مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره ششم شماره چهارم. ص ۳۰۹-۳۱۹، ۱۳۹۸

ارائه یک الگوریتم برای پیش‌بینی عود بیماران مبتلا به سرطان پستان با استفاده از الگوریتم ژنتیک و الگوریتم نزدیک‌ترین همسایگی

ستایش صادقی[۱]، امین گلاب‌پور[۲]*

• **دریافت مقاله:** ۱۳۹۷/۱۰/۱ • **پذیرش مقاله:** ۱۳۹۸/۳/۲۷

مقدمه: بیماری سرطان پستان یکی از شایع‌ترین انواع سرطان و شایع‌ترین نوع بدخیمی در زنان است که در سال‌های اخیر روند رو به رشدی داشته است. در مبتلایان به این بیماری همواره احتمال عود مجدد وجود دارد. عوامل زیادی میزان این احتمال را کاهش یا افزایش می‌دهند. داده‌کاوی از روش‌هایی است که در تشخیص یا پیش‌بینی سرطان‌ها به کار می‌رود و یکی از بیشترین کاربردهای آن، پیش‌بینی عود مجدد سرطان پستان است.

روش: در این مطالعه گذشته‌نگر از داده‌های ۶۹۹ بیمار مبتلا به سرطان پستان با ۱۴ ویژگی استفاده شد که از این تعداد ۴۵۸ نفر (۶۶ درصد) سرطان آن‌ها عود نکرد و ۲۴۱ نفر (۳۴ درصد) سرطان آن‌ها عود کرده است. این اطلاعات از سال ۱۳۹۱ تا ۱۳۹۴ از پرونده بیماران سرطان پستان جهاد دانشگاهی جمع‌آوری شد. در این پژوهش از ترکیب دو الگوریتم نزدیک‌ترین همسایگی و الگوریتم ژنتیک برای پیش‌بینی عود بیماران مبتلا به سرطان پستان استفاده گردید. ابتدا الگوریتم نزدیک‌ترین همسایگی برای پیش‌بینی عود سرطان پستان ارائه شد سپس به کمک الگوریتم ژنتیک متغیرهای وابسته کاهش یافت تا مدل صحت مناسب‌تری داشته باشد.

نتایج: تعداد متغیرهای وابسته ۱۴ متغیر بود که به کمک الگوریتم ژنتیک به ۶ متغیر کاهش پیدا نمود تا مدل پیش‌بینی کارایی بهتری داشته باشد. جهت ارزیابی مدل از پارامتر صحت استفاده شد که مقدار آن برای مدل پیشنهادی ۷۷/۱۴ درصد است که نسبت به روش‌های دیگر خروجی مناسب‌تری دارد.

نتیجه‌گیری: در این مطالعه الگوریتم پیشنهادی با روش‌های دیگر پیش‌بینی مورد بررسی قرار گرفت و مشخص گردید الگوریتم پیشنهادی دارای صحت بهتر است.

کلید واژه‌ها: عود سرطان پستان، الگوریتم ژنتیک، الگوریتم نزدیک‌ترین همسایگی

• **ارجاع:** صادقی ستایش، گلاب‌پور امین. ارائه یک الگوریتم برای پیش‌بینی عود بیماران مبتلا به سرطان پستان با استفاده از الگوریتم ژنتیک و الگوریتم نزدیک‌ترین همسایگی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۸؛ ۶(۴): ۳۰۹-۱۹.

۱. کارشناسی ارشد، مهندسی کامپیوتر، گروه مهندسی کامپیوتر، دانشگاه آزاد اسلامی واحد کرمان، کرمان، ایران

۲. دکتری تخصصی انفورماتیک پزشکی، استادیار، دانشگاه علوم پزشکی شاهرود، دانشکده پیراپزشکی، شاهرود، ایران

* **نویسنده مسئول:** شاهرود، میدان هفتم تیر، دانشگاه علوم پزشکی شاهرود، گروه فناوری اطلاعات سلامت

• **شماره تماس:** ۳۲۳۹۵۰۵۴ - ۰۲۳ • **Email:** a.golabpour@shmu.ac.ir

مقدمه

سرطان پستان شایع‌ترین سرطان زنان در آمریکا و دومین علت مرگ در اثر سرطان در این کشور است. بر اساس آمار سازمان بهداشت جهانی در سال ۲۰۱۸ سرطان پستان پنجمین علت مرگ ناشی از سرطان در جهان با حدود ۵۰۸,۰۰۰ مورد مرگ از بین حدود ۵۴ میلیون مرگ در جهان می‌باشد. این سرطان در بین زنان، اولین علت مرگ ناشی از سرطان را به خود اختصاص می‌دهد [۱].

بر اساس برآورد انجمن سرطان آمریکا در سال ۲۰۱۸ تعداد موارد جدید ابتلا به سرطان پستان در بین زنان در این کشور ۲۳۲,۳۴۰ نفر خواهد بود و تعداد مرگ در اثر سرطان پستان در زنان این کشور را ۳۹,۶۲۰ مورد برآورد کرده است؛ به طور کلی یک نفر از هر هشت زن در طول عمر خود به سرطان پستان مبتلا می‌شود و از هر ۳۶ نفر یک نفر به علت سرطان پستان می‌میرد [۲،۳].

سرطان پستان همچنین یکی از شایع‌ترین سرطان‌ها در زنان ایرانی است، در عین حال وجوه اپیدمیولوژیک سرطان در بین بیماران ایرانی نامشخص است. نتایج یک مرور نظام‌مند نشان داده است که این سرطان از سن ۱۵ تا ۸۴ سال مشاهده شده است؛ اما در سنین ۴۰ تا ۴۹ ساله در بین زنان نسبت به سنین دیگر شایع‌تر است. میزان بروز سرطان پستان در صد هزار نفر جمعیت حدود ۲۲ نفر و میزان شیوع آن نیز ۱۲۰ در صد هزار نفر جمعیت است؛ هر چند جمعیت ایران جوان است و عمده جمعیت یعنی ۸۶/۱ درصد کمتر از ۵۰ سال دارند، حتی پس از تحلیل با در نظر گرفتن سن تعداد بیماران مشاهده شده در سنین ۴۰ تا ۴۹ سال به طور معنی‌داری بالاتر از انتظار است [۴،۵]. تشخیص عود سرطان پستان نقش مؤثری در زمان شروع درمان و استفاده از داروهای مناسب برای افزایش بقای بیمار دارد؛ لذا این ضرورت احساس می‌شود که دانستن عود یا عدم عود سرطان پستان بیماران کمک بسیار مناسب به پزشکان برای تصمیم‌گیری می‌نماید.

