

## مقایسه‌ی عملکرد الگوریتم‌های نزدیکترین همسایه و ماشین بردار پشتیبان برای پیش‌بینی سرطان سینه در زنان فعال

لیلا فصیحی<sup>۱</sup>، حمید آقاعلی‌نژاد<sup>۲</sup>، رضا قراخانو<sup>۳</sup>، فرانسیسکو خوزه آمارو گهیتی<sup>۳</sup>

### مقاله پژوهشی

### چکیده

**مقدمه:** سرطان سینه، جزء سرطان‌های شایع زنان است که تشخیص بموقع آن در ادامه حیات و درمان نقش مهمی دارد. یادگیری ماشینی پتانسیل پیش‌بینی سرطان سینه را بر اساس ویژگی‌های پنهان در داده‌ها دارد. هدف اصلی این مطالعه، پیش‌بینی سرطان سینه با استفاده از الگوریتم‌های نزدیکترین همسایه و ماشین بردار پشتیبان در زنان فعال بود.

**روش‌ها:** در این مطالعه‌ی توسعه‌ای- کاربردی جمع‌آوری داده‌ها مربوط به تعداد ۶۴۱ پرونده متعلق به بیماران مبتلا به سرطان سینه از بیمارستان‌های امام خمینی و پژوهشکده سرطان معتمد، طی سال‌های ۱۴۰۳-۱۳۹۳ در محدوده‌ی سنی ۲۵ تا ۷۵ سال بود، پس از پیش پردازش اولیه در مجموعه داده‌ها، الگوریتم‌های نزدیکترین همسایه و ماشین بردار پشتیبان به کار گرفته شدند.

**یافته‌ها:** نتایج نشان داد که الگوریتم ماشین بردار پشتیبان با دقت ۸۷/۲ درصد و صحت ۸۶/۳۵، حساسیت ۸۶/۸۸ درصد، تشخیص‌پذیری ۸۵/۶۸، عملکرد بهتری نسبت به الگوریتم نزدیکترین همسایه برای پیش‌بینی سرطان سینه در زنان فعال داشت.

**نتیجه‌گیری:** با استفاده از الگوریتم‌های داده‌کاوی می‌توان سیستم‌های نوینی برای کمک به پزشکان طراحی نمود که موجب تسهیل در فرایندهای تشخیصی و درمانی شود. ترکیب عوامل خطر متعدد در مدل‌سازی برای پیش‌بینی سرطان سینه می‌تواند به تشخیص زودهنگام بیماری کمک کند.

**واژگان کلیدی:** سرطان سینه؛ نزدیکترین همسایه؛ ماشین بردار پشتیبان؛ زنان؛ پیش‌بینی

**ارجاع:** فصیحی لیلا، آقاعلی‌نژاد حمید، قراخانو رضا، خوزه آمارو گهیتی فرانسیسکو. مقایسه‌ی عملکرد الگوریتم‌های نزدیکترین همسایه و ماشین بردار پشتیبان برای پیش‌بینی سرطان سینه در زنان فعال. مجله دانشکده پزشکی اصفهان ۱۴۰۴؛ ۴۳ (۸۳۳): ۱۲۱۳-۱۲۱۹.

### مقدمه

در طی چندین سال اخیر، فعالیت بدنی انسان، با وجود زندگی ماشینی، به حداقل رسیده است و بشر را با فقر حرکتی روبه‌رو کرده است (۱). بین روش زندگی افراد و میزان ابتلا به سرطان سینه ارتباط قوی و مستقیم وجود دارد، به طوریکه نقش سبک زندگی در بروز سرطان‌های شایعی چون سرطان پروستات، سینه و روده از سایر علت‌ها پررنگ‌تر می‌باشد (۲). پزشکان همواره کم تحرکی را یکی از عوامل خطر بیماری سرطان عنوان می‌کنند و معتقدند که در میان عوامل خطر این بیماری مانند سن زیاد، چاقی، سابقه‌ی خانوادگی

سرطان و عدم فعالیت بدنی، تنها فعالیت بدنی و چاقی قابلیت اصلاح دارند (۳). بر اساس تحقیقات انجام شده، زنانی که فعالیت بدنی منظم دارند (به‌ویژه زنانی که در دوره یائسگی به‌سر می‌برند) تا ۲۵ درصد کمتر از افرادی که ورزش نمی‌کنند، در معرض خطر ابتلا به سرطان سینه قرار می‌گیرند (۴).

