کاربرد تکنیک‌های داده‌کاوی در پیش‌بینی احتمال بقای بیماران مبتلا به سرطان پستان: *مقاله مروری*

حسین باقریان[1]، شقایق حق‌جوی جوانمرد[2]، مهران شریفی[3]، محمد ستاری[1*]

۱- مرکز تحقیقات فناوری اطلاعات در امور سلامت، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.
۲- مرکز تحقیقات فیزیولوژی کاربردی، پژوهشکده قلب و عروق، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.
۳- گروه هماتولوژی و انکولوژی دانشکده پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.

[*] نویسنده مسئول: اصفهان، دانشگاه علوم پزشکی اصفهان، مرکز تحقیقات فناوری اطلاعات در امور سلامت.
تلفن: ۰۳۱-۳۷۹۲۵۱۵۲
E-mail: msattarimng.mui@gmail.com

دریافت: ۱۳۹۹/۱۱/۲۶ ویرایش: ۱۳۹۹/۱۲/۰۳ پذیرش: ۱۴۰۰/۰۲/۲۵ آنلاین: ۱۴۰۰/۰۳/۰۱

چکیده

این مقاله مروری نظام‌مند در سال ۹۸ در دانشگاه علوم پزشکی اصفهان انجام شده است. بررسی مطالعات مختلف مشخص کرد که از چه تکنیک‌های داده‌کاوی برای پیش‌بینی احتمال بقا، از چه ریسک فاکتورهایی برای این پیش‌بینی‌ها، ازچه معیارهایی برای ارزیابی تکنیک‌های داده‌کاوی و در نهایت چه منابع داده‌ای برای پیش‌بینی احتمال بقا مبتلایان به سرطان پستان استفاده شده است. این مطالعه از نوع مروری نظام‌مند است. مطابق با بیانیه پریزما، مطالعات منتشر شده در حوزه‌ی پیش‌بینی احتمال بقا بیماران مبتلا به سرطان پستان با استفاده از تکنیک‌های داده‌کاوی به زبان انگلیسی و در فاصله‌ی سال‌های ۲۰۰۵ تا ۲۰۱۸ در پایگاه‌های داده Science direct Medline، Web of science و Embase و پایگاه استنادی Scopus مورد جستجو و بازیابی قرار گرفت. پس از جستجو، تعداد ۵۲۷ مقاله بازیابی شد. پس از حذف موارد تکراری و ارزیابی مقالات در نهایت ۲۱ مقاله مورد استفاده قرار گرفت. سه تکنیک رگرسیون لجستیک، درخت تصمیم و ماشین بردار پشتیبان بیشترین کاربرد را در مقالات داشته اند. در بیشتر مطالعات، ریسک فاکتورهای سن، گرید تومور، استیج تومور و اندازه تومور استفاده شده بودند. معیار صحت در مطالعات بیشتری استفاده شده بود. اکثر مطالعات مورد بررسی از پایگاه داده برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی که متعلق به کشور آمریکا می‌باشد، استفاده کرده بودند. به‌طور معمول در حوزه‌ی پیش‌بینی احتمال بقا، تکنیک‌های داده‌کاوی در حوزه‌ی کلاسه‌بندی به‌دلیل انطباق با این حوزه بیشتر مورد توجه قرار می‌گیرند. بر همین اساس، تکنیک‌های درخت تصمیم، رگرسیون لجستیک و ماشین بردار پشتیبان در مطالعات مورد بررسی، بیشترین کاربرد را داشته‌اند.

کلمات کلیدی: سرطان پستان، داده‌کاوی، احتمال بقا.