در زمینه پیش‌بینی عود سرطان پستان مطالعات زیادی در داخل و خارج از کشور انجام شده است. در مطالعه‌ای Ojha و Goel برای پیش‌بینی عود سرطان پستان ارائه گردید [۶]. این پژوهش بر پایه روش‌های داده‌کاوی بوده و داده‌ها بر اساس پایگاه (https://archive.ics.uci.edu/ml/index.php) UCI مورد مطالعه قرار گرفت؛ همچنین برای پیش‌بینی عود سرطان پستان از ترکیب دو الگوریتم درخت تصمیم و ماشین بردار پشتیبان استفاده شده است. نتایج استخراج شده از این مطالعه تعیین متغیرهای تأثیرگذار برای عود سرطان پستان و ارائه یک مدل پیش‌بینی است و به ۱۱ پارامتر تأثیرگذار برای پیش‌بینی عود سرطان پستان رسیدند و همچنین مدل پیش‌بینی دارای صحت ۸۱ درصد است. در پژوهش دیگری Silva و همکاران برای پیش‌بینی عود سرطان پستان ارائه نمودند [۷]. در این پژوهش از داده‌های استاندارد دانشگاه کالیفرنیا استفاده گردید. این داده‌ها در سال ۱۹۸۷ جمع‌آوری گردید و شامل ۲۷۲ نمونه بوده و همچنین دارای ۹ متغیر مستقل و یک متغیر وابسته دودویی بوده است؛ در این مدل الگوریتم پیش‌بینی به دقت ۹۵/۸ درصد رسیده است. در پژوهش دیگری کیانی و آتشی در جهاد دانشگاهی تهران انجام دادند [۵]. در این پژوهش از الگوریتم درخت تصمیم جهت پیش‌بینی عود سرطان پستان استفاده کردند و الگوریتم به صحت ۷۵ درصد رسید و نشان داده شد که این الگوریتم نسبت به الگوریتم‌های دیگر بهتر عمل کرده است. در پژوهش قربانی و همکاران از اطلاعات ۳۷۷ بیمار مبتلا به سرطان پستان استفاده شد از این تعداد بیماران ۳۴ درصد با عود سرطان پستان مواجه شده بودند. در این پژوهش نشان داده شد که الگوریتم رگرسیون لجستیک یک روش مناسب برای پیش‌بینی عود سرطان پستان است [۸].

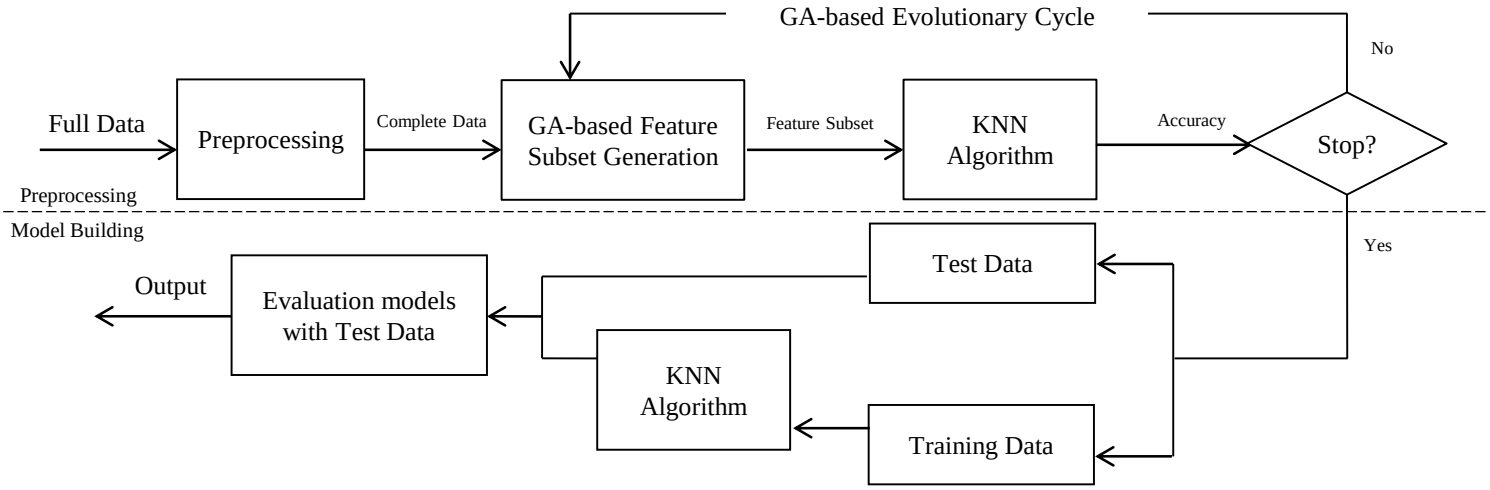
سرطان پستان نوعی بیماری وابسته به هورمون هست که در نتیجه تغییر و رشد غیرقابل‌کنترل سلول‌های پستان بروز می‌نماید. نسبت ابتلا به این بیماری در زنان ۱۵۰ برابر مردان می‌باشد در سال ۲۰۱۷ حدود ۱۸۰ هزار مورد از سرطان مهاجم پستان و ۴۰ هزار مورد مرگ ناشی از آن در ایالات‌متحده روی داد. علاوه بر این، نیز حدود دو هزار نفر از مردان به سرطان پستان مبتلا بودند. بدخیمی‌هایی با منشأ اپیتلیال پستان، شایع‌ترین علت سرطان در زنان (بعد از سرطان پوست) است که حدود یک سوم همه سرطان‌های زنان را تشکیل می‌دهد. با بهبود وضعیت درمان و تشخیص زودرس، مرگ‌ومیر ناشی از سرطان پستان به‌طور قابل‌توجهی در حال کاهش است [۹].