سرطان سینه، بیماری است که میلیون‌ها زن در سراسر جهان از آن رنج می‌برند و علت اصلی مرگ و میر ناشی از سرطان در زنان است (۵). بروز سرطان سینه در سال‌های اخیر به طور چشم‌گیر افزایش یافته است و در ایران بیشترین میزان ابتلا را در بین سرطان‌ها

۱- دانشجوی دکتری فیزیولوژی ورزشی، گروه تربیت بدنی و علوم ورزشی، دانشکده‌ی علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران.

۲. استاد فیزیولوژی ورزشی، گروه تربیت بدنی و علوم ورزشی، دانشکده‌ی علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران.

۳. دانشیار، گروه فیزیولوژی، دانشکده‌ی علوم پزشکی، دانشگاه گرانادا، گرانادا، اسپانیا

نویسنده‌ی مسؤول: : حمید آقا علی نژاد: استاد فیزیولوژی ورزشی، گروه تربیت بدنی و علوم ورزشی، دانشکده‌ی علوم انسانی، دانشگاه تربیت مدرس، تهران، ایران

Email: halinejad@modares.ac.ir

مربوط به زنان فعال به طور دردسترس برای شرکت در مطالعه به عنوان آزمودنی انتخاب شدند. معیارهای ورود به مطالعه شامل: جنسیت زن، سن بین ۲۵ تا ۷۵ سال، دارای سوابق پزشکی و آزمایشات بالینی در بیمارستان و در دسترس از طریق تلفن یا اینترنت بودند. معیارهای خروج از مطالعه شامل: خستگی غیرعادی، کم خونی، اختلال عملکرد فیزیکی ناشی از بیماری و مبتلا به بیماری‌های کلیه، کبد، پاراتیروئید تیروئید، دیابت قندی، بودند.

با بررسی پرونده‌های پزشکی ابتدا ۴۶ متغیر و سپس، با استفاده از نظر دو پزشک متخصص در زمینه‌ی سرطان سینه (دکتر اکرمی و یوسفی) و بررسی نتایج مقالات علمی، نهایتاً ۲۰ ویژگی آنترپوومتری، آزمایشگاهی و فیزیولوژیکی شامل: سن، وزن، شاخص توده‌ی بدن، سن اولین زایمان، بتا لیوپروئتین، سطوح سرمی سلنیم، کلسترول، فشار خون سیستولیک، فشارخون دیاستولیک، کلسترول لیپوپروتئین با چگالی بالا، سن قاعدگی، تعداد زایمان، سفتی قابل لمس در سینه، فعالیت جسمانی در اوقات فراغت، فعالیت جسمانی در هنگام کار، استفاده از داروهای ضد بارداری خوراکی، وجود توده، استفاده از هورمون درمانی جایگزین خوراکی، سابقه‌ی هیستریکتومی بیمار و ترشح در سینه به عنوان ویژگی‌های ورودی الگوریتم انتخاب شدند. در پژوهش حاضر، منظور از فعال زنانی بودند که به طور منظم در یک سال منتهی به پژوهش، هفته‌ای سه جلسه و هر جلسه به مدت ۹۰ دقیقه فعالیت بدنی (پیاده‌روی سریع و حرکات کششی ایستا و پویا) داشتند (۱۵). دو الگوریتم نزدیکترین همسایه و ماشین بردار پشتیبان، با به کارگیری متغیرهای ورودی و تعیین متغیر هدف ایجاد شدند (۱۶). مجموعه داده‌ها در قالب اکسل به نرم‌افزار متلب (MATLAB) نسخه‌ی ۲۰۲۴ انتقال داده شد و مورد تجزیه و تحلیل قرار گرفت.

### الگوریتم ماشین بردار پشتیبان

ماشین بردار پشتیبان (SVM)، یک الگوریتم یادگیری ماشینی نظارت شده است که برای کارهای طبقه‌بندی و رگرسیون استفاده می‌شود. هدف SVM یافتن هایپرپلان بهینه در فضای  $N$  بعدی برای جداسازی نقاط داده در کلاس‌های مختلف در فضای ویژگی است. این الگوریتم حاشیه بین نزدیک‌ترین نقاط کلاس‌های مختلف را به حداکثر می‌رساند (۱۷).

### الگوریتم نزدیکترین همسایه (KNN)

یک «یادگیری مبتنی بر نمونه» یا یادگیری غیر تعمیم‌کننده است که به عنوان الگوریتم «یادگیری تنبلی» نیز شناخته می‌شود KNN از داده‌ها استفاده می‌کند و نقاط داده جدید را بر اساس معیارهای شباهت (به عنوان مثال، تابع فاصله اقلیدسی) طبقه‌بندی می‌کند. طبقه‌بندی از رای اکثریت ساده  $k$  نزدیک‌ترین همسایه هر نقطه محاسبه می‌شود (۱۸).

