مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره اول شماره اول. ص ۳۱-۲۶، ۱۳۹۳

ایجاد یک مدل پیش‌آگهی مبتنی بر داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان

بهزاد کیانی[۱]، علیرضا آتشی[۱،۲]*

• دریافت مقاله: ۹۳/۷/۲۷ • پذیرش مقاله: ۹۳/۹/۱۲

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

روش: در این مطالعه گذشته‌نگر، از داده‌های ۸۰۹ بیمار مبتلا به سرطان پستان و ۱۸ ویژگی از هر بیمار، استفاده شد. با توجه به گمشدگی نسبتاً زیاد داده‌های این مجموعه، تنها اطلاعات ۶۶۵ بیمار قابل استفاده بود. به‌دلیل وجود مقادیر تهی در رکوردهای باقیمانده، این مقادیر از طریق الگوریتم EM و با استفاده از نرم‌افزار SPSS.V20، به‌عنوان یکی از فازهای پیش‌پردازش و آماده‌سازی داده‌ها، تخمین زده شده و در پایان، یک مدل پیش‌آگهی عود مجدد سرطان پستان در بین بیماران با به کارگیری درخت J48 بر روی داده‌ها ارائه شده‌است.

نتایج: ویژگی و حساسیت مدل توسعه یافته به‌ترتیب ۵۳ و ۸۵ درصد بود. این مدل، تنها ۱۴ درصد از بیماران دچار عود مجدد را به اشتباه، مستعد عود مجدد نمی‌داند.

نتیجه‌گیری: ایجاد مدل پیش‌بینی با ویژگی و حساسیت مناسب می‌تواند در مورد عود بیماری و انجام به موقع اقدامات پیشگیرانه برای جلوگیری از پیشرفت سرطان، هشدار مناسب را به بیماران بدهد. درصد منفی کاذب نیز در مدل‌های پیش‌بینی پزشکی بسیار اهمیت دارد، زیرا می‌تواند عواقب خطرناکی داشته باشد که در پژوهش حاضر این مقدار ۱۴ درصد بوده که از لحاظ مدلینگ، مقدار قابل قبولی به نظر می‌رسد.

کلید واژه‌ها: سرطان پستان، داده‌کاوی، مدل پیش‌آگهی

ارجاع: 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): ۳۱-۲۶.

۱. دانشجوی دکتری تخصصی انفورماتیک پزشکی، کمیته تحقیقات دانشجویی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران
۲. گروه پژوهشی انفورماتیک سرطان، مرکز تحقیقات سرطان پستان جهاد دانشگاهی، تهران، ایران
* **نویسنده مسؤول:** تهران، میدان ونک، ابتدای خیابان گاندی جنوبی، مرکز تحقیقات سرطان پستان جهاد دانشگاهی، گروه پژوهشی انفورماتیک سرطان.
• **شماره تماس:** ۸۸۶۷۷۴۷۸-۰۲۱ • **Email:** AtashiA901@mums.ac.ir

مقدمه

سرطان پستان یکی از شایع‌ترین سرطان‌ها در زنان است که تقریبا ده درصد از زنان را در مراحل مختلف زندگی تحت تأثیر قرار می‌دهد [۴-۱]. این سرطان، شایع‌ترین بدخیمی در زنان ایرانی و کانون اصلی توجهات در ایران است. در سال‌های اخیر، میزان شیوع بیماری روند رو به رشدی داشته و بررسی‌ها نشان می‌دهد که میزان بقای بیماران تا پنج و ده سال پس از تشخیص، به ترتیب ۸۸ و ۸۰ درصد بوده است [۲].

تمام تومورها سرطانی نیستند و ممکن است خوش‌خیم یا بدخیم باشند. تومورهای خوش خیم، رشد غیر طبیعی دارند ولی به ندرت کشنده هستند. با این حال، تعدادی از توده‌های خوش‌خیم پستان نیز می‌توانند خطر ابتلا به سرطان پستان را افزایش دهند. همچنین در برخی از زنان دارای سابقه نمونه‌برداری از توده‌های خوش‌خیم پستان نیز، خطر سرطان پستان افزایش یافته است. از طرف دیگر، تومورهای بدخیم، جدی‌تر بوده و سرطانی محسوب می‌شوند ولی تشخیص زودهنگام این نوع از سرطان‌ها شانس درمان موفقیت‌آمیز را بالا برده است [۴].