سه تاریخ مهم زندگی هر زنی که اثر عمده‌ای بر میزان بروز سرطان پستان دارد، عبارت‌اند از: سن اولین قاعدگی، سن اولین حاملگی و سن یائسگی. به‌عنوان مثال، خطر ابتلا به سرطان پستان در زنانی که اولین بار در ۱۶ سالگی قاعده شده‌اند؛ فقط ۵۰ تا ۶۰ درصد زنانی است که اولین بار در ۱۲ سالگی قاعده شده‌اند و این کاهش خطر ابتلا به سرطان پستان، در طول زندگی ادامه خواهد یافت. به همین ترتیب، اگر یائسگی ۱۰ سال قبل از میانه سن یائسگی ۵۲ (سالگی) روی دهد، خطر

ابتلا به سرطان پستان را در طول زندگی ۳۵ درصد کمتر می‌کند. یکی از اجزای مهم تعیین کننده خطر کلی ابتلا به سرطان، طول مدتی است که بیمار قاعده می‌شود. این سه فاکتور (اولین قاعدگی، سن اولین حاملگی و یائسگی) قادر است ۷۰ تا ۸۰ درصد از تغییراتی را که در فراوانی سرطان پستان در کشورهای مختلف وجود دارد، توجیه کند. طبق بررسی‌ها، مدت‌زمان شیردهی نیز ریسک ابتلا به این سرطان را به مقدار قابل‌توجهی کاهش می‌دهد [۱۰،۱۱].

داده‌های این پژوهش از نوع مطالعات پیش‌بینی عود بیماران سرطان پستان است، داده‌ها از پرونده تمام ۶۹۹ بیماران سرطان پستان که در سال‌های ۱۳۹۱ تا ۱۳۹۴ برای درمان به مرکز تحقیقات سرطان پستان جهاد دانشگاهی، ایران مراجعه کرده بودند استخراج شد. اطلاعات شامل ۱۴ متغیر مستقل و یک متغیر وابسته است. این متغیرها شامل ۲ متغیر اسمی، ۱۰ متغیر رتبه‌ای و ۳ متغیر فاصله‌ای هستند.

در این قسمت به کمک الگوریتم ژنتیک ویژگی‌های داده کاهش یافت، بعد از کاهش ابعاد، داده‌ها به دو قسمت آموزشی و آزمون تقسیم شدند، سپس مانند مرحله قبل بر روی داده‌های آموزشی الگوریتم نزدیک‌ترین همسایگی اجرا شد و مدل توسط داده‌های آزمون ارزیابی گردید. در شکل ۱ نحوه کاهش ابعاد توسط الگوریتم ژنتیک و اعمال الگوریتم نزدیک‌ترین همسایگی بیان شد.



شکل ۱: طراحی مدل پیشنهادی به کمک الگوریتم ژنتیک و الگوریتم نزدیک‌ترین همسایگی

همان‌گونه که در شکل ۱ مشاهده شد، ابتدا بر روی داده‌ها پیش‌پردازش انجام گرفت سپس تمام ۶۹۹ رکورد وارد الگوریتم ژنتیک شد در ادامه الگوریتم ژنتیک زیر مجموعه‌ای از رکوردها را انتخاب نمود، این زیر مجموعه توسط الگوریتم نزدیک‌ترین همسایگی ارزیابی شد این فرآیند آن‌قدر تکرار شد که بهترین زیر مجموعه به عنوان خروجی و پارامترهای تأثیرگذار بر روی عود سرطان پستان انتخاب شود. سپس داده‌ها به دو قسمت آموزشی و آزمون تقسیم شد در ادامه الگوریتم نزدیک‌ترین همسایگی بر روی داده‌های آموزشی اجرا و یک مدل طراحی گردید. مدل طراحی شده توسط داده‌های آزمون ارزیابی شد.

ساختار استفاده از الگوریتم ژنتیک جهت کاهش ابعاد مسئله

۱. کاهش ابعاد با الگوریتم ژنتیک

برای بهینه‌سازی متغیرهای مستقل، از الگوریتم ژنتیک دودویی استفاده شد. کروموزوم به طول متغیرهای مستقل تعریف شد، اگر محتوای خانه iام کروموزوم یک باشد بدان معنی است که متغیر مستقل iام در محاسبات در نظر گرفته شود و اگر محتوای خانه iام کروموزوم صفر باشد به این معنی است که متغیر مستقل iام در محاسبات در نظر گرفته نشود. الگوریتم ژنتیک حالت‌های متفاوت از وجود یا عدم وجود متغیر وابسته را تولید نمود، سپس زیر مجموعه‌ای از متغیرهای مستقل که دارای بهترین مدل پیش‌بینی برای الگوریتم نزدیک‌ترین همسایگی هستند، انتخاب شد (جدول ۱).

جدول ۱: ساختار کروموزوم پیشنهادی

	۱	۲	۳	…	n	Fitness
کروموزوم	[۰,۱]	[۰,۱]	[۰,۱]	…	[۰,۱]	

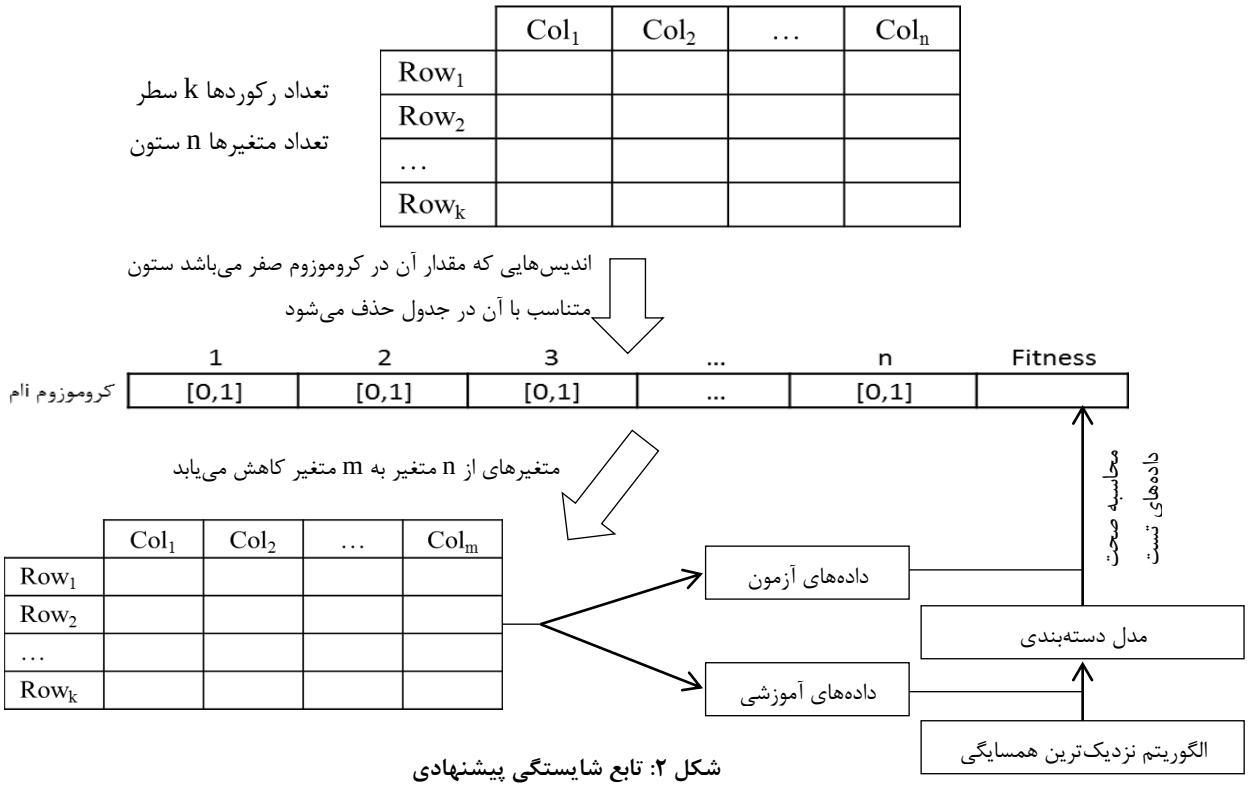
همان‌طور که در جدول ۱ مشاهده شد n تعداد متغیرهای وابسته و هر ژن کروموزوم دارای مقدار ۰ یا ۱ است.