به خود اختصاص داده است (۶). متوسط سن ابتلا به سرطان سینه در زنان کشورهای توسعه یافته ۴۸ سال است، اما یافته‌های حاصل از جامعه ایرانی حاکی از پایین‌تر بودن سن ابتلا به این بیماری می‌باشد (۷). هنگامی که سلول‌های سینه رشد می‌کنند و به شکل نامحدود تقسیم می‌شوند، توده‌ای از بافت به نام تومور را تشکیل می‌دهند (۸). اگر درمان مناسب انجام نشود، تومور ممکن است به سرطان تبدیل شود (۹). برخی عوامل خطر ابتلا به سرطان سینه را افزایش می‌دهند، از جمله افزایش سن، چاقی، مصرف مضر الکل، سابقه‌ی خانوادگی سرطان سینه، سابقه‌ی قرار گرفتن در معرض اشعه، سابقه‌ی باروری، مصرف دخانیات و هورمون درمانی پس از یائسگی (۱۰).

داده کاوی و یادگیری ماشین، از حوزه‌هایی هستند که باعث ایجاد تغییرات سریع در بسیاری از رشته‌ها بخصوص پزشکی شده‌اند (۱۱). داده کاوی بر اساس الگوریتم‌های خاصی با کشف الگوهای پنهان، شروع به کنکاش در رابطه بین علائم و ویژگی‌ها با متغیر هدف کرده و سعی می‌کند مدلی را استخراج کند که بتواند بر اساس آن بین افراد بیمار و سالم تمایز قائل شده و افراد مشکوک به بیماری را شناسایی کند (۱۲). رایج‌ترین اهداف این مدل‌ها، طبقه‌بندی و پیش‌بینی است که از علائم و نشانه‌ها برای پیش‌بینی بیماری استفاده می‌کند (۱۳). موضوعی که به آن کمتر پرداخته شده است استفاده از الگوریتم‌های داده‌کاوی در تشخیص بیماری سرطان سینه است (۱۴). این الگوریتم‌ها به پزشک کمک می‌کند که قبل از گسترش سرطان به سایر قسمت‌های بدن، پیش‌بینی درستی داشته باشد و اقدامات لازم را برای درمان بیمار در زمان مناسب انجام دهد. بعلاوه، شواهد حاکی از آن است که ابتلا به سرطان سینه اغلب با کاهش فعالیت جسمانی در فرد همراه است. بنابراین نیاز به برنامه‌ریزی به منظور افزایش فعالیت بدنی در مبتلایان به سرطان سینه بیشتر احساس می‌گردد. بنابراین، هدف از مطالعه‌ی حاضر، مقایسه‌ی عملکرد الگوریتم‌های نزدیک‌ترین همسایه و ماشین بردار پشتیبان برای پیش‌بینی سرطان سینه در زنان فعال بود.

### روش‌ها

جامعه‌ی آماری مطالعه‌ی حاضر را کلیه‌ی زنان مراجعه‌کننده به بیمارستان‌های امام خمینی و پژوهشکده سرطان معتمد، طی سال‌های ۱۳۹۳-۱۴۰۳ در در محدوده‌ی سنی ۲۵ تا ۷۵ سال، تشکیل دادند. تعداد پرونده‌های اولیه ۲۷۸۹ پرونده بود. این حجم از پرونده طی ۱۶ ماه توسط محقق و دو تن از کارکنان زده آزمایشگاهی مورد بررسی قرار گرفت. بعد از بررسی کلیه پرونده‌ها و ثبت مشخصات و آزمایشگاهی و تکمیل رضایت‌نامه و پرسشنامه فعالیت بدنی (از طریق وسایل ارتباط جمعی از قبیل واتس‌آپ و غیره)، نهایتاً تعداد ۶۴۱ پرونده

## یافته‌ها

در مجموع ۶۴۱ پرونده حاوی ۲۰ ویژگی دموگرافیک، آزمایشگاهی و آنترپومتری مرتبط با سرطان سینه در این مطالعه استفاده شد. داده‌ها با استفاده از نرم‌افزار Matlab R2024, The Mathworks Inc., (USA) توصیف، و تجزیه تحلیل شده است. جدول ۱ و ۲ متغیرهای کمی و کیفی پرونده‌های آزمودنی‌ها را نشان می‌دهد.