داده‌کاوی از روش‌هایی است که برای تشخیص یا پیش‌بینی سرطان‌ها به کار می‌رود. این روش یکی از پرطرفدارترین رویکردهای پیش‌بینی عود مجدد سرطان پستان است. بنابراین، رویکردهای داده‌کاوی می‌توانند با کاهش تعداد نتایج مثبت و منفی کاذب، پزشکان را در شناسایی بهتر سرطان پستان کمک کنند [۵،۶]. در نتیجه، رویکردهای جدید مانند کشف دانش از پایگاه داده (Knowledge Discovery and Data Mining= KDD) که شامل تکنیک‌های داده‌کاوی هستند، روز به روز محبوبیت بیشتری یافته و به یک ابزار تحقیقاتی مطلوب برای پژوهشگران علوم پزشکی تبدیل شده‌اند. به کمک آنها پژوهشگران می‌توانند الگوها و روابط بین تعداد زیادی از متغیرها را شناسایی کرده و پیش‌بینی نتایج حاصل از یک بیماری با استفاده از ذخایر اطلاعاتی موجود در پایگاه‌های داده، برای آنها امکان‌پذیر شده است [۷]. بررسی‌ها و پژوهش‌های گوناگونی در زمینه مشکلات ناشی از پیش‌بینی بقای بیماران مبتلا به سرطان پستان، با استفاده از روش‌های آماری و شبکه‌های عصبی مصنوعی، صورت گرفته و پژوهش‌ها در حوزه پزشکی برای یافتن روابط بین داده‌ها با استفاده از روش‌های داده‌کاوی، افزایش یافته است [۹-۷].

در علوم پزشکی پژوهش‌هایی با استفاده از داده‌کاوی و با رویکرد پیش‌بینی انجام شده است. به عنوان مثال، دلن (Delen) و همکاران با تجزیه و تحلیل پایگاه‌های بزرگ داده، از شبکه‌های عصبی مصنوعی، درخت تصمیم‌گیری و از رگرسیون لجستیک برای توسعه مدل‌های پیش‌بینی سرطان پستان استفاده کردند. نتایج تحقیقات آنها نشان داد که الگوریتم درخت تصمیم برای استخراج دانش از داده‌های موجود، بر سایر روش‌ها مقدم است و نتایج به دست آمده از تحقیق، نزدیک به واقعیت بود [۸]. همچنین لند (Land) و ورهگن (Verheggen)، تنها از ماشین بردار پشتیبان (Support Vector Machine=SVM) برای طبقه‌بندی تراکم تومور سرطان پستان، استفاده کردند. نتایج به دست آمده نشان داد که SVM، روش مناسبی بوده و نتایج به دست آمده با شواهد موجود و واقعی، مطابقت داشت [۱۰]. لاندین (Lundin) و همکارانش از مدل شبکه‌های عصبی مصنوعی و رگرسیون لجستیک برای پیش‌بینی پنج، ده و پانزده ساله بقای بیماران مبتلا به سرطان پستان استفاده کردند. آنها ۹۵۱ بیمار مبتلا به سرطان پستان را مورد مطالعه قرار داده و اندازه تومور، وضعیت گره‌های لنفی، نوع بافت، تشکیل توبول، نکروز تومور و سن را به عنوان متغیرهای ورودی در نظر گرفتند. سپس به این نتیجه رسیدند که درختان طبقه‌بندی و همچنین رگرسیون لجستیک برای تفسیر بالینی، بسیار آسان‌تر است [۱۱].

پندهارکر (Pendharkar) و همکاران از چندین روش داده‌کاوی برای بررسی الگوهای موجود در سرطان پستان استفاده کردند. در این مطالعه، آنها نشان دادند که می‌توان از داده‌کاوی به عنوان ابزاری ارزشمند برای شناسایی شباهت‌ها (الگوها) در مورد سرطان پستان به منظور تشخیص، پیش‌آگهی و درمان، استفاده کرد [۱۲]. از طرفی، در کشور ایران هم پژوهش‌هایی با رویکردهای داده‌کاوی در حوزه سرطان پستان انجام شده است. برای مثال، طلوعی اشلقی و همکاران در یک پژوهش، از سه تکنیک مختلف داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان استفاده و آنها را مقایسه کردند. در نهایت هر سه راهبرد، در پیش بینی عود مجدد سرطان پستان، دقت بالایی داشتند و SVM (Support Vector Machine)، بیشترین دقت در پیش‌بینی عود مجدد را داشت [۱۳]. در پژوهش دیگری عظیمیان و همکاران، از روش‌های داده‌کاوی برای تشخیص بیماری سرطان پستان در زنان استفاده کردند که تشخیص آنان با دقت بالایی صورت گرفت [۱۴]. این پژوهش‌ها نمونه‌هایی

از کاربرد داده‌کاوی در علـوم پزشـکی برای پیش‌بینی بیماری‌ها هستند.