۲. تولید جمعیت اولیه

بعد از این مرحله اندازه جمعیت باید تعیین گردد، یعنی چه تعداد کروموزوم برای اجرا الگوریتم ژنتیک باید موجود باشد. اندازه جمعیت عامل مهمی در کارایی الگوریتم به حساب می‌آید. اگر اندازه جمعیت کوچک باشد، بخش کوچکی از فضای جواب جستجو خواهد شد و جواب به سرعت و با احتمال زیاد به یک بهینه محلی همگرا می‌شود و اگر اندازه جمعیت بزرگ باشد محاسبات بسیار زیادی انجام خواهد شد که نسبت به جواب به دست آمده نامتناسب است، در نتیجه زمان اجرای بسیار طولانی خواهد داشت. طی ارزیابی‌های انجام شده تعداد جمعیت اولیه برای مسئله عود سرطان پستان ۱۰ در نظر گرفته شده است که این مقدار با سعی و خطا به دست آمد.

۳. تابع شایستگی

تابع شایستگی، برازندگی و عملکرد هر عضو جمعیت را در حل مسئله ارزیابی می‌کند و برای حل مسئله عود سرطان پستان نیز باید تابعی برای اندازه‌گیری مناسب بودن مدل پیشگویی جهت برازش جواب‌های به دست آمده، تعریف شود. برای اندازه‌گیری میزان شایستگی هر کروموزوم، زیر مجموعه‌ای از متغیرهای مستقل انتخاب شد سپس بر روی آن الگوریتم نزدیک‌ترین همسایگی اعمال و میزان دقت الگوریتم به عنوان تابع شایستگی در نظر گرفته شد (شکل ۲).



شکل ۲: تابع شایستگی پیشنهادی

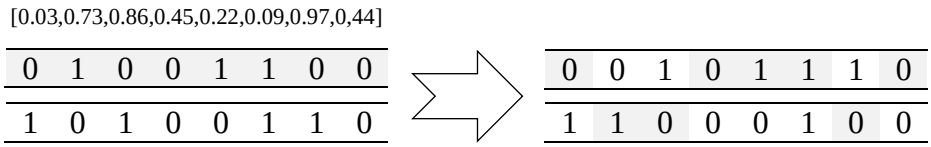
٤. انتخاب

برای انتخاب والدین از روش انتخاب رتبه‌ای استفاده شد. دلیل انتخاب این است که در ابتدای الگوریتم از همگرایی زودرس الگوریتم ژنتیک جلوگیری شود و همچنین قابلیت واگرایی در الگوریتم ایجاد گردد [۱۲]. سپس بعد از طی ۵۰ نسل (تعداد نسل‌ها توسط آزمایش و خطا به دست آمده است) می‌بایست الگوریتم به همگرایی برسد به همین دلیل الگوریتم انتخاب را به الگوریتم انتخاب چرخ رولت تغییر می‌یابد.

٥. ترکیب

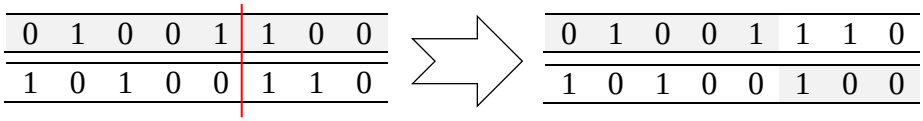
ساختار الگوریتم ژنتیک به این صورت بود که ابتدا باید سعی در واگرایی داشته باشد سپس الگوریتم شروع به همگرایی نماید؛ در قسمت اول ترکیب در حالت انتخاب رتبه‌ای از روش ترکیب یکنواخت استفاده شد (شکل ۳).

در عملگر ترکیب یکنواخت، مقدار ژن فرزند با توجه به مقادیر ژن‌های متناظر هر دو والد انتخاب شد. در این روش، مقادیر ژن‌های هر کدام از والدین، شانسی برابر برای حضور در ژن متناظر فرزند دارند. در عملگر باز ترکیب یکنواخت، بر اساس یک توزیع تصادفی دودویی مشخص شد که مقدار هر ژن فرزند از مقدار ژن متناظر کدام والد انتخاب گردد.



شکل ۳: ساختار ترکیب یکنواخت [۱۳]

در قسمت دوم ترکیب در حالت انتخاب چرخ رولت از ترکیب تک نقطه‌ای استفاده گردید (شکل ۴).

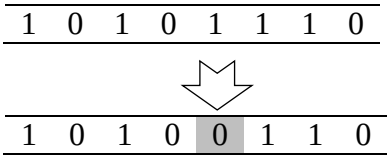


شکل ٤: ساختار ترکیب یک نقطه‌ای [۱۳]

٦. جهش

در عملگر باز ترکیب تک نقطه‌ای، ابتدا یک نقطه تصادفی در دنباله کروموزوم‌های والدین انتخاب شد و سپس از محل انتخاب شده، کروموزوم هر دو والد برش خورد. فرزند دوم، شامل بخش اول والد دوم و بخش دوم والد اول است.