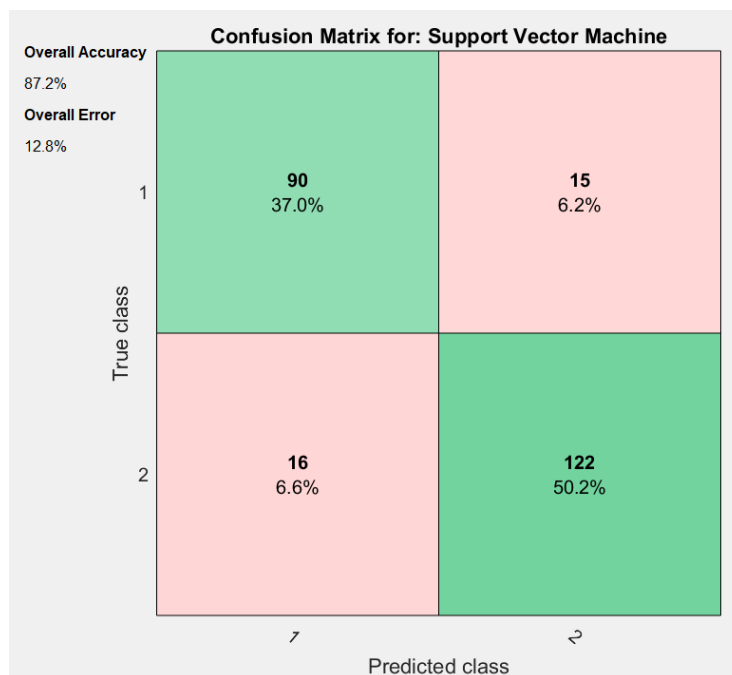
جدول ۱. متغیرهای کمی آزمودنی‌ها

| شماره | شاخص‌ها                                             | میانگین و انحراف استاندارد |
|-------|-----------------------------------------------------|----------------------------|
| ۱     | سن (سال)                                            | $49/76 \pm 11/17$          |
| ۲     | وزن (کیلوگرم)                                       | $62/31 \pm 12/52$          |
| ۳     | شاخص توده‌ی بدنی (کیلوگرم بر مترمربع)               | $27/59 \pm 2/12$           |
| ۴     | سن اولین زایمان                                     | $27/14 \pm 5/50$           |
| ۵     | بنا لیوپروتئین (میلی گرم بر کیلوگرم)                | $14/14 \pm 2/46$           |
| ۶     | فشارخون سیستولی (میلی متر جیوه)                     | $146/9 \pm 28/6$           |
| ۷     | فشارخون دیاستولیک (میلی متر جیوه)                   | $89/13 \pm 7/9$            |
| ۸     | سلنیوم سرم (میکروگرم در دسی‌لیتر)                   | $16/24 \pm 8/21$           |
| ۹     | کلسترول (میلی گرم بر دسی‌لیتر)                      | $277/54 \pm 41/43$         |
| ۱۰    | سن قاعدگی                                           | $14/03 \pm 3/07$           |
| ۱۱    | تعداد زایمان                                        | $4/50 \pm 4/50$            |
| ۱۲    | کلسترول لیوپروتئین با چگالی بالا (میلی مول بر لیتر) | $11/1 \pm 0/09$            |

