

## افزایش دقت پیش‌بینی سرطان پستان با استفاده از الگوریتم ژنتیک و داده‌کاوی

علی محمد لطیف: دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، یزد، ایران  
 محمد مومنی: دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، یزد، ایران  
 رابعه صرام: دانشگاه علوم پزشکی شهید صدوقی یزد، یزد، ایران  
 مهدی آقا صرام\*: دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، یزد، ایران  
 علی پوراحمدی: دانشگاه علم و هنر یزد، یزد، ایران  
 زهرا حاج‌ابراهیمی: دانشگاه علوم پزشکی شهید صدوقی یزد، یزد، ایران

### چکیده

**مقدمه:** سرطان پستان یکی از شایع‌ترین علت مرگ و میر در زنان محسوب می‌شود. پیش‌بینی صحیح سرطان پستان دارای اهمیت است. وجود علائم و ویژگی‌های مختلف این بیماری، تشخیص را برای پزشکان دشوار می‌کند. داده‌کاوی امکان تحلیل داده‌های بالینی بیماران برای تصمیم‌گیری‌های پزشکی را فراهم می‌کند. هدف این مقاله، ارائه یک مدل برای افزایش دقت پیش‌بینی سرطان پستان است.

**روش بررسی:** در این مطالعه، پرونده پزشکی ۵۷۴ بیمار مبتلا به سرطان پستان با تعداد ۳۲ ویژگی مورد بررسی قرار گرفته است. اطلاعات بیماران از پایگاه داده استاندارد بیمارستان فوق تخصصی مرتاض یزد جمع‌آوری شده است. هر یک از بیماران حداقل به مدت یک سال تحت پیگیری بوده‌اند. به منظور ارائه مدل پیش‌بینی سرطان پستان از الگوریتم ژنتیک و داده‌کاوی استفاده می‌شود.

**یافته‌ها:** مدل پیشنهادی با روش‌های درخت تصمیم‌گیری، نایو بیز و نزدیک‌ترین همسایه مورد مقایسه قرار گرفت. نتایج نشان می‌دهد که دقت پیش‌بینی مدل پیشنهادی برابر با ۰/۹۷۳ بوده است. همچنین برای روش‌های نایو بیز، درخت تصمیم‌گیری و نزدیک‌ترین همسایه دقت پیش‌بینی به ترتیب برابر با ۰/۹۱۳، ۰/۹۲۹ و ۰/۹۵۱ می‌باشد.

**نتیجه‌گیری:** در پیش‌بینی سرطان پستان، مدل پیشنهادی نسبت به سایر مدل‌های مورد مقایسه دارای حداقل میزان خطا و بیش‌ترین دقت و صحت است. روش نایو بیز، حداکثر میزان خطا و کم‌ترین دقت را دارا می‌باشد.

**واژه‌های کلیدی:** سرطان پستان، الگوریتم ژنتیک، داده‌کاوی، افزایش دقت پیش‌بینی.

\* نشانی نویسنده پاسخگو: یزد، دانشکده مهندسی برق و کامپیوتر، گروه کامپیوتر، دانشگاه یزد، مهدی آقا صرام.  
 نشانی الکترونیک: sarram@yazd.ac.ir

## مقدمه

۹۳/۶٪، شبکه عصبی مصنوعی با دقت ۹۱/۲٪ و مدل رگرسیون لجستیک با دقت ۸۹/۲٪ شبیه‌سازی گردید. آرونا و همکاران با استفاده از بانک اطلاعاتی<sup>۱</sup> WDBC توسط مدل‌های داده‌کاوی از قبیل درخت تصمیم و بیز ساده و شبکه عصبی به ترتیب با دقت ۹۲/۹۷٪، ۹۲/۶۱٪ و ۹۳/۶۷٪ به پیش‌بینی سرطان پستان رسیدند (۱۱).

استر و همکاران روش تجزیه و تحلیل گسسته خطی در تشخیص سرطان پستان با استفاده از بانک اطلاعاتی WDBC با دقت ۹۶/۸٪ ارایه دادند (۱۲). دقت تشخیص و پیش‌بینی روش‌های داده‌کاوی را می‌توان با استفاده از الگوریتم ژنتیک افزایش داد. الگوریتم ژنتیک یکی از زیر گروه‌های محاسبات فرا ابتکاری است که از قوانین تکامل بیولوژیک طبیعی تبعیت می‌کند. الگوریتم ژنتیک با استفاده از قانون بقای برترین‌ها، در یک زیرمجموعه از پاسخ‌های مساله، به دنبال به‌دست آوردن پاسخ‌های بهتر است. یک زیر مجموعه از پاسخ‌های ممکن برای مساله، جمعیت اولیه را تشکیل می‌دهد. متناسب با ارزش هر یک از پاسخ‌ها، فرآیند انتخاب از جمعیت اولیه و تولید مثل برای ایجاد نسل جدید انجام می‌شود. در هر نسل با ترکیب و تولیدمثل پاسخ‌های انتخاب شده، به کمک عملگرهایی که از ژنتیک طبیعی پیروی می‌کنند، تقریب‌های بهتری از جواب نهایی به‌دست می‌آید. این فرایند باعث می‌شود که نسل‌های جدید با شرایط مساله سازگارتر باشد (۱۳).

در این مطالعه با طراحی پرسش‌نامه، تکمیل و گردآوری مجموعه داده‌های موبوط به پرونده‌های بیماران در بیمارستان فوق تخصصی مرتاض یزد، مدلی برای افزایش دقت تشخیص و پیش‌بینی روش‌های داده‌کاوی توسط الگوریتم ژنتیک معرفی می‌گردد.