در عملگر جهش با تغییر بیت، یک ژن به طور تصادفی انتخاب شد و مقدارش معکوس شد (اگر صفر باشد، یک می‌شود و برعکس). مشخص است که این عمل به دلیل عدم بهره‌برداری از اطلاعات موجود در جمعیت، سنخیت کاملی با تعریف عملگر جهش دارد و سعی در تولید جمعیت جدید الگوریتم دارد (شکل ۵).



شکل ٥: ساختار جهش [۱۳]

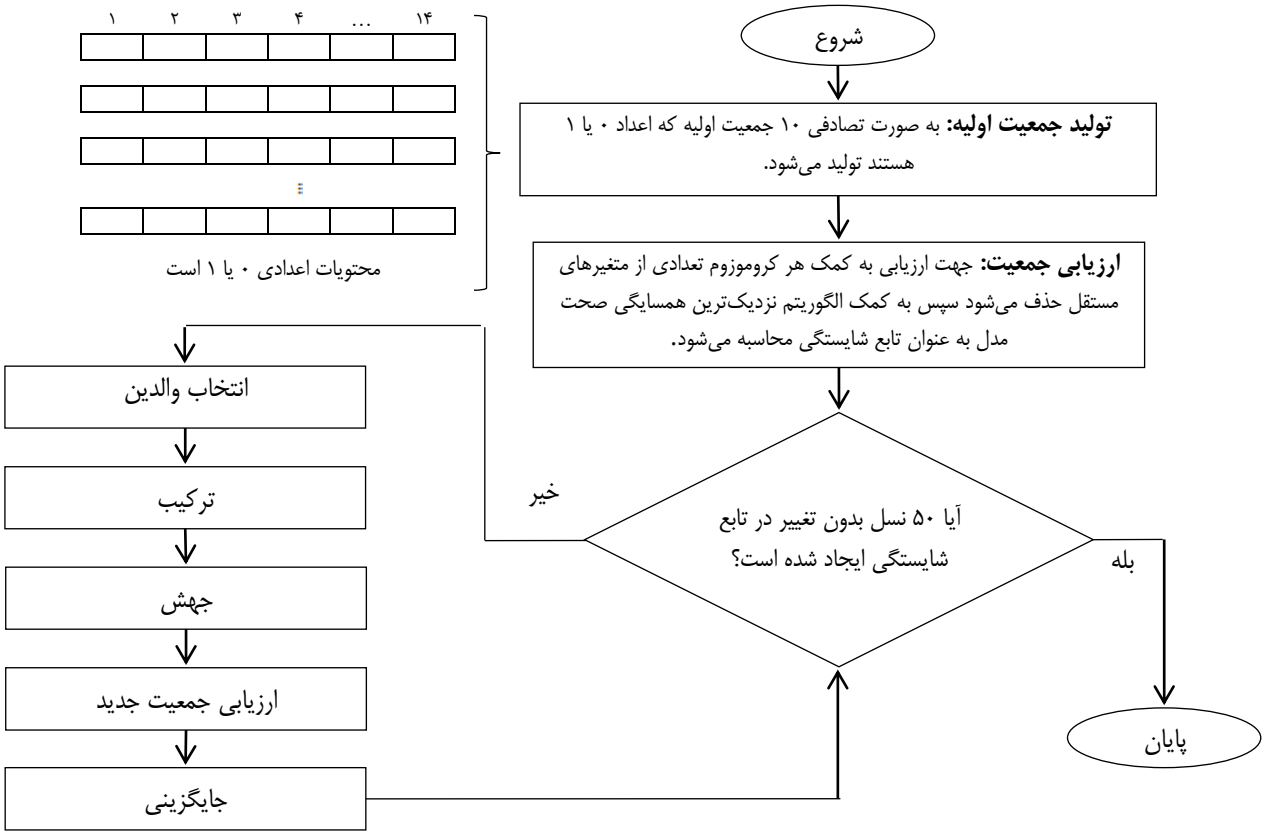
در ابتدای الگوریتم درصد جهش را بالا گرفته تا میزان واگرایی الگوریتم زیاد باشد سپس بعد از تولید چندین نسل درصد جهش الگوریتم را پایین گرفته تا به همگرایی کروموزوم‌ها برسد [۱۲].

۷. جایگزینی

از روش جایگزینی نسلی استفاده شد. در ابتدای شروع الگوریتم ۵۰٪ از والدین و ۵۰٪ از فرزندان را به نسل بعد انتقال می‌داد که باعث واگرایی در مسئله گردد [۱۴]. بعد از گذشت هر نسل درصد فرزندان کاهش و درصد والدین را افزایش داده شد تا مسئله به همگرایی ختم شود. میزان افزایش درصد والدین از طریق سعی و خطا (تجربه) محاسبه گردید.