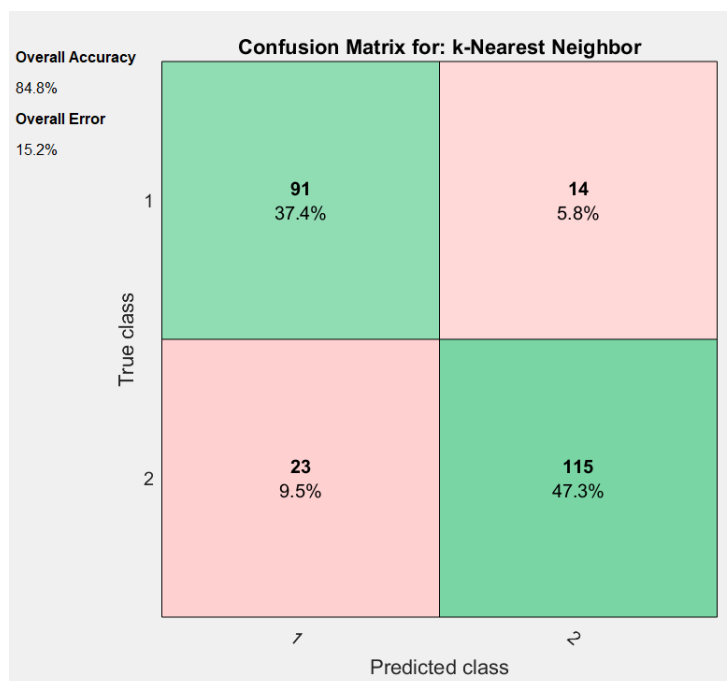
جدول ۲. متغیرهای کیفی آزمودنی‌ها

| شماره | ویژگی                                  | نوع (ناپیوسته)    | درصد     |
|-------|----------------------------------------|-------------------|----------|
| ۱     | فعالیت جسمانی در اوقات فراغت           | ۱ دارد<br>۰ ندارد | ۶۴<br>۳۶ |
| ۲     | فعالیت جسمانی در هنگام کار             | ۱ دارد<br>۰ ندارد | ۵۵<br>۴۵ |
| ۳     | سابقه‌ی هیستریکتومی بیمار              | ۱ دارد<br>۰ ندارد | ۶۰<br>۴۰ |
| ۴     | سفتی قابل لمس در سینه                  | ۱ دارد<br>۰ ندارد | ۶۵<br>۳۵ |
| ۵     | توده                                   | ۱ دارد<br>۰ ندارد | ۸۴<br>۱۶ |
| ۶     | ترشح در سینه                           | ۱ دارد<br>۰ ندارد | ۶۱<br>۳۹ |
| ۷     | استفاده از هورمون خوراکی درمان جایگزین | ۱ دارد<br>۰ ندارد | ۲۳<br>۷۷ |
| ۸     | استفاده از قرص‌های ضد بارداری خوراکی   | ۱ دارد<br>۰ ندارد | ۶۷<br>۳۳ |

از بین داده‌ها ۲۰ درصد برای تست و ۸۰ درصد برای آموزش الگوریتم در نظر گرفته شدند. نتایج حاصل از ماتریس درهم ریختگی این الگوریتم‌ها در شکل ۱ و ۲ نشان داده شده است. نتایج نشان داد که الگوریتم ماشین بردار پشتیبان، می‌تواند با دقت ۸۷/۲ درصد، صحت ۸۶/۳۵ درصد، حساسیت ۸۶/۸۸ درصد،



شکل ۱. ماتریس درهم ریختگی برای ماشین بردار پشتیبان



شکل ۲. ماتریس درهم‌ریختگی برای K-نزدیکترین همسایه

در مطالعه‌ی خود با عنوان الگوریتم‌های یادگیری ماشینی برای پیش‌بینی و تشخیص سرطان سینه از پنج الگوریتم یادگیری ماشین، مانند ماشین بردار پشتیبان (SVM)، جنگل تصادفی، رگرسیون لجستیک، درخت تصمیم (C4.5) و K-نزدیک‌ترین همسایه (KNN) بر روی مجموعه داده تشخیصی سرطان پستان ویسکانسین، استفاده کردند. آن‌ها پس از به دست آوردن نتایج، ارزیابی عملکرد و مقایسه بین نتایج را انجام دادند. هدف اصلی آن‌ها پیش‌بینی و تشخیص سرطان پستان با استفاده از الگوریتم‌های یادگیری ماشینی و یافتن مؤثرترین آن‌ها از نظر ماتریس سردرگمی، دقت و صحت بود. نهایتاً دقت ۹۷/۲۰ درصد را برای الگوریتم ماشین بردار پشتیبان گزارش کردند (۱۹).

در مطالعه‌ی Al-Quraishi و همکاران، با عنوان انتخاب ویژگی بهینه سازی ازدحام ذرات برای پیش بینی عود سرطان سینه، از ۱۹۸ آزمودنی با ۳۴ ویژگی بالینی و الگوریتم ژنتیک برای پیش بینی عود سرطان سینه استفاده کردند و در نهایت دقت و حساسیت، به ترتیب ۷۰ و ۸۱ درصد گزارش شد (۲۰).

Afshar و همکاران، بقای بیماران سرطان سینه را با استفاده از یک مجموعه داده با ۸۵۶ رکورد و ۱۵ بالینی با استفاده از مدل‌های یادگیری ماشین مورد مطالعه قرار دادند. در نتایج این مطالعه، الگوریتم درخت تصمیم‌گیری بالاترین حساسیت (۹۲/۲۱ درصد) و دقت (۸۴ درصد) را نشان داد (۲۱). Bagherian و همکاران، در مطالعه‌ای با استفاده از یک پایگاه داده با ۵۴۷۱ آزمودنی، از مشخصات دموگرافیک و عوامل خطر آزمایشگاهی برای پیش‌بینی سرطان سینه استفاده کردند

تشخیص‌پذیری ۸۵/۶۸ درصد و  $F1score$  ۶/۸۴ درصد زنان فعال دارای بیماری سرطان سینه را پیش‌بینی کند.

