مدل ریاضی برای طبقه بندی کردن الگوهای سرطانی انجام گرفته، ولی هیچ

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 <sup>2</sup> / 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)

میانگین، خطای استاندارد و بزرگترین مقدار (میانگین سه تا از بزرگ‌ترین مقادیر) این ویژگی‌ها محاسبه شده و به این ترتیب ۳۰ ویژگی با مقدار عددی حقیقی برای هر نمونه به دست 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 دو گروه آموزش (Train) و آزمون (Test) قرار می‌گیرند. برای بهینه‌سازی پارامترهای مدل، با استفاده از تکنیک اعتبارسنجی

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 ویژگی‌ها نتایج کمی آزمایش اسپیراسیون 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 اسپیراسیون سوزنی توده پستان به دست 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

بدست آمده بر روی مجموعه داده‌های آزمون انجام می‌دهیم. روش اعتبارسنجی متقاطع 4-Fold نمونه‌های آموزشی را به صورت تصادفی به چهار گروه نسبتاً مساوی تقسیم می‌کند و چهار بار عملیات آموزش را برای سه گروه و عملیات آزمون را برای گروه چهارم تکرار می‌کند. دقت کل مدل به وسیله میانگین چهار دقت مجزا در پیش‌بینی‌ها محاسبه می‌شود. در مورد تعیین مقادیر بهینه پارامترهای مدل با استفاده از تکنیک اعتبارسنجی متقاطع 4-fold بر روی داده‌های آموزشی، مقادیر مختلفی را به ازای پارامترهای مدل آزمایش می‌نماییم و مقداری که بالاترین دقت را دارد به عنوان مقدار بهینه در نظر می‌گیریم.

#### مرحله سوم: طبقه‌بندی

در این مرحله با استفاده از روش‌های طبقه‌بندی ناپارامتری مبتنی بر برآورد چگالی هسته‌ای به طبقه‌بندی مجموعه داده‌های آزمون با استفاده از مقادیر بهینه پارامترهای مدل می‌پردازیم. در ادامه روش‌های ناپارامتری استفاده شده در این مقاله شرح داده می‌شوند.

برآورد چگالی ناپارامتری: تابع توزیع چگالی، مفهومی بنیادی در آمار است. متغیر تصادفی  $X$  را در نظر بگیرید که تابع توزیع چگالی آن  $P$  است. اگر تابع  $P$  را داشته باشیم می‌توانیم تخمینی از توزیع  $X$  داشته باشیم. فرض کنید مجموعه‌ای از داده‌های مشاهده شده از نمونه‌ها وجود دارند که تابع توزیع چگالی آن ناشناخته است. برآورد چگالی (Density estimation) به فرآیند تخمین تابع چگالی احتمال یک متغیر تصادفی با استفاده از نمونه‌های مشاهده‌شده از آن متغیر گفته می‌شود. معمولاً برای برآورد تابع چگالی احتمال از روش‌های پارامتری و ناپارامتری استفاده می‌شود. در روش‌های پارامتری فرض می‌شود که داده‌ها از یک توزیع مشخص پیروی می‌کنند و تنها باید مقادیر پارامترهای آن تخمین زده شوند. در برآورد ناپارامتری، فرضیات کمتری درباره‌ی توزیع داده‌های مشاهده شده صورت می‌گیرد. روش‌های طبقه‌بندی پارامتری و ناپارامتری بر اساس قانون بیز عمل می‌کنند.

روش‌های مختلفی برای برآورد چگالی ناپارامتری

متقاطع 4-fold بر روی مجموعه داده‌های آموزشی به تعیین مقدار بهینه پارامترهای مدل می‌پردازیم. پس از آن با استفاده از پارامترهای بهینه تعیین شده، فرآیند یادگیری بر روی داده‌های گروه آموزش انجام می‌شود و مدل مورد نظر با استفاده از داده‌های گروه آزمون سنجیده خواهد شد. مدل ارائه شده دارای سه مرحله پیش پردازش داده‌ها، اعتبارسنجی و طبقه‌بندی است که در ادامه به توضیح هریک از مراحل پرداخته می‌شود.

#### مرحله اول: پیش پردازش داده‌ها

اولین مرحله در ایجاد هر مدلی بر اساس تکنیک‌های داده‌کاوی، مرحله پیش‌پردازش (Preprocessing) می‌باشد که جهت بهبود کیفیت داده‌های واقعی برای داده‌کاوی لازم است. این مرحله شامل برخورد با داده‌های گمشده و نرمال‌سازی داده‌ها است.

در پایگاه داده WBCD، ۱۶ نمونه با مقادیر گمشده (Missing) وجود دارد. در مرحله پیش پردازش داده‌های این پایگاه، ابتدا شماره شناسه بیمار و نمونه‌های دارای مقادیر گمشده را حذف نموده و آزمایشات را با ۶۸۳ نمونه و ۹ ویژگی ادامه می‌دهیم. از ۶۸۳ نمونه مذکور، ۴۴۴ نمونه دارای برچسب خوش‌خیم (۶۵٪) و ۲۳۹ نمونه دارای برچسب بدخیم (۳۵٪) هستند.