## مواد و روش‌ها

## الف- گردآوری مجموعه داده‌ها:

برای تهیه مجموعه داده‌های مربوط به سرطان پستان، در ابتدا پرسش‌نامه استاندارد بر اساس ویژگی‌های جدید طراحی شده، سپس در بیمارستان فوق تخصصی مرتاض یزد تکمیل و جمع‌آوری گردیده است. داده‌ها مربوط به سال‌های ۱۳۹۳ تا ۱۳۹۴ است. این مجموعه داده شامل ۵۷۴ نمونه بوده که تعداد ۱۷ نمونه فاقد اطلاعات کامل

سرطان پستان تهدید بزرگی بر سلامت زنان بوده و از عوامل شایع در کاهش عمر زنان به شمار می‌رود (۱). سرطان پستان، ناشی از رشد خارج از قاعده سلول‌های غیرطبیعی در پستان است (۲). برای پیش‌بینی، تشخیص و درمان این بیماری عوامل بسیاری از قبیل وجود تومور، درگیری غدد لنفاوی، تو رفتگی نوک پستان، بروز ترشح در پستان و ... استفاده می‌شود (۳). وجود شباهت زیاد در علائم بالینی و آزمایشگاهی سرطان پستان احتمال تشخیص نادرست را افزایش می‌دهد (۴). توده، شایع‌ترین علامت سرطان پستان می‌باشد که در اغلب موارد توسط خود بیمار بصورت اتفاقی کشف می‌شود و در بقیه موارد توسط پزشک در معاینه بالینی مشخص می‌شود. این توده ممکن است دردناک باشد ولی در اغلب موارد بدون درد است. در بعضی موارد سرطان پستان بصورت توده‌های متعدد بروز می‌کند (۵-۷). تشابه علائم بالینی و آزمایشگاهی سرطان پستان، احتمال بروز خطا در تشخیص را افزایش می‌دهد. تشخیص و پیش‌بینی انواع بیماری‌ها با استفاده از تکنیک‌های داده‌کاوی امکان‌پذیر است. داده‌کاوی در پزشکی به فرایند استخراج اطلاعات معتبر از پیش‌شناخته، قابل فهم و قابل اعتماد از پایگاه داده‌های پزشکی و استفاده از آن جهت پیش‌بینی، تشخیص و کمک به درمان بیماری گفته می‌شود. کشف الگوهای مفید بین بیماری و علائم بالینی و آزمایشگاهی بیمار از کاربردهای داده‌کاوی در پزشکی است. منظور از الگوی مفید، مدلی در داده‌ها است که ارتباط میان یک زیرمجموعه از داده‌های بیمار و تشخیص بیماری را بیان می‌کند (۸). در سال‌های اخیر با پیشرفت‌هایی که در تشخیص زودرس این بیماری به وجود آمده، درمان آن نیز با موفقیت بیشتری همراه شده است. اگر توده‌های پستان در اندازه کوچک کشف شوند به خوبی قابل درمان هستند. داده‌کاوی از روش‌های جدید برای تشخیص زودرس سرطان پستان است. شیخ‌پور و همکاران تشخیص سرطان پستان با استفاده از کاهش دو مرحله‌ای ویژگی‌های استخراج شده اسپیراسیون سوزنی و الگوریتم‌های داده‌کاوی را معرفی کردند (۹). استفاده از سه تکنیک داده‌کاوی برای تشخیص سرطان پستان توسط دلن و همکاران معرفی شد که پیش‌بینی سرطان توسط درخت تصمیم (C5) با دقت

<sup>1</sup> Wisconsin Diagnostic Breast Cancer

۲- در هر تکرار الگوریتم، جمعیت توسط تابع برازندگی ارزیابی می‌شود. سپس تعدادی از بهترین کاندیداها برای نسل بعد گزینش می‌شوند و جمعیت جدید را تشکیل می‌دهند.

۳- تعدادی از این جمعیت با استفاده از اپراتورهای ژنتیکی نظیر تقاطع<sup>۳</sup> و جهش<sup>۴</sup> برای تولید فرزندان جدید استفاده می‌شوند.

۴- مراحل فوق تا رسیدن به یک پاسخ مناسب ادامه می‌یابد.

مراحل مطرح شده برای اجرای الگوریتم ژنتیک در قالب یک روندنما در شکل ۱ مشاهده می‌شود.

### مدل پیشنهادی

هر یک از ویژگی‌ها و یافته‌ها در تشخیص و پیش‌بینی سرطان پستان از اهمیت خاصی برخوردار هستند. به بیان دیگر همه ویژگی‌ها دارای ارزش یکسان نیستند. به عنوان مثال در تشخیص بیماری دو ویژگی BMI و وجود تومور دارای اهمیت متفاوتی هستند. این که هر یک از ویژگی‌ها دارای چه ارزشی است و چقدر در تشخیص بیماری نقش دارد، مساله مهمی است. در این مقاله روشی ارائه می‌شود که ارزش و نقش هر یک از ویژگی‌ها به طور دقیق مشخص شده و بیماری تشخیص داده می‌شود.

در روش پیشنهادی برای مشخص کردن ارزش و نقش هر یک از ویژگی‌ها در تشخیص بیماری از الگوریتم ژنتیک استفاده می‌شود. برای هر یک از ویژگی‌ها، یک ژن<sup>۵</sup> تعریف می‌شود. برای مقداردهی ژن‌ها، به طور تصادفی برای هر ژن یک عدد در بازه (۰-۱) اختصاص داده می‌شود که نشان دهنده درجه اهمیت ویژگی متناظر با آن ژن است. هر اندازه مقدار ژن بزرگ‌تر باشد، نشان‌دهنده ارزش و اهمیت بیشتر ویژگی متناظر است. در جدول ۳ نحوه وزن‌دار کردن ویژگی‌ها بر اساس ژن متناظر نشان داده شده است.

می‌باشد. نمونه‌هایی که فاقد اطلاعات کامل بودند از طریق روش بیش‌ترین فراوانی تخمین زده شدند. برای هر بیمار تعداد ۳۲ ویژگی ثبت شده است. بازه مقادیر اولیه ویژگی‌های بالینی بیمار در جدول ۱ نمایش داده شده است. اسامی کلاس‌ها و تعداد نمونه‌های موجود در هر کلاس در جدول ۲ نشان داده شده است.

قالب مناسب داده‌ها به عنوان ورودی داده‌کاوی در نتایج و خروجی تاثیرگذار است. اگر مقادیر ویژگی‌های مجموعه داده در دامنه متفاوتی قرار داشته باشند، احتمال بروز خطا در یافته‌ها افزایش می‌یابد. به قرار دادن داده‌های یک جامعه آماری در دامنه مشابه، نرمال‌سازی گفته می‌شود (۱۴). در مدل پیشنهادی نحوه نرمال‌سازی به روش Max/Min و در بازه (۰-۱) است (۱۵).