نتایج نشان داد که الگوریتم نزدیک‌ترین همسایه، می‌تواند با دقت ۸۴/۸۰ درصد، صحت ۸۳/۲۳ درصد، حساسیت ۸۲/۱۴ درصد، تشخیص‌پذیری ۸۳/۳۹ درصد و  $F1score$  ۸۳/۷۷ درصد زنان فعال دارای بیماری سرطان سینه را پیش‌بینی کند.

با توجه نتایج الگوریتم ماشین بردار پشتیبان در کلیه ویژگی‌های ارزیابی دقت، صحت، حساسیت، تشخیص‌پذیری و  $F1score$  عملکرد بهتری داشت.

### بحث

در این مقاله، ما یک روش پیش‌بینی سرطان سینه را با استفاده از تکنیک‌های یادگیری ماشینی مانند ماشین بردار پشتیبان و k-نزدیک‌ترین همسایه معرفی کرده‌ایم. هدف، استفاده از چند ویژگی منتخب است که منجر به پیش‌بینی دقیق‌تر توسط الگوریتم‌های یادگیری ماشین می‌شود. همچنین ایجاد یک راه‌حل قوی و موثر برای بهبود دقت پیش‌بینی سرطان سینه بود. نتایج این پژوهش نشان داد که الگوریتم ماشین بردار پشتیبان با دقت و صحت حساسیت، عملکرد بهتری نسبت به الگوریتم نزدیک‌ترین همسایه دارد.

چندین مطالعه با مجموعه داده متفاوت و از ویژگی‌های مشابه و در نمونه‌های مختلف تشخیص و پیش‌بینی سرطان سینه را با استفاده از تکنیک‌های یادگیری ماشینی مورد بررسی قرار داده‌اند. Naji و همکاران

از محدودیت‌های این مطالعه، نبود اطلاعات کافی مربوط به افراد سالم است، زیرا در اکثر داده‌های ثبتی تعداد نمونه افراد سالم بسیار کمتر از بیماران است. وجود این عدم تقارن می‌تواند منجر به عدم یادگیری کامل الگوریتم‌ها و کاهش دقت تشخیص شود.

### نتیجه‌گیری

ما معتقدیم که الگوریتم‌های یادگیری ماشین به زودی به عنوان ابزاری توانمند در بسیاری از محیط‌های بالینی و بیمارستانی برای تصمیم‌گیری درمان بالینی سرطان سینه رایج خواهد شد.

### تشکر و قدردانی

این مقاله منتج از پایان‌نامه‌ی مقطع دکتری رشته‌ی فیزیولوژی ورزشی با کد IR.MODARES.REC.1402.185 می‌باشد که در دانشگاه تربیت مدرس به تصویب رسیده و با حمایت مالی دانشگاه تربیت مدرس به انجام رسیده است. بدین‌وسیله از زحمات اساتید محترم و دانشگاه تربیت مدرس تقدیر و تشکر می‌شود.

و برای الگوریتم درخت تصمیم‌گیری (دقت ۸۲ درصد، حساسیت ۸۶ درصد و ویژگی ۷۷ درصد) را گزارش کردند (۲۲).

همچنین، Li و Chen، یک رویکرد تشخیص سرطان سینه با پنج مدل طبقه‌بندی مجزا ارائه کردند که شامل: درخت تصمیم، جنگل تصادفی، ماشین‌های بردار پشتیبان، شبکه‌ی عصبی مصنوعی و رگرسیون لجستیک. این روش از مجموعه داده‌های سرطان سینه کویمبرا و سرطان سینه ویسکانسین استفاده می‌کند. تجزیه و تحلیل تجربی مقایسه‌ای نشان داد که مدل جنگل تصادفی از چهار روش دیگر با دقت ۷۴/۳ درصد در سرطان سینه کویمبرا و ۹۶/۱ درصد در سرطان سینه ویسکانسین بهتر عمل کرد (۲۳).

تحقیقات نشان داد که روش‌هایی که از انتخاب ویژگی استفاده می‌کنند، مانند روش‌هایی که در (۲۴-۲۷) بیان شد، نسبت به طبقه‌بندی‌کننده‌هایی که بدون جزء انتخاب ویژگی در پیش‌بینی سرطان سینه کار می‌کنند، به دقت پیش‌بینی بهتری دست می‌یابند. الگوهای مشابهی در سایر تکنیک‌های بهینه‌سازی ویژگی در منابع دیگر مشاهده شده است (۲۸، ۲۹).