با عنایت به اهمیت ویژه بیماری سرطان پستان و با توجه به مزیت مضاعف روش‌های داده‌کاوی، پژوهشگران این مطالعه در صدد برآمدند تا به وسیله شیوه‌های داده‌کاوی و با استفاده از داده‌های موجود مربوط به سرطان پستان یک پایگاه داده در کشورمان، یک مدل پیش‌بینی مبتنی بر دسته‌بندی داده‌ها برای پیش‌بینی عود مجدد سرطان پستان را ارائه دهند. پیش‌بینی عود مجدد سرطان پستان از جهات زیادی اهمیت دارد، چرا که در مورد عود مجدد سرطان، هشدار لازم به بیمار داده می‌شود و بیمار نیز اقدامات تشخیصی و درمانی لازم را زودتر شروع می‌کند.

روش

در این پژوهش گذشته‌نگر، داده‌ها از مرکز تحقیقات سرطان دانشگاه شهید بهشتی دریافت شد. این داده‌ها مربوط به ۸۰۹ بیمار مبتلا به سرطان پستان و دارای ۱۸ ویژگی برای هر بیمار بود. با توجه به گمشدگی بسیار زیاد داده‌ها در این مجموعه، تنها اطلاعات مربوط به ۶۶۵ بیمار قابل استفاده بود. بنابراین، در این پژوهش، از ۶۶۵ رکورد مربوط به افرادی که مبتلا به سرطان پستان بودند، استفاده شد. با توجه به اینکه تعدادی از فیلدهای موجود در رکوردهای باقیمانده، دارای مقادیر تهی بودند، این مقادیر از طریق الگوریتم EM (Expectation Maximization Algorithm) و با استفاده از نرم‌افزار SPSS.V20، به عنوان یکی از فازهای پیش‌پردازش و آماده‌سازی داده‌ها، تخمین زده شد. سپس داده‌های پیوسته به مقادیر گسسته تبدیل شدند. پس از این مرحله، دسته‌بندی (Classification) داده‌ها با در نظر گرفتن فیلد عود مجدد سرطان پستان به عنوان برچسب کلاس، انجام شد. برای این منظور از الگوریتم درخت J48 برای پیش‌بینی عود مجدد سرطان و برای ارزیابی مدل ایجاد شده از روش K-Fold با K=10 استفاده شد. این الگوریتم از طریق ابزار داده‌کاوی WEKA اجرا شد.

نتایج

پس از ساخت درخت دسته‌بندی توسط ابزار داده‌کاوی WEKA، ماتریس برخورد، تشکیل شد که با استفاده از آن، ویژگی و حساسیت مدل ایجادشده، محاسبه شد (جدول ۱).

جدول ۱: ماتریس برخورد مدل پیش‌بینی ارائه شده

کلاس واقعی	کلاس پیش‌بینی شده توسط مدل	
	بیماری عود می‌کند	بیماری عود نمی‌کند
بیماری عود کرده است	۳۶۵ مثبت واقعی (True Positive=TP)	۶۱ منفی کاذب (False Negative=FN)
بیماری عود نکرده است	۱۱۱ مثبت کاذب (False Positive=FP)	۱۲۸ منفی واقعی (True Negative=TN)

پارامتر حساسیت مدل عبارت از نسبت تعداد افرادی که در مدل به عنوان مستعد عود مجدد پیش‌بینی شده‌اند، به کل افرادی که واقعا دچار عود مجدد شده‌اند. پارامتر ویژگی، مشخص‌کننده تعداد افرادی است که توسط مدل به درستی به عنوان افرادی که دچار عود مجدد سرطان پستان نمی‌شوند، تشخیص داده شده‌اند. پارامتر مثبت کاذب مشخص‌کننده درصدی است که دچار عود مجدد سرطان پستان نشده‌اند اما توسط مدل به اشتباه مستعد عود مجدد سرطان پستان تشخیص داده شده‌اند و پارامتر منفی کاذب، مشخص‌کننده درصد افرادی است که دچار عود مجدد سرطان پستان شده‌اند اما توسط مدل، پیش‌بینی شده است که دچار عود نمی‌شوند. مقادیر پارامترهای فوق و صحت (Accuracy) و دقت (Precision) و نحوه محاسبه آنها در جدول ۲ نشان داده شده است.