۸. شرط به پایان رسیدن الگوریتم ژنتیک

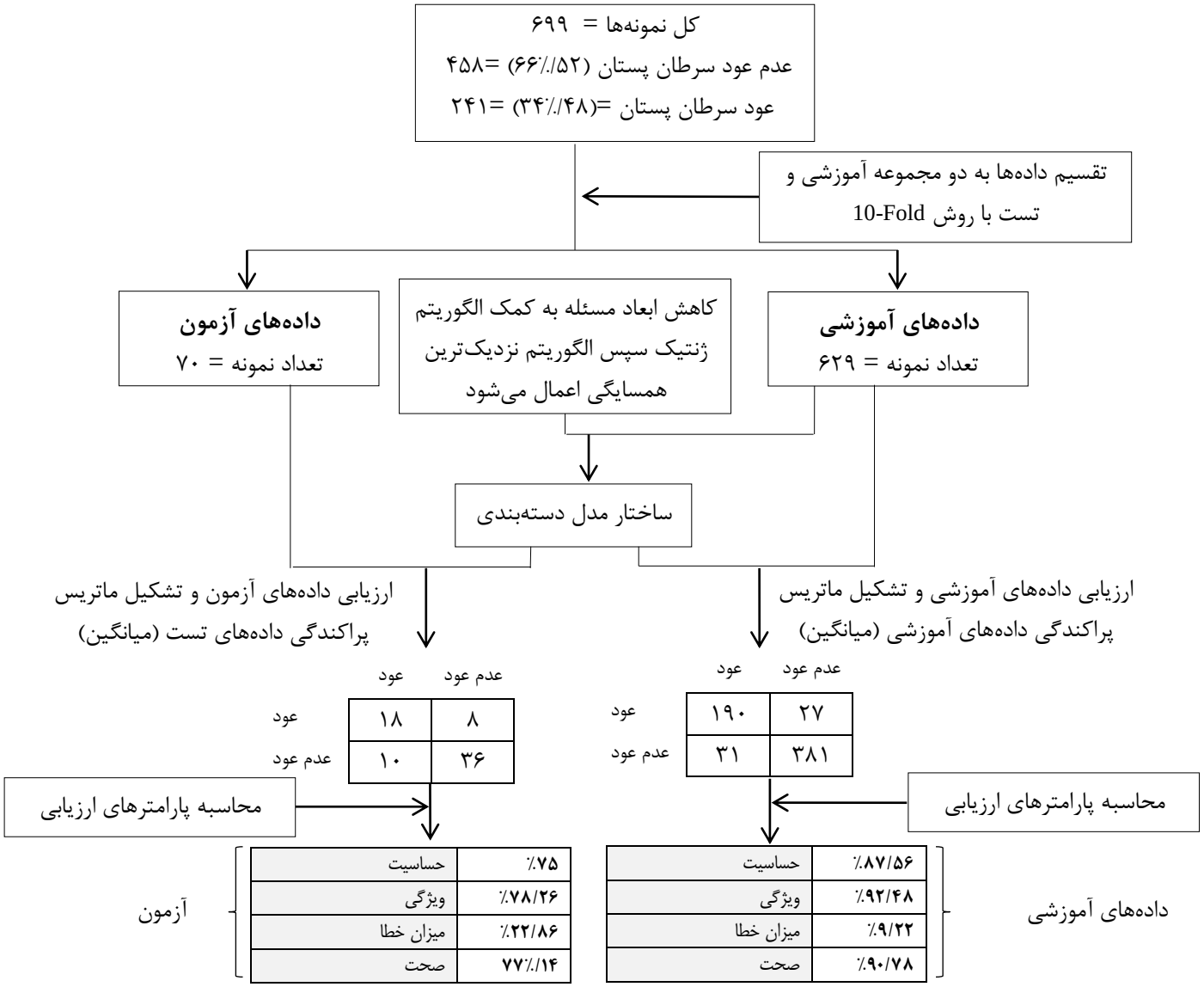
شرط پایان الگوریتم تولید ۵۰ نسل بدون تغییر تابع شایستگی است، این مقدار با سعی و خطا محاسبه شد؛ در ادامه در شکل ۶ فلوچارت کلی الگوریتم ژنتیک برای کاهش ابعاد نشان داده شد.



شکل ٦: الگوریتم ژنتیک پیشنهادی برای کاهش تعداد متغیرها مستقل

۹. ارزیابی

الگوریتم ترکیبی پیشنهادی از ترکیب الگوریتم نزدیک‌ترین همسایگی و الگوریتم ژنتیک تشکیل شده است. در ارزیابی الگوریتم پیشنهادی داده‌ها به دو مجموعه آموزشی و آزمون تقسیم شد. در این مسئله داده‌ها از روش 10-Fold استفاده می‌شود [۱۵]. الگوریتم ۱۰۰ بار اجرا گردید. در شکل ۷ خروجی اجرا الگوریتم پیشنهادی بر روی عود سرطان پستان نشان داده شد. در ارزیابی، مقادیر ماتریس پراکندگی، حساسیت، ویژگی، میزان خطا و صحت محاسبه شد.



شکل ۷: خروجی الگوریتم پیشنهادی برای داده عود سرطان پستان

ابتدا به کمک ترکیب دو الگوریتم نزدیک‌ترین همسایگی و ژنتیک متغیرهای مستقل کاهش پیدا کرد سپس با کمک داده‌های آموزشی یک مدل دسته‌بندی ارائه شد در ادامه داده‌های آموزشی و داده‌های آزمون مدل ارزیابی و ماتریس پراکندگی برای آن اجرا شد و در ادامه چهار پارامتر دقت، ویژگی، حساسیت و درصد خطا برای داده‌های آموزشی و آزمون محاسبه شد. پارامترهای داده‌های آزمون مهم‌تر است همان‌طور که مشاهده شد در داده‌های عود سرطان پستان درمانی دقت ۷۷/۱۴ درصد حاصل شد که خروجی نهایی الگوریتم نزدیک‌ترین همسایگی است (جدول ۲).

جدول ۲: خروجی الگوریتم پیشنهادی برای داده عود سرطان پستان

نام داده	داده آموزشی						داده آزمون					
	ماتریس پراکندگی		حساسیت	ویژگی	میزان خطا	صحت	ماتریس پراکندگی		حساسیت	ویژگی	میزان خطا	صحت
داده‌های سرطان پستان	۲۷	۱۹۰	٪۸۷/۵۶	٪۹۲/۴۸	٪۹/۲۲	٪۹۰/۷۸	۸	۱۸	٪۷۵	٪۷۸/۲۶	٪۲۲/۸۶	٪۷۷/۱۴
	۳۸۱	۳۱					۳۶	۱۰				

الگوریتم‌های ژنتیک برای کاهش متغیرهای مستقل استفاده گردید، با استفاده از آن متغیرهای مستقل به ۶ متغیر کاهش یافت (جدول ۳). این شش پارامتر شامل سن بیمار، میانگین شعاع توده سرطانی، سن تشخیص بیمار، وجود سابقه قبلی سرطان پستان در خانواده، وجود سابقه قبلی سرطان‌ها دیگر در خانواده و سن تشخیص سرطان در خانواده است که به عنوان پارامترهای تأثیرگذار استخراج شد.

جدول ۳: شش متغیر کاهش یافته به کمک الگوریتم ژنتیک

ردیف	نام متغیر
۱	سن بیمار
۲	میانگین شعاع توده سرطانی
۳	سن تشخیص بیمار
۴	وجود سابقه قبلی سرطان پستان در خانواده
۵	وجود سابقه قبلی سرطان‌ها دیگر در خانواده
۶	سن تشخیص سرطان در خانواده

مقایسه الگوریتم پیشنهادی با الگوریتم‌های دیگر

کارایی الگوریتم پیشنهادی با روش‌های درخت تصمیم C4.5، بیزین، ماشین بردار پشتیبان، رگرسیون لجستیک و شبکه عصبی چندلایه مورد مقایسه قرار گرفت. جهت اجرا این پنج الگوریتم از نرم‌افزار Weka V3.9 استفاده گردید (جدول ۴).