### ب- الگوریتم ژنتیک:

الگوریتم ژنتیک در سال ۱۹۶۲ میلادی توسط جان هلند<sup>۲</sup> ارائه شد. این الگوریتم در گروه الگوریتم‌های بهینه‌سازی تصادفی قرار دارد و برای بهینه‌سازی مسایل پیچیده با فضای جست‌وجوی ناشناخته مناسب است (۱۶).

ایده اصلی الگوریتم ژنتیک از نظریه تکاملی داروین گرفته شده است. نظریه داروین به این شرح است که آن دسته از صفات طبیعی که با قوانین طبیعی سازگاری بیش‌تری دارند، شانس بقای بیش‌تری دارند. شایان ذکر است که نظریه تکاملی داروین هیچ اثبات تحلیلی و قطعی ندارد؛ اما از نظر تجربی و آماری تأیید شده است (۱۷).

افراد جدید یک جامعه از طریق زاد و ولد تولید می‌شوند. شانس بقای یک فرد در نسل جدید به ترکیب خاص کروموزومی وابسته است. در مراحل زاد و ولد ممکن است جهش‌هایی در خصوصیات یک فرد نسل جدید رخ دهد که در نتیجه موجودی با خصوصیات عالی و سازگاری بالا تولید شود. در روند زاد و ولد به گونه‌های برتر در هر نسل اجازه تولید مثل داده می‌شود و گونه‌های نامطلوب به تدریج از بین خواهند رفت و افراد نسل‌های جدید با گذشت زمان تکامل می‌یابند. الگوریتم ژنتیک در زیر به صورت خلاصه بیان شده است.

۱- مجموعه‌های تصادفی از کاندیداها را جواب به عنوان جمعیت اولیه تولید می‌شوند، تولید و در هر نسل با کاندیداها جدیدی جایگزین می‌شوند.

<sup>3</sup> Crossover

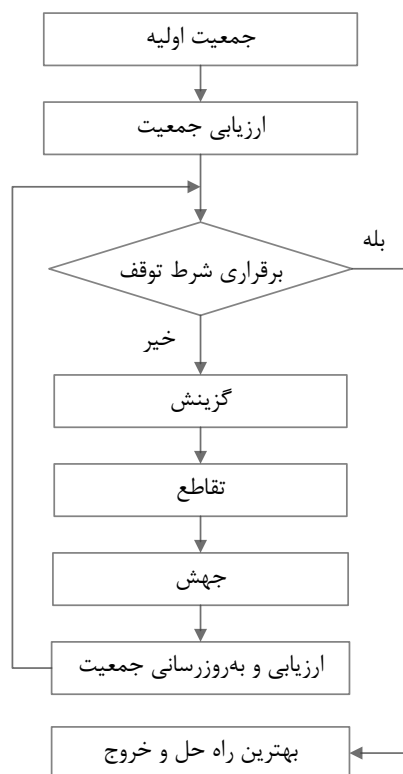
<sup>4</sup> Mutation

<sup>5</sup> Gene

<sup>2</sup> John Holland

جدول ۱: بازه مقادیر ویژگی‌های بالینی بیماران

بازه مقادیر در مجموعه داده‌ها		ویژگی	بازه مقادیر در مجموعه داده‌ها		ویژگی
ندارد = ۰	۱ = دارد و قدیمی است.	۲۱	[۲۵ ۶۷]	۱	سن
۲ = دارد و جدید است.		N/R	[۰ ۵۱]	۲	مدت تاهل
کم = ۱	ندارد = ۰	۲۲	[۱۱۵ ۱۸۹]	۳	قد
۳ = زیاد	۲ = متوسط	درد پستان	[۴۹ ۱۱۰]	۴	وزن
ندارد = ۰	۱ = در یک ناحیه	۲۳	[۱۶ ۳۷]	۵	BMI
۲ = منتشر		ناحیه درد در پستان	[۰ ۹]	۶	تعداد زایمان طبیعی
ندارد = ۰	۱ = دارد (ترشح خود به خود)	۲۴	[۰ ۵]	۷	تعداد زایمان سزارین
۲ = دارد (ترشح با فشار)		ترشح در پستان	[۰ ۳]	۸	تعداد سقط جنین
ندارد = ۰	۱ = خونی	۲۵	[۸ ۱۶]	۹	سن اولین قاعدگی
۲ = سفید و شیری	۳ = سبز و سیاه	رنگ ترشح پستان	[۱۶ ۳۷]	۱۰	سن اولین زایمان طبیعی
۴ = قهوه‌ای یا صورتی	۵ = بیرنگ		[۱۸ ۴۱]	۱۱	سن اولین زایمان سزارین
ندارد = ۰	۱ = دارد	۲۶	[۱۵ ۳۲]	۱۲	سن اولین سقط جنین
۱ = تومور در پستان سمت راست	۲ = تومور در پستان سمت چپ	۲۷	۰ = سابقه ندارد	۱۳	سابقه نازایی بیمار
۳ = تومور دوطرفه		سمت قرارگیری تومور	۱ = سابقه دارد	۱۴	سن شروع یائسگی
[۰ ۶]		۲۸	۰ = وجود ندارد		
ندارد = ۰	۱ = دارد	۲۹	۱ = فامیل درجه ۳	۱۵	سابقه فامیلی ابتلا به سرطان پستان در بیمار
۱ = دارد		سفتی قابل لمس در پستان	۲ = دایی یا عمو		
۰ = شیردهی نداشته است.	۱ = یک سال	۳۰	۳ = عمه یا خاله		
۲ = دو سال	۳ = سه سال	مدت شیردهی	۴ = برادر		
۴ = چهار سال	۵ = پنج و بیش از پنج سال		۵ = خواهر		
۰ = قاعدگی قطع نشده است	۱ = یائسگی طبیعی	۳۱	۶ = فرزند		
۲ = یائسگی زودرس	۳ = برداشتن رحم	علت قطع قاعدگی طبیعی	۷ = پدر		
۳ = برداشتن رحم	۴ = سایر علل		۸ = مادر		
ندارد = ۰	۱ = بدخیمی درجه یک	۳۲	۰ = کمتر از یک سال	۱۸	مدت مصرف قرص ضد بارداری بیمار
۲ = بدخیمی درجه دو	۳ = بدخیمی درجه سه	درجه بدخیمی	۱ = از ۱ تا ۵ سال		
۴ = بدخیمی درجه چهار	۵ = بدخیمی درجه پنج		۲ = از ۶ تا ۱۰ سال		
			۳ = از ۱۱ تا ۱۵ سال		
			۴ = بیش از ۱۵ سال		
			۰ = ندارد	۱۹	سابقه هیستریکتومی بیمار
			۱ = دارد		
			۰ = ندارد	۲۰	جایگزینی HRT
			۱ = از ۱ تا ۵ سال دارد		
			۲ = از ۶ تا ۱۰ سال دارد		
			۳ = از ۱۱ تا ۱۵ سال دارد		
			۴ = بیش از ۱۵ سال دارد		