### References

- Pancholi S, Wachs JP, Duerstock BS. Use of artificial intelligence techniques to assist individuals with physical disabilities. *Annu Rev Biomed Eng* 2024; 26(1): 1-24.
- Justiz-Vaillant A, Gardiner L, Mohammed M, Surajbally M, Maharaj L, Ramsingh L, et al. Narrative Literature Review on Risk Factors Involved In Breast Cancer, Brain Cancer, Colon Rectal Cancer, Gynecological Malignancy, Lung Cancer, and Prostate Cancer. *Preprints* 2021; 2021040334.
- Marino P, Mininni M, Deiana G, Marino G, Divella R, Bochicchio I, et al. Healthy lifestyle and cancer risk: modifiable risk factors to prevent cancer. *Nutrients* 2024; 16(6): 800.
- Miyamoto T, Nagao A, Okumura N, Hosaka M. Effect of post-diagnosis physical activity on breast cancer recurrence: a systematic review and meta-analysis. *Curr Oncol Rep* 2022; 24(11): 1645-59.
- Abu Bashar MD, Begam N. Breast cancer surpasses lung cancer as the most commonly diagnosed cancer worldwide. *Indian J Cancer* 2022; 59(3): 438-9.
- Kazeminia M, Salari N, Hosseini-Far A, Akbari H, Bazrafshan M-R, Mohammadi M. The prevalence of breast cancer in Iranian women: a systematic review and meta-analysis. *Indian Journal of Gynecologic Oncology* 2022; 20(1): 14.
- Alizadeh M, Ghofazadeh M, Piri R, Mirza-Aghazadeh-Attari M, Mohammadi S, Naghavi-Behzad M. Age at diagnosis of breast cancer in Iran: a systematic review and meta-analysis. *Iran J Public Health* 2021; 50(8): 1564-76.
- Zaylaa AJ, Kourtian S. Advancing Breast Cancer Diagnosis through Breast Mass Images, Machine Learning, and Regression Models. *Sensors (Basel)* 2024; 24(7): 2312.
- Faruk M. Breast cancer resistance to chemotherapy: When should we suspect it and how can we prevent it? *Ann Med Surg (Lond)* 2021; 70: 102793.
- Mohanty SS, Mohanty PK. Obesity as potential breast cancer risk factor for postmenopausal women. *Genes Dis* 2021; 8(2): 117-23.
- Sahu M, Gupta R, Ambasta RK, Kumar P. Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis. *Prog Mol Biol Transl Sci* 2022; 190(1): 57-100.
- Moreno Gutiérrez S. ATOPE+: Supporting Personalized Exercise Interventions in Breast Cancer Care using Mobile Technologies and Machine Learning. [Thesis]. Department of Computer Architecture and Technology University of Granada; 2022.
- Mokoatle M, Marivate V, Mapiye D, Bormman R, Hayes VM. A review and comparative study of cancer detection using machine learning: SBERT and SimCSE application. *BMC Bioinformatics* 2023; 24(1): 112.
- Pal Choudhury P, Wilcox AN, Brook MN, Zhang Y, Ahearn T, Orr N, et al. Comparative validation of breast cancer risk prediction models and projections for future risk stratification. *J Natl Cancer Inst* 2020; 112(3): 278-85.
- Bula A, Tatar K, Wysocka R, Chyrek K, Piejko L, Nawrat-Szołtyś A, et al. Effect of physical activity on static and dynamic postural balance in women treated for breast cancer: A systematic review. *Int J Environ Res Public Health* 2023; 20(4): 3722.