جدول ٤: مقایسه پنج الگوریتم با الگوریتم پیشنهادی

نام الگوریتم	حساسیت	ویژگی	میزان خطا	صحت
درخت تصمیم	٪۷۴/۶۵	٪۷۰/۴۳	٪۲۴/۰۶	٪۷۵/۹۴
بیزین	٪۷۱/۷۸	٪۷۰/۴۴	٪۲۶/۵۸	٪۷۳/۴۲
ماشین بردار پشتیبان	٪۷۷/۵۲	۷۰/۰۸	٪۲۳/۹۹	٪۷۶/۰۱
رگرسیون لجستیک	٪۷۴/۸۴	٪۷۰/۴۱	٪۲۴/۵۲	٪۷۵/۴۸
شبکه عصبی چندلایه	٪۷۵/۶۸	٪۷۰/۳۹	٪۲۴/۴	٪۷۵/۶
الگوریتم پیشنهادی	**٪۷۵**	**٪۷۸/۲٦**	**٪۲۲/۸٦**	**۷۷٪/۱٤**

بحث و نتیجه‌گیری

با توجه به جدول ۴ الگوریتم پیشنهادی بهترین خروجی را برای مسئله پیش عود سرطان پستان درمانی اخذ نمود. الگوریتم پیشنهادی در ویژگی، میزان خطا و صحت خروجی بهتری دارد در محاسبه حساسیت الگوریتم ماشین بردار پشتیبان بهتر عمل کرده است در ویژگی الگوریتم پیشنهادی با مقدار ۷۸/۲۶٪ دارای بیشتر ویژگی است و بعد از آن الگوریتم بیزین با مقدار ویژگی ۷۰/۰۸٪ بهتر از چهار الگوریتم دیگر عمل کرده است و محاسبه صحت الگوریتم پیشنهادی با مقدار ۷۷٪/۱۴ دارای بیشترین صحت است و بعد از آن الگوریتم ماشین بردار پشتیبان با صحت ۷۶/۰۱٪ بهتر عمل کرده است؛ لذا الگوریتم پیشنهادی ۱/۱۳٪ از الگوریتم ماشین بردار پشتیبان بهتر عمل کرده است که نشان دهنده مناسب بودن روش پیشنهادی است.

هدف از این پژوهش، ارائه روشی برای دسته‌بندی مجموعه داده‌های عود بیماران سرطان پستان با استفاده از ترکیب دو الگوریتم ژنتیک و الگوریتم نزدیک‌ترین همسایگی است؛ به طوری که بتوان از طریق این الگوریتم پیشنهادی، بیماری سرطان پستان با کارایی بالا تشخیص داده شود. الگوریتم‌های مختلفی جهت دسته‌بندی مجموعه داده‌ها موجود هستند، که از این میان، در این تحقیق از الگوریتم‌های ژنتیک پیشنهادی و الگوریتم نزدیک‌ترین همسایگی استفاده شد. باید خاطر نشان کرد که طی تحقیقات انجام شده در گذشته، الگوریتم‌های منفرد در زمینه تشخیص عود بیماری سرطان پستان تا حد زیادی نمی‌توانند کارایی دسته‌بندی را افزایش دهند [۷].

تمرکز این پژوهش بر معیار صحت است این معیار از جمله معیارهای مهم در ارزیابی کارایی سیستم‌های تشخیصی در

حوزه پزشکی است [۱۶]. در این تحقیق با ثابت فرض کردن مجموعه داده و الگوریتم یادگیری، هدف بررسی تأثیر ویژگی‌های استخراج شده از داده‌ها پستان بر کارایی تشخیص عود پستان است. جهت انجام این کار از الگوریتم ژنتیک استفاده گردید و متغیرهای تأثیرگذار برای عود بیماری استخراج گردید در پژوهش [۱۷] مشخص شده بود که الگوریتم ژنتیک روشی مناسب برای کاهش ابعاد است.

در این پژوهش مشخص شد شش پارامتر سن بیمار، میانگین شعاع توده سرطانی، سن تشخیص بیمار، وجود سابقه قبلی سرطان پستان در خانواده، وجود سابقه قبلی سرطان‌ها دیگر در خانواده و سن تشخیص سرطان در خانواده پارامترهایی برای تعیین عود سرطان پستان نقش تعیین کنند دارد همچنین مشخص گردید در مطالعه [۱۸] نیز تأکید شده بود سن یکی از عامل اصلی عود مجدد سرطان پستان است. در مطالعه [۵] هم مشخص شده بود سابقه قبلی یکی از عوامل تأثیرگذار در عود مجدد سرطان پستان است که در این پژوهش هم مشخص گردید.

در این پژوهش مشخص گردید الگوریتم پیشنهادی دارای صحت بیشتر نسبت به درخت تصمیم، بیزین، ماشین بردار پشتیبان، شبکه عصبی مصنوعی و رگرسیون لجستیک است و میزان صحت بین یک تا چهار درصد افزایش یافته است.

الگوریتم پیشنهادی می‌توان با احتمال ۷۷ درصد به پزشک پیشنهاد بدهد که سرطان این بیمار عود می‌کند یا خیر، نه نسبت از روش‌های دیگر روش قابل قبولی است. همچنین مدل پیشنهادی دارای حساسیت ۷۵ درصد است از این روی می‌توان انتظار داشت که نزدیک ۷۵ درصد از عود مجدد سرطان پستان بیماران را کشف نماید.

با توجه به بررسی‌های انجام شده، یکی از راه‌های افزایش کارایی دسته‌بندی، استفاده از دسته‌بندهای ترکیبی است. استفاده از دسته‌بند گروهی باعث افزایش کارایی دسته‌بندی توسط ترکیب چند دسته‌بند منفرد می‌شود. در روش پیشنهادی استفاده از داده‌های تعیین عود سرطان پستان مورد ارزیابی قرار گرفت. در این تحقیق، علاوه بر روش پیشنهادی، الگوریتم‌های دسته‌بندی به کار برده شده در روش پیشنهادی، به صورت منفرد و در قالب یک سیستم ناهمگن جهت مقایسه با روش پیشنهادی از لحاظ کارایی نیز، پیاده‌سازی گردیدند. نتایج به دست آمده نشان‌دهنده کارایی مناسب روش پیشنهادی است.