سرطان پستان نوعی تومور بدخیم است که طی آن سلول‌های بافت پستان به‌علت اختلالات ژنتیکی مانند جهش، افزایش کروموزومی، حذف، بازسازمانی، جابه‌جاشدگی و تکرارشدگی کروموزومی بدون هیچ کنترلی شروع به تکثیر شدن می‌کنند.[1] در حال حاضر سرطان سینه، شایع‌ترین سرطان در میان زنان در سراسر جهان می‌باشد.[2و5] در واقع بسیاری از زنان در معرض این بیماری هستند به‌گونه‌ای که ۲۴/۲٪ سرطان‌هایی که در زنان در سال ۲۰۱۸ شناسایی شده، سرطان پستان می‌باشد.[6و7و8و9] این سرطان در میان زنان در سطح جهانی، رتبه نخست مرگ و میر ناشی از سرطان را نیز به خود اختصاص داده است. طبق گزارش سازمان جهانی بهداشت در سال

۲۰۱۸ در حدود ۱۵٪ مرگ‌های ناشی از سرطان در زنان مربوط به سرطان پستان بوده است.[۹] برای آن‌که بتوان میزان مرگ و میر ناشی از این نوع سرطان را کاهش داد بایستی روش‌های مختلف درمانی ارتقا یابند. بررسی میزان بقا در بیماران مبتلا به سرطان سینه، به‌عنوان شاخصی جهت ارزیابی و ارتقا روش‌های مختلف درمانی محسوب می‌شود.[۱۰] تحلیل بقا رویکردی مهم در تعیین عوامل مرتبط با بقای بیماری است و روش‌های تحلیل بقا با نوعی از داده‌ها سر و کار دارند که در آن‌ها زمان انتظار تا رخداد یک حادثه را اندازه می‌گیرند.[۱۱] احتمال بقا بیماران مبتلا به سرطان پستان با توجه به ویژگی‌های بالینی افراد متفاوت می‌باشد. در مقایسه با سایر انواع سرطان‌ها احتمال بقا بیماران مبتلا به سرطان پستان بالاتر بوده به‌ویژه اگر این بیماری زود تشخیص داده شود.[۴] بهبود میزان بقا این بیماری مسائل جدیدی از جمله ابداع روش‌های مناسب‌تر جهت غربالگری و پیشگیری از وقوع بیماری و تشخیص به‌موقع عود و متاستاز، بهبود مراقبت از عوارض دراز مدت درمان و حفظ کیفیت زندگی آنان و تصمیم‌گیری‌های پیچیده‌تر در درمان بیماران را پیش روی پزشکان و محققین قرار می‌دهد.[۱۲]

در سال‌های اخیر رویکردهای داده‌کاوی مانند کشف دانش جدید از پایگاه‌های داده به‌عنوان ابزار تحقیقاتی مناسب برای پژوهشگران حوزه پزشکی تبدیل شده است و محققین از تکنیک‌های مختلف این حوزه برای پیش‌بینی بقا یا عود مجدد سرطان‌ها استفاده می‌کنند.[۱۳] داده‌کاوی از روش‌هایی است که به‌منظور یافتن الگو از درون پایگاه‌های بزرگ داده و نیز پیش‌بینی نتایج سرطان‌ها با استفاده از الگوریتم‌های پیش‌بینی‌کننده بکار می‌رود. با کمک داده‌کاوی می‌توان الگوهایی که به سختی قابل تشخیص هستند را از پایگاه‌های داده استخراج نمود. مطالعات مختلفی از تکنیک‌های داده‌کاوی برای پیش‌بینی احتمال بقا در بیماران مبتلا به سرطان استفاده کرده اند.[۱۴و۱۷]

Delen و همکاران در مطالعه خود از دو تکنیک درخت تصمیم (decision tree) و شبکه عصبی مصنوعی (Artificial Neural Networks–ANN) به همراه تکنیک آماری رگرسون لجستیک (logistic regression) برای پیش‌بینی احتمال بقا بیماران مبتلا به سرطان پستان استفاده کردند.[۱۸]