شکل ۱: روندنمای الگوریتم ژنتیک

جدول ۲: اسامی و تعداد نمونه‌های کلاس‌ها

تعداد نمونه‌ها	نام کلاس
۱۸۳	سالم
۸۵	درجه بدخیمی یک
۱۱۴	درجه بدخیمی دو
۱۳۰	درجه بدخیمی سه
۴۵	درجه بدخیمی چهار
۱۷	درجه بدخیمی پنج

جدول ۳: وزن دار کردن ویژگی‌ها

ویژگی	مقدار اولیه	مقدار نرمال	وزن	مقدار وزن دار
سن	۵۵	۰/۷۱	۰/۳	۰/۲۱
قد	۱۵۶	۰/۵۵	۰/۸	۰/۴۴
تعداد زایمان طبیعی	۳	۰/۳۳	۰/۲	۰/۰۷
...	...	...	...	...
BMI	۳۴/۹	۰/۹	۰/۷	۰/۶۳

تابع برازندگی نشان دهنده شایستگی یا توانایی هر کروموزوم است. با فرض متغیرهای تعریف شده در جدول ۱، تابع برازندگی در شکل ۲ نمایش داده شده است. عمل کلاسه‌بندی با مقادیر ویژگی‌های وزن دار، که بر اساس هریک از کروموزوم‌ها محاسبه می‌گردد، به روش نزدیک‌ترین همسایه (۱۸) انجام می‌شود. هر کروموزوم که کلاسه‌بندی را با خطای پیش‌بینی کم‌تری انجام دهد، شایسته‌تر خواهد بود. تخمین خطا بر مبنای نمونه‌گیری به روش 10-Fold Cross Validation (۱۹) انجام می‌شود. عملکرد انتخاب، تعدادی کروموزوم را برای تولید مثل از جمعیت انتخاب می‌کند. در این مدل از روش انتخاب نخبگرا (Elitist Selection) استفاده شده است. در این

به ازای هر یک از کروموزوم‌های موجود در جمعیت، مراحل زیر تکرار می‌شود:

- ۱- مجموعه داده‌ها فراخوانی می‌شود.
- ۲- یک رکورد از مجموعه داده انتخاب می‌شود.
- ۳- مقادیر ویژگی‌های رکورد نرمال می‌شود.
- ۴- مقدار ژن مربوط به هر ویژگی از کروموزوم استخراج می‌شود.
- ۵- به طور متناظر هر یک از ویژگی‌ها در ژن مخصوص به خود ضرب شده تا مقدار ویژگی‌ها وزن دار شوند.
- ۶- مراحل یک تا پنج برای همه رکوردها تکرار می‌شود.
- ۷- بر اساس مقادیر وزن دار ویژگی‌ها، عمل کلاسه‌بندی انجام می‌شود.

کلاس‌های واقعی و کلاس‌های پیش‌بینی شده با استفاده از ماتریس Confusion قابل محاسبه است. در شکل ۵، پارامترهای مورد نیاز ماتریس Confusion ذکر شده است. برای مقایسه مدل پیشنهادی با سایر روش‌ها از معیارهای Accuracy، Sensitivity، Specificity، Precision و F-Measure با توجه به شکل ۵ طبق روابط زیر استفاده می‌شود (۱۵):

$$\text{Accuracy} = (\text{TP} + \text{TN}) / \text{All} \quad (1)$$

$$\text{Sensitivity} = \text{TP} / (\text{TP} + \text{FN}) \quad (2)$$

$$\text{Specificity} = \text{TN} / (\text{FP} + \text{TN}) \quad (3)$$

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP}) \quad (4)$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN}) \quad (5)$$

$$\text{F Measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (6)$$

روش بهترین کروموزوم‌ها در هر نسل برای تولید مثل انتخاب می‌شوند. انتخاب نخبه‌گان به میزان قابل ملاحظه‌ای کارایی الگوریتم ژنتیک را افزایش می‌دهد (۲۰). در عمل تقاطع به صورت تصادفی بخش‌هایی از کروموزوم‌ها با یکدیگر ترکیب شده و فرزند جدید متولد می‌شود. این موضوع باعث ارتزبری فرزندان از خصوصیات والدین خود می‌شود. در شکل ۳ نحوه تولید فرزند جدید نمایش داده شده است. عملگر جهش پس از اتمام عمل تقاطع بر روی فرزندان جدید اعمال می‌شود. این عملگر یک ژن از یک کروموزوم را به طور تصادفی انتخاب نموده و سپس محتوای آن ژن را تغییر می‌دهد. شکل ۴ نشان‌دهنده نحوه جهش می‌باشد.

پس از اتمام عمل جهش، کروموزوم‌های تولید شده به عنوان نسل جدید در نظر گرفته می‌شوند. پس از گذشت چندین نسل، مقادیر کروموزوم‌ها همگرا شده و پاسخ نهایی به دست می‌آید.