در مرحله پیش پردازش داده‌های پایگاه داده WDBC، ابتدا شماره شناسه بیمار را حذف نموده و سپس داده‌ها را نرمال‌سازی می‌نماییم. نرمال‌سازی تغییر مقیاس داده‌ها به گونه‌ای است که آنها را به یک دامنه کوچک و معین نگاشت کند. نرمال‌سازی داده‌ها معمولاً منجر به کسب نتایج بهتر می‌شود. به علت متفاوت بودن مقیاس داده‌های پایگاه WDBC، با استفاده از روش نرمال‌سازی ماکزیمم-می‌نیمم (Min-max) داده‌ها را نرمال‌سازی می‌نماییم.

#### مرحله دوم: اعتبارسنجی

در این مرحله، با استفاده از تکنیک اعتبارسنجی متقاطع 4-fold بر روی مجموعه داده‌های آموزشی به تعیین مقدار بهینه پارامترهای مدل می‌پردازیم و سپس آزمایشات را با مقادیر بهینه

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

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

در این رابطه  $K(\cdot)$  تابع هسته،  $N$  تعداد  
نمونه‌های آموزشی و  $h$  اندازه‌ی پنجره است که به  
آن پارامتر هموارسازی یا پهنا 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 (*Mahalanobis*)  
استفاده کرد تا همبستگی را نیز در نظر  
بگیرد (۲۹). برآوردگر هسته‌ای مبتنی بر فاصله  
ماهالانوبیس به صورت رابطه‌ی زیر تعریف می‌شود:

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

در این رابطه،  $S$  ماتریس کوواریانس نمونه‌ها  
است.

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

$$(۵) \sum_{j=1}^d 1(x_j \neq x_j^t) = \text{Hamming distance}(x, x^t)$$

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

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  $P(\cdot)$  هستند.  
 $\hat{P}(\cdot)$  برآوردگر  $P(\cdot)$  است. برآورد ناپارامتری به  
صورت رابطه زیر تعریف می‌شود:

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

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  $(x_0)$  و  
طول بازه  $(h)$  را انتخاب کرد. این دو انتخاب بر  
روی برآورد تاثیر می‌گذارند. انتخاب طول بازه،  
تعیین کننده‌ی مقدار هموارسازی است و با داشتن  
 $h$ های مختلف، هموارسازی‌های مختلفی وجود  
خواهد داشت. هیستوگرام ارائه شده به نقطه  $x_0$   
انتخابی نیز بستگی دارد و انتخاب‌های مختلف  $x_0$   
می‌تواند نتایج مختلفی داشته باشد. از معایب  
هیستوگرام می‌توان به گسسته بودن و وابستگی  
شکل ارائه شده برای توزیع به  $x_0$  اشاره کرد.

برآوردگر ساده: یکی دیگر از روش‌های برآورد  
چگالی، برآوردگر ساده (Naïve estimator) است.  
با استفاده از این برآوردگر نیاز به انتخاب نقطه‌ی  
 $x_0$  نمی‌باشد. در واقع این برآوردگر یکی از  
روش‌های موجود برای برطرف کردن مشکل  
وابستگی هیستوگرام به نقطه‌ی  $x_0$  است.

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

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

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): برابر است با تعداد بیماران مبتلا به تومور خوش خیم 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 WBDC، ابتدا در مرحله اعتبارسنجی مقادیر بهینه پارامتر  $h$  در روش برآورد چگالی هسته‌ای گوسین و پارامتر  $k$  در روش  $k$  نزدیکترین همسایه را با استفاده از روش اعتبارسنجی متقاطع 4-fold بر روی داده‌های مجموعه آموزش به دست 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  $\{0.1, 0.05, 0.07, 0.08, 0.09, 0.1, 0.11, 0.12, 0.13, 0.14, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2, 3, 4\}$  و برای تعیین مقدار بهینه‌ی پارامتر  $k$  از بازه‌ی  $\{1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 31, 41, 51, 55, 61, 65, 71, 73, 75, 77, 79, 81, 83, 85, 87, 89, 91, 93, 95, 97, 99\}$  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

توجه به چگالی محلی داده‌ها انجام می‌گیرد. این روش، ناحیه‌ی اطراف  $x$  را تا یافتن  $k$  امین همسایه‌ی 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) شاخص‌های 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 مشخص کردن 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 مشخص کردن 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.