Liu و همکاران در مطالعه " Decision tree based predictive models for breast cancer survivability on imbalanced data" بر پایه داده‌های نامتعادل از درخت تصمیم و الگوریتم بگینگ (bagging algorithm) برای پیش‌بینی احتمال بقا در مبتلایان به سرطان پستان استفاده کردند.[۱۹] بسیاری از مطالعات برای پیش‌بینی احتمال بقا بیماران مبتلا به سرطان پستان از تکنیک‌های مختلف داده‌کاوی در شرایط متفاوتی استفاده کرده‌اند و هر یک از این تکنیک‌ها یک یا نهایتا دو تکنیک داده‌کاوی بصورت مستقل یا در مقایسه با یکدیگر را مورد استفاده قرار داده‌اند. با در نظر گرفتن اهمیت پیش‌بینی احتمال بقا بیماران مبتلا به سرطان پستان و با توجه به اهمیت شناسایی عوامل تاثیرگذار بر احتمال بقای این بیماران، پژوهشگران قصد دارند در این پژوهش کلیه‌ی مطالعاتی را که از تکنیک‌های داده‌کاوی برای پیش‌بینی احتمال بقای بیماران مبتلا به سرطان پستان استفاده کرده‌اند، مورد جستجو، بررسی و تجزیه و تحلیل قرار داده در نهایت یک چهارچوب علمی برای انجام پژوهش‌های آتی در این زمینه ارایه نمایند.

این مطالعه از نوع مروری نظام‌مند است. مطابق با بیانیه پریزما، PRISMA Statement (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) مطالعات منتشر شده در حوزه‌ی پیش‌بینی احتمال بقا بیماران مبتلا به سرطان پستان با استفاده از تکنیک‌های داده‌کاوی به زبان انگلیسی و در فاصله‌ی بین سال‌های ۲۰۰۵ تا ۲۰۱۸ در پایگاه‌های داده Medline، Science direct، Web of science و Embase و پایگاه استنادی Scopus مورد جستجو قرار گرفت.[۲۰]

پس از مشخص شدن کلید واژه‌های اصلی و کلید واژه‌های مترادف آن‌ها با بهره‌گیری از or، هر یک از کلید واژه‌های اصلی با مترادف‌هایش ترکیب و مورد جستجو قرار گرفت و نتیجه جستجو ذخیره شد. در ادامه نتایج حاصل از جستجوی هر یک از کلید واژه‌های اصلی و مترادف آنها با استفاده از عملگر and، با یکدیگر ترکیب و برای جستجو مورد استفاده قرار گرفتند. لازم به ذکر است که از علائم ستاره و علامت سوال برای پوشش ترکیب‌های مختلف واژه‌های مورد جستجو استفاده شد.

مطالعاتی که به زبان انگلیسی در زمینه‌ی استفاده از تکنیک‌های داده‌کاوی در پیش‌بینی احتمال بقای بیماران مبتلا به سرطان پستان در فاصله سال‌های ۲۰۰۵ تا ۲۰۱۸ انجام شده بودند در این مطالعه مورد استفاده قرار گرفتند. مطالعاتی که به سایر زبان‌های غیر از انگلیسی

منتشر شده بودند و موضوع آنها استفاده از تکنیک‌های داده‌کاوی درحوزه‌های مانند تشخیص، عود و تعیین بدخیم و خوش‌خیم بودن سرطان پستان استفاده کرده بودند، کنار گذاشته شدند.

پس از جستجوی اولیه تعداد ۵۲۷ مطالعه بازیابی شد. از این تعداد، ۴۷ مطالعه تکراری بود که حذف گردید. در ادامه عناوین و چکیده مطالعات باقی‌مانده (۴۸۰ مطالعه) مطالعه شد و برهمین اساس ۴۵۱ مطالعه که فاقد شرایط ورود به این پژوهش بودند (عنوان و هدف آن‌ها با معیارهای این مطالعه همخوانی نداشت)، کنار گذاشته شدند. پس از مطالعه متن کامل مطالعات باقی‌مانده (۲۹ مورد)، تعداد نه مطالعه شرایط ورود به این پژوهش را نداشتند و حذف شدند و در نهایت تعداد ۲۰ مطالعه انتخاب و مورد استفاده قرار گرفت (جدول ۱).

از میان۵۲۷ مطالعه‌ای که در ابتدا بازیابی شده بود، در نهایت ۲۰ مطالعه مورد استفاده قرار گرفت. در ادامه هر یک از این مطالعات به‌صورت اجمالی شرح داده شده است. پس از تجزیه و تحلیل مطالعات منتخب، مشخص شد که در مجموع ۸۶ ریسک فاکتور در این ۲۰ مطالعه مورد استفاده قرار گرفته بود. تعداد ۶۵ ریسک فاکتور در سه یا کمتر از سه مطالعه استفاده شدند. اکثر مطالعات مورد بررسی از پایگاه داده برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی (SEER) که متعلق به کشور آمریکا می‌باشد، استفاده کرده‌اند. در هشت مطالعه مورد بررسی از پایگاه داده SEER استفاده شده است. یافته‌ها نشان می‌دهد که در مجموع از هشت تکنیک داده‌کاوی در این ۲۰ مطالعه استفاده شده است.