مدل پیشنهادی با سه روش نایو بیس (۲۱)، درخت تصمیم (۲۲) و نزدیک‌ترین همسایه مقایسه شده است. ارتباط بین

#### ALGORITHM

- 1) Read the training data from a file
- 2) Read the testing data from a file
- 3) Normalize the attribute values in the range of 0 to 1.
- 4) Let  $x_1, x_2, \dots, x_m$  denote the  $m$  instances from data set  $\{f_1, f_2, \dots, f_n\}$ ,  $n$  = Number of features
- 5) Let  $Ch$  denote the current chromosome from population  $\{g_1, g_2, \dots, g_r\}$ ,  $r$  = Number of genes
- 6) Assign weight  $Ch$  to each instance  $x_i$  in the training set
- 7) Train the weights on the whole training data set
  - For every training instance
    - Calculate the weighted value as
    - $Ch_j * x_{ij}$ , where  $j$  is the attribute
    - Find the  $K$  nearest neighbors based on the Euclidean distance
    - Calculate the class value
  - End for
- 8) For each testing instance in the testing data set
  - Find the  $K$  nearest neighbors in the training data set based on the Euclidean distance
  - Predict the class value by finding the maximum class represented in the  $K$  nearest neighbors
- End for
- 9) Calculate the error rate as
 
$$\text{Error Rate} = 1 - (\# \text{ of correctly classified examples} / \text{All}) * 100$$
- 10) Fitness Function = Minimize (Error Rate)

شکل ۲: تابع برازندگی

- ۱- بردار  $r$  در ابعاد  $(1 \times 34)$  به طور تصادفی در بازه  $[0, 1]$  مقداردهی می‌گردد.  
 ۲- به صورت تصادفی دو کروموزوم  $Ch_1$  و  $Ch_2$  از جمعیت انتخاب می‌شوند.  
 ۳-  $فرزند\ جدید\ اول = r * Ch_1 + (1-r) * Ch_2$   
 ۴-  $فرزند\ جدید\ دوم = (1-r) * Ch_1 + r * Ch_2$

شکل ۳: نحوه تولید فرزند جدید

- ۱- بردار  $r$  در ابعاد یک کروموزوم  $(1 \times 34)$  تعریف می‌شود.  
 ۲- بردار  $r$  با توزیع نرمال در بازه  $[-1, 1]$  مقداردهی می‌گردد.  
 ۳- به صورت تصادفی یک کروموزوم از جمعیت انتخاب می‌شوند.  
 ۴- کروموزوم جهش‌یافته از جمع کروموزوم انتخاب شده با  $r$  ایجاد می‌شود.

شکل ۴: نحوه اعمال عملگر جهش در روش پیشنهادی

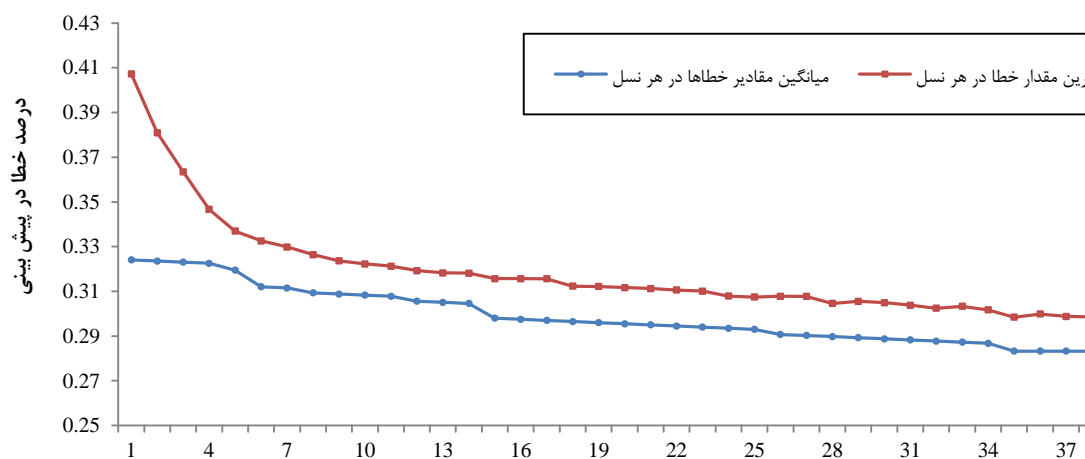
- TP: تعداد رکوردهایی که به درستی، مثبت تشخیص داده می‌شوند.  
 TN: تعداد رکوردهایی که به درستی، منفی تشخیص داده می‌شوند.  
 FP: تعداد رکوردهایی که به غلط، مثبت تشخیص داده می‌شوند.  
 FN: تعداد رکوردهایی که به غلط، منفی تشخیص داده می‌شوند.

شکل ۵: پارامترهای مورد نیاز برای ارتباط بین کلاس‌های واقعی و کلاس‌های پیش‌بینی شده

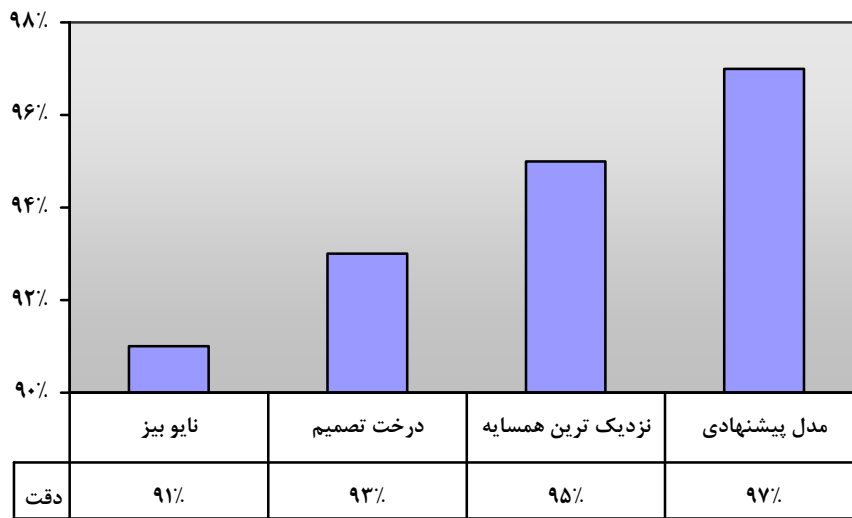
## یافته‌ها

معیارهای Sensitivity و Specificity به ترتیب در جدول‌های ۴ و ۵ نمایش داده شده‌است. جدول ۶ و ۷ نتایج روش‌ها به ترتیب با معیار Precision و F-Measure مورد مقایسه قرار گرفته است. نتایج مقایسه نشان‌دهنده برتری عملکرد مدل پیشنهادی می‌باشد. در جدول ۸، روش پیشنهادی توسط معیارهای F-Measure، Precision، Sensitivity، Specificity با سایر روش‌ها مقایسه شده است. مقادیر جدول نشان‌دهنده عملکرد بهتر روش پیشنهادی است.