16. Bansal M, Goyal A, Choudhary A. A comparative analysis of K-nearest neighbor, genetic, support vector machine, decision tree, and long short term memory algorithms in machine learning. *Decision Analytics Journal* 2022; 3: 100071.
17. Egwom OJ, Hassan M, Tanimu JJ, Hamada M, Ogar OM. An LDA-SVM machine learning model for breast cancer classification. *BioMedInformatics* 2022; 2(3): 345-58.
18. Assegie TA. An optimized K-Nearest Neighbor based breast cancer detection. *J Robot Control* 2021; 2(3): 115-8.
19. Naji MA, El Filali S, Aarika K, Benlahmar EH, Abdelouhahid RA, Debauche O. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Computer Science* 2021; 191: 487-92.
20. Al-Quraishi T, Abawajy JH, Chowdhury MU, Rajasegarar S, Abdalrada AS. Breast cancer recurrence prediction using random forest model. *Recent Advances on Soft Computing and Data Mining: Proceedings of the 3rd International Conference on Soft Computing and Data Mining (SCDM 2018)*, Johor, Malaysia, February 06-07, 2018.
21. Afshar HL, Jabbari N, Khalkhali HR, Esnaashari O. Prediction of breast cancer survival by machine learning methods: An application of multiple imputation. *Iran J Public Health* 2021; 50(3): 598-605.
22. Bagherian H, Haghjoo S, Mosayebi A, Arabzades S, Sharifi M, Sattari M. Identification of Effective Factors in Breast Cancer Survival in Isfahan Using Machine Learning Techniques. *J Adv Med Biomed Res* 2024; 32(154): 350-60.
23. Li Y, Chen Z. Performance evaluation of machine learning methods for breast cancer prediction. *Appl Comput Math* 2018; 7(4): 212-6.
24. Zhang D, Zou L, Zhou X, He F. Integrating feature selection and feature extraction methods with deep learning to predict clinical outcome of breast cancer. *Ieee Access* 2018; 6: 28936-44.
25. Fogliatto FS, Anzanello MJ, Soares F, Brust-Renck PG. Decision support for breast cancer detection: classification improvement through feature selection. *Cancer Control* 2019; 26(1): 1073274819876598.
26. Pudjihartono N, Fadason T, Kempa-Liehr AW, O'Sullivan JM. A review of feature selection methods for machine learning-based disease risk prediction. *Front Bioinform* 2022; 2: 927312.
27. Singh LK, Khanna M, Singh R. Efficient feature selection for breast cancer classification using soft computing approach: A novel clinical decision support system. *Multimed Tools Appl* 2024; 83(14): 43223-76.
28. Singh LK, Khanna M, Singh R. An enhanced soft-computing based strategy for efficient feature selection for timely breast cancer prediction: Wisconsin Diagnostic Breast Cancer dataset case. *ultimed Tools Appl* 2024; 83: 76607-72.
29. Pérez NP, López MAG, Silva A, Ramos I. Improving the Mann-Whitney statistical test for feature selection: An approach in breast cancer diagnosis on mammography. *Artif Intell Med* 2015; 63(1): 19-31.

## Comparison of the Performance of Nearest Neighbor and Support Vector Machine Algorithms for Breast Cancer Prediction in Active Women

Leila Fasihi<sup>1</sup>, Hamid Agha-Alinejad<sup>2</sup>, Reza Gharakhanlou<sup>2</sup>, Francisco J. Amaro Gahete<sup>3</sup>

### Original Article

### Abstract

**Background:** Breast cancer is one of the most common cancers in women, and its timely diagnosis plays an important role in survival and treatment. Machine learning has the potential to predict breast cancer based on hidden features in the data. The main objective of this study was to predict breast cancer in active women using Nearest Neighbor and Support Vector Machine algorithms.

**Methods:** In this developmental-applied study, data were collected from the medical records of 641 breast cancer patients at Imam Khomeini Hospital and the Motamed Cancer Research Institute between 2014 and 2024, within the age range of 25 to 75 years. After initial preprocessing of the dataset, the Nearest Neighbor and Support Vector Machine algorithms were applied.

**Findings:** The results showed that the Support Vector Machine algorithm performed better than the Nearest Neighbor algorithm for predicting breast cancer in active women, with an accuracy of 87.2%, a precision of 86.35%, a sensitivity of 88.86%, and a specificity of 68.85%.

**Conclusion:** Data mining algorithms can be used to design novel systems that assist physicians in facilitating diagnostic and therapeutic processes. Combining multiple risk factors in modeling for breast cancer prediction can help in the early diagnosis of the disease.

**Keywords:** Breast cancer; Nearest neighbor; Support vector machine; Women; Prediction

**Citation:** Fasihi L, Agha-Alinejad H, Gharakhanlou R, Amaro Gahete FJ. **Comparison of the Performance of Nearest Neighbor and Support Vector Machine Algorithms for Breast Cancer Prediction in Active Women.** J Isfahan Med Sch 2025; 43(833): 1213-19.

1- PhD. Student of Exercise Physiology, Department of Physical Education & Sport Sciences, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran.

2- Professor of Exercise Physiology, Department of Physical Education & Sport Sciences, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran.

3- Associate Professor, Department of Physiology, Faculty of Medical Sciences, University of Granada, Granada, Spain

**Corresponding Author:** Hamid Agha-Alinejad, Professor of Exercise Physiology, Department of Physical Education & Sport Sciences, Faculty of Humanities, Tarbiat Modares University, Tehran, Iran; Email: halinejad@modares.ac.ir