جدول ۲: مشخصات مدل پیش‌بینی عود مجدد سرطان پستان

نام پارامتر	فرمول	مقدار
حساسیت (True Positive Rate or Sensitivity or Recall Rate)	$\frac{TP}{TP+FN}$	۸۵
ویژگی (True Negative Rate)	$\frac{TN}{TN+FP}$	۵۳
مثبت کاذب (False Positive Rate)	$\frac{FP}{TP+FP}$	۴۶
منفی کاذب (False Negative Rate)	$\frac{FN}{FN+TN}$	۱۴
دقت (Precision)	$\frac{TP}{TP+FP}$	۷۷
صحت (Accuracy)	$\frac{TP+TN}{N}$	۷۵

بحث و نتیجه‌گیری

در این مطالعه با استفاده از الگوریتم‌های داده‌کاوی (درخت تصمیم)، یک مدل پیش‌آگهی مبتنی بر داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان با توجه به داده‌های موجود مرکز تحقیقات سرطان پستان، ایجاد شد. سرطان‌ها از جمله بیماری‌هایی هستند که عوامل بسیار زیادی در ابتلا به آنها مؤثر می‌باشند. این ویژگی بیانگر مؤثر بودن استفاده از تکنیک‌های داده‌کاوی می‌باشد.

پیش‌بینی عود مجدد سرطان پستان از جهات متعددی اهمیت دارد. برای نمونه، در مورد بیمارانی که فقط تومور و یا یکی از پستان‌ها به طور کامل برداشته می‌شود، در صورت بالا بودن میزان احتمال عود سرطان پستان، می‌توان قبل از گسترش سرطان به بقیه قسمت‌های بدن، اقدامات خاصی را انجام داد. مدل‌های پیش‌بینی از این حیث می‌توانند بسیار مفید باشند. اما باید توجه کرد که در حوزه ارزیابی مدل‌های پیش‌بینی پزشکی باید به دو پارامتر ویژگی و حساسیت مدل با هم توجه نمود، زیرا در نظر گرفتن یکی از آنها به تنهایی می‌تواند گمراه‌کننده باشد. علاوه بر این باید به مقدار پارامتر منفی کاذب توجه خاصی داشت. درصد این پارامتر در مدل‌های پیش‌بینی در حوزه پزشکی بسیار اهمیت دارد چون فرد بیمار به اشتباه، سالم در نظر گرفته می‌شود که می‌تواند عواقب بسیار خطرناکی داشته باشد. در مدل پیش بینی ارائه شده در این تحقیق این مقدار ۱۴ درصد بود که مقدار نسبتا کمی است و از این حیث می‌توان این مدل را قابل قبول دانست.

پژوهش‌های مشابهی که در حوزه سرطان پستان انجام شده، بیشتر شامل مدل‌های وابسته به ژن‌های ایجاد کننده سرطان می‌باشد اما نتایج این پژوهش از جهت میزان نتایج، جالب توجه و قابل مقایسه با نتایج پژوهش طلوعی اشلقی و همکاران می‌باشد، به طوری‌که این پژوهش در تشخیص میزان منفی کاذب، دقت کمتری دارد. اگر چه مطالعه مذکور یک مطالعه مقایسه‌ای بین روش‌های داده‌کاوی است، اما هر سه تکنیک مورد استفاده در آن مطالعه دارای دقت بیشتری نسبت به این پژوهش هستند که احتمالا مربوط به نوع داده‌های پایگاه داده می‌باشد [۱۳].

یکی از محدودیت‌های این پژوهش، مقادیر زیاد داده‌های از دست‌رفته بود. همان گونه که در روش اجرای تحقیق توضیح داده شد، این مقادیر توسط الگوریتم EM تخمین زده شد. در صورتی که مقادیر داده‌های از دست‌رفته زیاد نبودند، می‌توان پیش‌بینی کرد که دقت و بازنمایی و سایر پارامترهای ارزیابی مدل پیش‌بینی ارائه شده به صورت قابل توجهی بهبود یابد. محدودیت دیگر، تعداد کم بیماران برای ایجاد مدل است که باید در مطالعات بعدی، این مدل با تعداد بیشتری از بیماران توسعه یابد. به هر حال مانند تمامی پژوهش‌های داده‌کاوی، نتایج این پژوهش فقط برای پایگاه داده مورد مطالعه (و نه سایر پایگاه‌ها، مگر به شرط توسعه) معتبر است. از طرفی، ارزیابی این مدل توسط روش K-Fold با K=10 انجام شده است که پیشنهاد می‌شود در مطالعات بعدی، مدل‌های ایجاد شده توسط بیماران واقعی که در آینده به مراکز پزشکی مراجعه می‌کنند، ارزیابی شود.