داده‌های موجود در پرونده بیماران با استفاده از نرم‌افزار Matlab (R2013b, The Mathworks Inc., USA) توصیف، شبیه‌سازی و تحلیل شده است. در شکل ۶ درصد خطا برای پیش‌بینی بیماری در الگوریتم پیشنهادی در ۴۰ نسل نمایش داده شده است. شکل ۷ نشان‌دهنده نمودار نتایج تشخیص روش‌های مختلف با معیار Accuracy است. همان‌گونه که مشاهده می‌شود مدل پیشنهادی دقت بیشتری نسبت به سایر روش‌ها دارد. هم‌چنین مقایسه نتایج پیش‌بینی بیماری با



شکل ۶: درصد خطا برای پیش‌بینی بیماری در الگوریتم پیشنهادی



شکل ۷: نمودار نتایج با معیار Accuracy

جدول ۴: مقایسه نتایج با معیار Sensitivity

مدل پیشنهادی	نزدیک ترین همسایه	درخت تصمیم	نایو بیز	
۰/۹۴۸	۰/۸۵۰	۰/۸۳۳	۰/۸۳۳	سالم
۰/۹۸۱	۰/۹۷۲	۰/۹۵۴	۰/۹۸۱	بدخیمی درجه یک
۰/۹۸۰	۰/۹۶۷	۰/۹۵۳	۰/۹۲۵	بدخیمی درجه دو
۰/۹۸۰	۰/۹۸۱	۰/۹۲۳	۰/۹۲۵	بدخیمی درجه سه
۰/۹۶۱	۰/۹۶۱	۰/۸۹۹	۰/۸۹۹	بدخیمی درجه چهار
۰/۹۸۲	۰/۹۸۱	۰/۹۳۱	۰/۸۳۱	بدخیمی درجه پنج

جدول ۵: مقایسه نتایج با معیار Specificity

مدل پیشنهادی	نزدیک ترین همسایه	درخت تصمیم	نایو بیز	
۰/۹۷۸	۰/۹۷۵	۰/۹۶۰	۰/۹۶۸	سالم
۰/۹۸۰	۰/۹۸۰	۰/۹۶۵	۰/۹۳۴	بدخیمی درجه یک
۰/۹۸۲	۰/۹۸۱	۰/۹۷۴	۰/۹۷۸	بدخیمی درجه دو
۰/۹۸۰	۰/۹۸۰	۰/۹۷۸	۰/۹۸۰	بدخیمی درجه سه
۰/۹۷۵	۰/۹۵۳	۰/۹۶۲	۰/۹۶۸	بدخیمی درجه چهار
۰/۹۸۱	۰/۹۸۱	۰/۹۷۸	۰/۹۷۵	بدخیمی درجه پنج

جدول ۶: مقایسه نتایج با معیار Precision

مدل پیشنهادی	نزدیک ترین همسایه	درخت تصمیم	نایو بیز	
۰/۹۶۴	۰/۹۴۵	۰/۸۹۸	۰/۹۱۰	سالم
۰/۹۸۱	۰/۹۸۰	۰/۹۴۶	۰/۸۸۱	بدخیمی درجه یک
۰/۹۸۰	۰/۹۸۱	۰/۹۵۳	۰/۹۶۷	بدخیمی درجه دو
۰/۹۸۰	۰/۹۸۰	۰/۹۶۱	۰/۹۸۱	بدخیمی درجه سه
۰/۹۴۱	۰/۸۲۳	۰/۸۶۳	۰/۸۹۹	بدخیمی درجه چهار
۰/۹۸۱	۰/۹۸۲	۰/۹۳۱	۰/۸۷۶	بدخیمی درجه پنج

جدول ۷: مقایسه نتایج با معیار F-Measure

مدل پیشنهادی	نزدیک ترین همسایه	درخت تصمیم	نایو بیز	
۰/۹۵۶	۰/۸۹۵	۰/۸۹۰	۰/۸۷۰	سالم
۰/۹۸۱	۰/۹۷۷	۰/۹۵۰	۰/۹۳۰	بدخیمی درجه یک
۰/۹۸۰	۰/۹۷۴	۰/۹۵۳	۰/۹۴۶	بدخیمی درجه دو
۰/۹۸۰	۰/۹۸۱	۰/۹۴۲	۰/۹۵۱	بدخیمی درجه سه
۰/۹۵۱	۰/۸۸۷	۰/۸۸۱	۰/۸۹۹	بدخیمی درجه چهار
۰/۹۸۲	۰/۹۸۱	۰/۹۳۱	۰/۸۵۳	بدخیمی درجه پنج

جدول ۸: مقایسه کلی روش‌های داده‌کاوی

Precision	Sensitivity	Specifity	F-Measure	
۰/۹۱۹	۰/۸۹۹	۰/۹۶۷	۰/۹۰۸	نایو بیز
۰/۹۲۵	۰/۹۲۴	۰/۹۷۰	۰/۹۲۵	درخت تصمیم
۰/۹۴۹	۰/۹۵۲	۰/۹۷۵	۰/۹۴۹	نزدیک ترین همسایه
۰/۹۷۲	۰/۹۷۲	۰/۹۷۹	۰/۹۷۲	مدل پیشنهادی

## بحث

در این مطالعه، مجموعه‌ای از الگوهای کمتر شناخته شده و موثر در بروز سرطان پستان بر اساس یک پایگاه داده بومی مورد پردازش قرار گرفته است. طراحی و تکمیل پرسش‌نامه‌ها، ثبت علائم بالینی و نتایج آزمایشگاهی در مجموعه داده‌های مورد استفاده، توسط نویسندگان این مقاله در بیمارستان فوق تخصصی مرتاض یزد جمع‌آوری شده است. پژوهش‌های مربوط به پیش‌بینی سرطان پستان محدودیت‌هایی از قبیل تعداد کم بیماران برای

ایجاد مدل، داده‌های از دست رفته و متغیرهای ناقص را دارا می‌باشند. در این پژوهش تعداد قابل قبولی از بیماران با متغیرهای مناسب با حداقل داده‌های از دست رفته به کار گرفته شده است. مطالعه حاضر با به‌کارگیری ویژگی‌های بالینی و آزمایشگاهی بیماران با استفاده الگوریتم ژنتیک و داده‌کاوی، سرطان پستان را پیش‌بینی می‌کند. مدل پیشنهادی این مطالعه شامل نرمال‌سازی ویژگی‌ها، انتخاب ویژگی‌ها، وزن‌دهی به ویژگی‌ها و معرفی روشی برای پیش‌بینی سرطان پستان با استفاده از ویژگی‌های وزن‌دار است.