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

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

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 مالهالانوبیس) ( $h=0.2$ )	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

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 ۹۶/۲۷٪ است و با استفاده از فاصله مهالانوبیس مقدار بهینه پارامتر  $h=0.2$  دارای دقت ۷۲/۳۴٪ 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، ۱۳ و ۱۵ با دقت ۹۳/۸۸٪ است. در جدول ۴ نتایج بررسی روش‌های غیرپارامتری مبتنی بر برآورد هسته‌ای بر

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 ۹۵/۳۶٪ است و با استفاده از فاصله همینگ مقدار بهینه این پارامتر ۰/۰۵ و دارای دقت ۹۶/۱۰٪ است. بنابراین در ادامه برای بررسی ارزیابی روش هسته‌ای گوسین بر روی داده‌های 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 ۹۶/۵۸٪ است. در جدول ۳ نتایج بررسی روش‌های غیرپارامتری

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 طبقه‌بندی‌کننده‌های یادگیری با نظارت نظیر روش بیزین ساده، SVM-RBF، شبکه عصبی RBF، درخت تصمیم J48 و CART بر روی مجموعه داده‌های پایگاه WBCD پرداختند تا بهترین طبقه‌بندی‌کننده 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 Lavanya و همکاران در سال ۲۰۱۱، کارایی طبقه‌بندی‌کننده 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 WBCD 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 داده‌های مجتمع و مدل آنتروپی ارائه دادند که با استفاده از روش‌های دسته‌بندی SVM، درخت تصمیم C5.0 و رگرسیون لجستیک بر روی پایگاه

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 داده‌های 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 متنوع‌تری را در زمان تشخیص بیماری یا انتخاب درمان در نظر بگیرد (۲۰). این روش‌ها می‌توانند قابلیت تصمیم‌گیری پزشکان را ارتقا بخشند و خطاهای احتمالی ناشی از خستگی یا بی‌تجربگی متخصصین بالینی را کاهش دهند (۲۱)، (۲۲). تحقیقات متعددی توسط محققان برای پیش‌بینی 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 روی پایگاه

استفاده از تکنیک‌های یادگیری ماشین و داده‌کاوی می‌توانند به عنوان یک سیستمی برای 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, Mohammadi 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-

WDBC به ترتیب به دقت ۸۹/۸۶٪، ۹۳/۹۲٪ و ۹۵/۹۵٪ دست یافتند. آنها همچنین با استفاده از روش رگرسیون لجستیک و تکنیک انتخاب ویژگی CfsSubsetEval به دقت ۹۵/۹۵٪ و با استفاده از روش رگرسیون لجستیک و تکنیک انتخاب ویژگی فیلتر به دقت ۹۶/۶۲٪ دست یافتند. در مطالعه آنها روش SVM با تکنیک‌های انتخاب ویژگی فیلتر و CfsSubsetEval با دقت ۸۷/۸۴٪ سرطان پستان را به درستی پیش‌بینی نمود (۳۵).

Salama و همکاران در سال ۲۰۱۲ با استفاده از روش بیزین ساده و درخت تصمیم J48 به ترتیب به دقت ۹۲/۹۷٪ و ۹۳/۱۵٪ بر روی پایگاه داده WDBC دست یافتند (۳۶). Yao و همکاران در سال ۲۰۱۳ با استفاده از روش‌های MARS، RF و RF&MARS بر روی داده‌های WDBC به ترتیب به دقت ۹۶/۲۶٪، ۹۶/۷٪ و ۹۶/۲۹٪ دست یافتند. آنها همچنین نشان دادند که روش درخت تصمیم C4.5 دارای دقت ۹۳/۱۶٪ و روش SVM دارای دقت ۹۵/۸۵٪ است (۳۷). Maldonado و همکاران در سال ۲۰۱۱ با استفاده از تکنیک انتخاب ویژگی Fisher و طبقه‌بندی کننده SVM به دقت ۹۴/۷٪ و با تکنیک انتخاب ویژگی RFE و طبقه‌بندی کننده SVM به دقت ۹۵/۲۵٪ دست یافتند (۳۸).

مدل ارائه شده در این مطالعه، با استفاده از روش برآورد چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی بر روی مجموعه داده‌های WDBC به دقت ۹۷/۹۳٪ دست یافت و با استفاده از روش‌های تخمین چگالی هسته‌ای گوسین مبتنی بر فاصله اقلیدسی و k نزدیکترین همسایه بر روی مجموعه داده‌های WBCD به دقت ۹۸/۱۷٪ دست یافت. بدین ترتیب با استفاده از مدل ارائه شده، دقت شناسایی سیستم‌های تشخیص سرطان پستان افزایش یافت. نتایج مطالعات ذکر شده محققان بر روی داده‌های پایگاه‌های WBCD و WDBC در مقایسه با مطالعه حاضر، حاکی از برتری مدل ارائه شده در این مطالعه است. نتایج این مطالعه نشان داد که روش‌های ناپارامتری برآورد چگالی هسته‌ای می‌توانند با دقت بالایی برای تشخیص سرطان پستان به کار روند. بنابراین

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. *JRNMI*; 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 diseases*; 2015. 7(4):43-51 [Persian].

17. Nahar J, Imam T, Tickle KS, Shawkat Ali ABM, Chen P. Computational 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 pezeski*; 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](mailto:r.sheikhpour@yahoo.com)  
**Razieh Sheikhpour**, Department of Computer Engineering, Yazd University, Yazd, Iran. [r\\_sheikhpour@stu.yazd.ac.ir](mailto: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