References

1. Hortobagyi GN, de la Garza Salazar J, Pritchard K, Amadori D, Haidinger R, Hudis CA,, et al. The global breast cancer burden: variations in epidemiology and survival.Clin Breast Cancer. 2005; 6(5):391-401.

2. Setayeshi S, Akbari ME, Darghahi R, Haghighat khah HR."Breast Cancer and Technical Analysis of its Diagnostics". Tehran: Bitarafan; 2011. Persian.
3. Jemal A, Bray F, Center MM, Ferlay J, Ward E, Forman D. Global cancer statistics. CA Cancer J Clin. 2011; 61(2):69-90.
4. American Cancer Society. Breast cancer facts & figures 2009-2010. [cited 2006 Feb 11]. Available from: http://www.cancer.org
5. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory; Pittsburgh: ACM Press. 1992. p. 144–52.
6. Calle J. Breast cancer facts and figures 2003-2004. American Cancer Society 2004. [cited 2006 Feb 11]. Available from: http://www.cancer.org
7. Breast cancer Q & A/ facts and statistics. [cited 2006 Mar 11]. Available from (http://www.komen.org/bci/bhealth/QA/q_and_a.asp).
8. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artif Intell Med. 2005; 34(2):113-27.
9. Lugo-Reyes SO, Maldonado-Colín G, Murata C .Artificial intelligence to assist clinical diagnosis in medicine. Rev Alerg Mex. 2014; 61(2):110-20.
10. Land WH Jr, Verheggen EA. Multiclass primal support vector machines for breast density classification. Int J Comput Biol Drug Des. 2009; 2(1):21-57.
11. Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology. 1999; 57(4):281-6.
12. Pendharkar PC, Rodger JA, Yaverbaum GJ, Herman N, Benner M. Associations statistical, mathematical and neural approaches for mining breast cancer patterns. Expert Syst Appl. 1999; 17(3):223-32.
13. Toloie Ashlaqi A, PourEbrahimi A, Ebrahimi M, GhasemAhmad L. Using Data Mining Techniques for Prediction 11. Breast Cancer Recurrenc. Iran J Breast Dis. 2012; 5(4):23-34. Persian.
14. Azimian F, Tadaion-Tabrizi GN, Jalali M. Breast Cancer Detection Using Data Mining Techniques. 4th Iran Data Mining Conference (IDMC), Tehran; 2010: 31- 1

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2014; 1(1): 26-31

A Prognostic Model Based on Data Mining Techniques to Predict Breast Cancer Recurrence

Behzad Kiani[1], Alireza Atashi[1,2]*

• **Received:** 19 Oct, 2014 • **Accepted:** 3 Dec, 2014

Introduction: Breast cancer is one of the most common cancers, and also it is the most common type of malignancy in Iranian women that has been growing in recent years. The risk of recurrence is usual in patient. Many factors may increase or decrease the recurrence rate. Data mining methods have been used to diagnose or predict cancer and one of the most application of data mining approaches is prediction of breast cancer recurrence

Method: This is a retrospective study. Collected data on 809 patients with breast cancer with 18 fields for each patient were used. Due to excessive missing data only about 665 cases have been used. Since the number of fields in the remaining records with null values have been observed, as a preprocessing and data preparation phases, these values have been estimated by the EM algorithm and using SPSS.v20 software. In this study, a model for prognosis of breast cancer recurrence among patients using J48 tree has been developed.

Results: The specificity and sensitivity of the developed model are 53% and 85%, respectively. Moreover, only 14% of patients who have relapsed are known as false negative with developed model.

Conclusion: Creating a predictive model with appropriate specificity and sensitivity can warn patients about recurrence and timely preventive measures to prevent progression of the cancer. The False Negative rate is very important in medical prediction models that can make serious results/consequences. In present study this rate is about 14% that seems reasonable amount in term of modeling.

Key words: Breast cancer, Data mining, Prognostic model

• **Citation:** Kiani B, Atashi A. A Prognostic Model Based on Data Mining Technics to Predict Breast Cancer Recurrence. Journal of Health and Biomedical Informatics 2014; 1(1): 26-31.

1. Ph.D Candidate of Medical Informatics, Research Committee, Department of Medical Informatics, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
2. Cancer Informatics Department, Breast Cancer Research Center, ACECR, Tehran, Iran.

***Correspondence**: Cancer Informatics Department, Breast Cancer Research Center 146, South Gandhi Ave., Vanak Sq., Tehran, Iran.
• **Tel:** 021-88677478 • **Email:** AtashiA901@mums.ac.ir