۸۹ ویژگی‌های توده، اندازه تومور، سابقه فامیلی، میزان تصویر ماموگرافی قبلی نسبت به کنونی بیمار بالاترین کارایی در دسته‌بندی مدل پیشنهادی دارا بودند. همچنین ویژگی‌هایی نوع زایمان، درد پستان و سن اولین سقط جنین حداقل تاثیر در دسته‌بندی را داشتند. متغیرهای نوع روغن خوراکی مصرفی، جنس پارچه سوتین و استفاده از اسپری زیر بغل هیچ تاثیری در تشخیص و پیش‌بینی درجه بدخیمی سرطان پستان را نداشتند. در این مطالعه علاوه بر شناسایی مهم‌ترین ویژگی‌ها، با استفاده از انتخاب ویژگی‌ها بر اساس الگوریتم ژنتیک به عملکردی بهتر بر حسب شاخص‌های دقت، حساسیت و ویژگی دست یافتیم.

آرونا و همکاران با استفاده از دسته‌بندی کننده‌های با ناظر مانند روش نایو بیز و درخت تصمیم بر روی مجموعه داده‌های پایگاه داده WBCD به پیش‌بینی سرطان پستان پرداختند. بر اساس معیارهای Sensitivity, Accuracy, Precision, Specificity و Recall در نتایج شبیه‌سازی، مدل پیشنهادی نتایج بهتری نسبت به این مقاله دارد (۲۴). لند و ورهگن از استفاده از ماشین بردار پشتیبان بر روی مجموعه داده‌های سرطان پستان برای پیش‌بینی استفاده کردند (۲۵). نتایج آزمایشات نشان داد که ماشین بردار پشتیبان دقت ۹۶/۷٪ را دارد. کیان و همکاران با به‌کارگیری روش RBF و MLP برای پیش‌بینی سرطان پستان به ترتیب به دقت ۹۶/۱۸٪ و ۹۵/۷۴٪ رسیدند (۲۶). چاراسیا و همکاران در (۲۷) با استفاده از روش SVM به دقت ۹۶/۴٪ دست یافتند. سروستانی و همکاران برای پیش‌بینی درجه بدخیمی سرطان پستان به مقایسه میانگین مربع خطا در شبکه‌های عصبی چند لایه، رقابتی و پایه شعاع پرداختند که بهترین دقت را شبکه عصبی پایه شعاعی داشت (۲۸). لاوانیا و همکاران با استفاده از داده‌های پایگاه داده WBCD و درخت تصمیم با دسته‌بندی دو مرحله‌ای به دقت ۹۴/۸۴٪ رسیدند (۲۹).

طلوعی و همکاران در (۳۰) پیش‌بینی عود مجدد سرطان پستان به کمک سه تکنیک داده کاوی را ارائه دادند. بررسی‌های صورت گرفته نشان می‌دهد که دقت در سه الگوریتم داده کاوی، یعنی درخت تصمیم گیری، ANN و SVM به ترتیب برابر ۰/۹۳۶، ۰/۹۴۷، ۰/۹۵۷ بوده است. کیانی و همکاران در مدل توسعه یافته (۳۱) به ویژگی و حساسیت به ترتیب ۶۵٪ و ۹۶٪ رسیدند. در این مطالعه گذشته‌نگر، از داده‌های ۸۰۹ بیمار مبتلا به سرطان پستان و

۸۹ ویژگی از هر بیمار، استفاده شد. آتشی و همکاران در (۳۲) تعداد ۱۰۰ رابطه انجمنی با ضرایب اطمینان بالاتر از ۹۰٪ کشف کردند. تعداد ۱۰ رابطه از این روابط معنی‌دار گزارش شد. کاوراسیا و همکاران در (۳۲) از درخت تصمیم جهت پیش‌بینی سرطان پستان با دقت ۷۴٪ استفاده کردند. دقت ۸۷٪ در تشخیص سرطان پستان با استفاده از درخت C4.5 نتیجه کار سوریا و همکاران بود (۳۳).

سلاما و همکاران روش شبکه فازی-عصبی را برای پیش‌بینی سرطان پستان به کار بردند. دقت پیش‌بینی این روش ۹۵/۰۶٪ بود (۳۴). در حالی که مدل پیشنهادی با دقت ۹۷/۳٪، حساسیت ۹۷/۹٪ و ویژگی ۹۷/۲٪ دارای عملکرد بهتری نسبت به روش نایو بیز، درخت تصمیم و نزدیک‌ترین در تشخیص سرطان پستان است.

در روش پیشنهادی با استفاده از کاهش تعداد متغیرها و وزن‌دار کردن ویژگی‌ها با به‌کارگیری الگوریتم ژنتیک برای افزایش دقت با هدف طراحی و ارزیابی یک مدل پزشک‌یار در تعیین درجه بدخیمی سرطان پستان انجام شد. مدل پزشک‌یار طراحی شده در این پژوهش در تشخیص درجه بدخیمی موفق بوده است و دسته‌بندی را با دقت قابل قبولی انجام می‌گردد. آزمایشات و شبیه‌سازی نشان داد سیستم پزشک‌یار معرفی شده در این پژوهش بر روی مجموعه داده بیماران بومی مبتلا به سرطان پستان بیمارستان مرتاض یزد به دقت ۹۷/۳٪ رسیده است که بالاتر از تحقیقات مشابه بر روی مجموعه داده‌های متفاوت بوده است.

### نتیجه‌گیری

پیش‌بینی و تشخیص صحیح سرطان پستان با استفاده از هوش مصنوعی و یادگیری ماشین، شانس درمان موفق را بالا می‌برد. در این مقاله برای پیش‌بینی و تشخیص سرطان پستان، از الگوریتم ژنتیک برای بهینه‌سازی نتایج داده‌کاوی استفاده شد و یک مدل جدید ارائه گردید. نتایج شبیه‌سازی نشان می‌دهد که مدل پیشنهادی با دقت پیش‌بینی ۰/۹۷۳ از روش‌های نایو بیز، درخت تصمیم و نزدیک‌ترین همسایه دقت بیشتری دارد. دقت بالا در تشخیص سرطان حاکی از برتری رهیافت پیشنهادی است. پیچیدگی و زمان‌بر بودن مدت اجرا از نقاط ضعف این روش می‌باشد.