از محدودیت‌های این مطالعه کند بودن الگوریتم پیشنهادی است نیست به الگوریتم‌های مشابه است و همچنین پایین بودن حساسیت الگوریتم پیشنهادی نسبت به سایر الگوریتم‌های دیگر است.

تشکر و قدردانی

این مقاله مستخرج از پایان‌نامه کارشناسی ارشد نویسنده اول با کد ۱۰۸۴۱۰۰۶۹۴۲۰۷۵ می‌باشد.

تعارض منافع

در این مطالعه هیچ‌گونه تضاد منافعی وجود نداشت.

References

1. Cameron D, Piccart-Gebhart MJ, Gelber RD, Procter M, Goldhirsch A, de Azambuja E, et al. 11 years' follow-up of trastuzumab after adjuvant chemotherapy in HER2-positive early breast cancer: final analysis of the HERceptin Adjuvant (HERA) trial. Lancet 2017;389(10075):1195-205. doi: 10.1016/S0140-6736(16)32616-2.

2. Denkert C, Liedtke C, Tutt A, von Minckwitz GJTL. Molecular alterations in triple-negative breast cancer-the road to new treatment strategies. Lancet 2017;389(10087):2430-42. doi: 10.1016/S0140-6736(16)32454-0.

3. DeSantis CE, Ma J, Goding Sauer A, Newman LA, Jemal A. Breast cancer statistics, 2017, racial disparity in mortality by state. CA Cancer J Clin 2017;67(6):439-48. doi: 10.3322/caac.21412.

4. Dehghan P, Mogharabi M, Zabbah I, Layeghi K, Maroosi A. Modeling Breast Cancer Using Data Mining Methods. Journal of Health and Biomedical Informatics 2018; 4(4):266-78. [In Persian]

5. Kiani B, Atashi A. A Prognostic Model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics 2014; 1(1):26-31.[In Persian]

6. Ojha U, Goel S. A study on prediction of breast cancer recurrence using data mining techniques. 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence; 2017 Jan 12-13; Noida, India: IEEE; 2017. doi: 10.1109/CONFLUENCE.2017.7943207

7. Silva J, Lezama OBP, Varela N, Borrero LA. Integration of Data Mining Classification Techniques and Ensemble Learning for Predicting the Type of

Breast Cancer Recurrence. International Conference on Green, Pervasive, and Cloud Computing; 2019; Cham: Springer International Publishing; 2019. p. 18-30.
8. Ghorbani N, Yazdani Cherati J, Anvari K, Ghorbani N. Factors Affecting Recurrence in Breast Cancer Using Cox Model. J Mazandaran Univ Med Sci 2015; 25(131):32-9.[In Persian]
9. Guyton AC, Hall JE. Textbook of Medical Physiology. 13th ed. Philadelphia: Saunders; 2015.
10. Wikipedia. Breast cancer: 2018 [cited 2018 Dec 18]. Available from: https://en.wikipedia.org/w/index.php?title=Breast_cancer&oldid=874275465.
11. Winchester DJ, Winchester DP, Hudis CA, Norton L. Breast Cancer. 2nd ed. Ontario, Canada: B.C. Decker Inc; 2006.
12. Sivanandam SN, Deepa SN. Introduction to Genetic Algorithms: Springer Berlin Heidelberg; 2007.
13. Man KF, Tang KS, Kwong S. Genetic Algorithms: Concepts and Designs. London: Springer; 2012.
14. Beasley D, Bull DR, Martin RR. An overview of genetic algorithms: Part 1, fundamentals. University Computing 1993;15(2):1-16.
15. Witten IH, Frank E, Hall MA. Data Mining: Practical Machine Learning Tools and Techniques. 4th ed. USA: Morgan Kaufmann; 2011.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2020; 6(4): 309-319

An Algorithm for Predicting Recurrence of Breast Cancer Using Genetic Algorithm and Nearest Neighbor Algorithm

Sadeghi Setayesh[1], Golabpour Amin[2]*

• **Received:** 22 Dec, 2018 • **Accepted:** 17 Jun, 2019

Introduction: breast cancer is the most prevalent malignancy, and one of the most common types of cancer among women around the world that has showed a growing trend in recent years. There is always a probability of recurrence in patients who suffer from this disease. There are many factors that increase or decrease this probability. Data mining is one of the methods that can be used to predict or diagnose cancers. Recurrence detection of breast cancer is one of the most common applications of data mining.

Method: In this retrospective study, data of 699 breast cancer patients with 14 characteristics were collected from patients' records of Jahad Daneshgahi University from 2012 to 2015 and used. From all, 458 patients (66%) did not have recurrence, while in 241 patients (34 %) recurrence was observed. In this study, through combining k nearest neighbor (KNN) and genetic algorithm (GA), a hybrid approach was proposed to predict recurrence of breast cancer. First, KNN was applied to predict recurrence of breast cancer and then, GA was used to reduce unnecessary independent variables to provide a more accurate model of accuracy.

Results: The number of independent variables was 14 variables, which was reduced to 6 variables by genetic algorithm to make the prediction model more efficient. We used accuracy as the criterion to evaluate performance of the model, and it was obtained 77.14% which is higher than the accuracy of alternative methods.

Conclusion: In comparison to other alternative methods, the proposed method is more accurate.

Keywords: Breast Cancer Recurrence, Genetic Algorithm, Nearest Neighbor Algorithm

•**Citation:** Sadeghi S, Golabpour A. An Algorithm for Predicting Recurrence of Breast Cancer Using Genetic Algorithm and Nearest Neighbor Algorithm. Journal of Health and Biomedical Informatics 2020; 6(4): 309-19. [In Persian]

1. M.Sc. in Computer Engineering, Computer Engineering Dept., Islamic Azad University, Kerman, Iran
2. Ph.D. in Medical Informatics, Assistant Professor, Shahroud University of Medical Sciences, School of Paramedical, Shahroud, Iran

***Correspondence:** Health Information Technology Dept., Shahroud University of Medical Sciences, Hafte Tir Square, Shahroud, Iran
• **Tel:** 023-32395054 • **Email:** a.golabpour@shmu.ac.ir