## References

1. Parkin DM. Estimates of the worldwide Incidence of Major Cancer in 1990. *International Journal of Cancer* 1999; 80: 827-41.
2. David J, Ashly. Evan's histological appearance of tumours. *Tumours of mammary gland*, Forth Ed Edingburgh, Churcill Livingstone 1990;440-55.
3. A. Patrick Forrest, et al, Randomised controlled trial of conservation therapy for breast cancer: 6-year analysis of the Scottish trial 1996; 348(9029):708-13.
4. Milovic B. Prediction and decision making in Health Care using Data Mining. *Int J Publ Health Sci (IJPHS)* 2012; 1(2): 69-78.
5. Hariz M, Adnan M, Husain W, Rashid N. A. Data Mining for Medical Systems: A Review. *Int Conf Adv Comput Inform Tech - ACIT* 2012; 17-22.
6. Richie RC, Swanson JO. Breast Cancer: A Review of the Literature. *J Insur Med* 2003; 35:85 -101.
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 Tech* 2014; 7(4): 472-9.
8. Cios KJ, Moore, Uniqueness of medical data mining, *Artificial intelligence in medicine*, 2002; Vol. 26, No. 1: 1-24.
9. Sheikhpour R, et al, Breast Cancer Detection Using Two-Step Reduction of Features Extracted From Fine Needle Aspirate and Data Mining Algorithms, *Iranian Quarterly Journal of Breast Disease* 2015; 7(4).
10. Delen D, Walker G, A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods, *Artificial Intelligence in Medicine*, 34(2) , 2005, 113–127.
11. Aruna S, Rajagopalan DS, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. *Comput Sci Inform Tech* 2011; 2: 37-45.
12. Ster B, Dobnikar A. Neural networks in medical diagnosis: Comparison with other methods. *Proc Int Conf Eng Appl neural networks*; 427-30.
13. Holland J. H. *Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence*, U Michigan Press; 1975.
14. Frank, Asuncion, *UCI Machine Learning Repository*, University of California; 2010.
15. Dodge, Y *the Oxford Dictionary of Statistical Terms*, OUP; 2003.
16. Han J, Kamber M, Pei J, *Data Mining: Concepts and Techniques*, 3rd edition, Morgan Kaufmann; 2011.
17. Goldberg D, Holland J. Genetic algorithms and machine learning. *Machine learning* 1961,3(2): 95-99.
18. Talbi E G, *Metaheuristics: From Design to Implementation*, Wiley 2009.
19. Alpaydin, *Voting over Multiple Condensed Nearest Neighbors*, *Artif. Intell. Rev.*, 1997; 11: 115-132.
20. Kohavi, Ron A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (San Mateo, CA: Morgan Kaufmann)* 1995; 2(12):1137-43.
21. Jong D, *An Analysis of the Behavior of a Class of Genetic Adaptive Systems*, Ph.D. dissertation, University of Michigan, Ann Arbor, MI, 1975.
22. Rish I. An empirical study of the naive Bayes classifier, *IJCAI-01 Workshop on Empirical Methods in Artificial Intelligence*; 2001.
23. Rokach L, Maimon O. *Data Mining with Decision Trees: Theory and Applications*, World Scientific Publishing; 2008.
24. Aruna S, Rajagopalan DS, Nandakishore LV. Knowledge based analysis of various statistical tools in detecting breast cancer. *Computer Science & Information Technology* 2011; 2: 37-45.

25. Land WH, Verheggen EA. Multiclass Primal Support Vector Machines For Breast Density Classification. *Int J Comput Biol Drug Des* 2009; 2(1):21-57.
26. Kiyani T, Yildirim T. Breast cancer diagnosis using statistical neural networks. *IU-Journal of Electrical & Electronics Engineering* 2011; 4(2): 1149-53.
27. Chaurasia S, Chakrabarti P. An Approach with Support Vector Machine using Variable Features Selection on Breast Cancer Prognosis. *International Journal of Advanced Research in Artificial Intelligence* 2013; 2(9): 38-42.
28. Sarvestan Soltani A, Safavi AA, Parandeh MN, Salehi M. Predicting Breast Cancer Survivability using data mining techniques. *Software Technology and Engineering (ICSTE). 2nd International Conference* 2010; 2: 227-31.
29. Lavanya D, Rani KU. Ensemble decision tree classifier for breast cancer data. *International Journal of Information Technology Convergence and Services* 2012; 2(1): 17-24.
۳۰. طلوعی اشلقی عباس، پورابراهیمی علی، ابراهیمی ماندانا، قاسم احمد لیلا. پیش بینی عود مجدد سرطان پستان به کمک سه تکنیک داده‌کاوی. فصلنامه بیماری‌های پستان ایران، ۱۳۹۱؛ ۵(۴): ۲۳-۳۴.
۳۱. کیانی بهزاد، آتشی علیرضا. ایجاد یک مدل پیش‌آگهی مبتنی بر داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان. *مجله انفورماتیک سلامت و زیست پزشکی*. ۱۳۹۳؛ ۱(۱): ۲۶-۳۱.
۳۲. آتشی علیرضا، کیانی بهزاد. کشف الگوهای پنهان در مجموعه داده‌های واقعی بیماران مبتلا به سرطان پستان با استفاده از تکنیک داده‌کاوی. *مجله انفورماتیک سلامت و زیست پزشکی*، ۱۳۹۴؛ ۸(۱)، ۶۵-۶۰.
33. Vikas C, Saurabh P. Data Mining Techniques: To Predict and Resolve Breast Cancer Survivability. *IJCSMC* 2014; 3(1): 10-22.
34. Emina A, Abdulhamit S. Comparison Of Decision Tree Methods For Breast Cancer Diagnosis. *ICIT 2013 The 6th International Conference on Information Technology*, 8(1), 2013:12-130.
35. Lavanya D and Usha R. Ensemble Decision Tree Classifier For Breast Cancer Data. *International Journal of Information Technology Convergence and Services (IJITCS)* 2012;2(1):268-74.
36. GI Salama, M Abdelhalim. Breast cancer diagnosis on three different datasets using multi-classifiers, *Breast Cancer (WDBC)* 2012; 190-6.