این هشت تکنیک شامل رگرسیون خطی (Linear regression)، ماشین بردار پشتیبان (support vector machine)، جنگل تصادفی (Random forest)، درخت تصمیم (decision tree)، ادابوست (AdaBoost)، بگینگ (Bagging)، K نزدیکترین همسایه (k-nearest neighbors, KNN)، دسته‌بند بیزین ساده (Naive Bayes Classifiers)، برای پیش‌بینی احتمال بقا استفاده شده است. همچنین چهار معیار Sensitivity، Accuracy، Specificity و AUC به‌عنوان پرکاربردترین معیارهای ارزیابی برای پیش‌بینی احتمال بقا به‌کار برده شده‌اند.

Delen و همکاران از سه تکنیک درخت تصمیم، شبکه‌های عصبی، رگرسیون خطی برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. این تکنیک‌ها بر روی داده‌هایی که در بازه‌ی زمانی سال‌های ۱۹۷۳ تا ۲۰۰۰ از پایگاه داده‌ی جهانی برنامه‌ی

جدول ۱: ریسک فاکتورهای مهم به همراه تعداد دفعات و درصد فراوانی

ریسک فاکتور	تعداد دفعات	درصد فراوانی
سن	۱۷	٪۸۵
گرید	۱۵	٪۷۵
استیج	۱۵	٪۷۵
اندازه تومور	۱۴	٪۷۰
وضعیت تاهل	۹	٪۴۵
ژنتیک	۹	٪۴۵
پرتودرمانی	۹	٪۴۵
نژاد	۸	٪۴۰
وضعیت گیرنده استروژن	۷	٪۳۵
محل اصلی	۸	٪۴۰
وضعیت گیرنده پروژسترون	۷	٪۳۵
درگیر شدن غدد لنفاوی	۵	٪۲۵
عمل جراحی	۶	٪۳۰
تعداد لنف نودهای مثبت	۵	٪۲۵
شیمی درمانی	۵	٪۲۵
رادیوتراپی	۵	٪۲۵

نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود، پیاده‌سازی شدند. داده‌های جمع‌آوری شده شامل ۴۳۳/۲۷۲ رکورد و ۷۲ متغیر است. نتایج نشان دهنده‌ی عملکرد بهتر رکورد و ۷۲ متغیر است. نتایج نشان دهنده‌ی عملکرد بهتر تکنیک درخت تصمیم نسبت به بقیه‌ی تکنیک‌های موجود بود.[۱۸]

Thongkam و همکاران اثر عمل پیش‌پردازشی برروی تکنیک‌های داده‌کاوی جهت پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان را بررسی کردند. پژوهشگران سعی در بهبود داده‌های قبل از پردازش توسط تکنیک‌های داده‌کاوی داشتند. در واقع با انجام یک عمل پیش‌پردازشی به دنبال مقابله با عدم توازن در داده‌ها بودند. این عمل پیش‌پردازشی شامل تشخیص داده‌های پرت و تکنیک‌های بیش نمونه‌برداری بود و سپس چهار تکنیک ادابوست، بگینگ، درخت تصمیم C4.5 و ماشین بردار پشتیبان برای تعیین میزان بقا در بیماران مبتلا به سرطان پستان استفاده شد. داده‌ها از یک بیمارستان در تایلند از سال ۱۹۹۰ تا سال ۲۰۰۱ جمع‌آوری شده و شامل ۳۳۷۸ بیمار است. نتایج نشان می‌دهد که پیش‌پردازش اعمال شده باعث بهبود عمکرد تکنیک‌های داده‌کاوی شده است.[۲۱]

Ture و همکاران تکنیک‌های مختلف مبتنی بر درخت تصمیم را برای تعیین میزان بقا در بیماران مبتلا به سرطان پستان بکار بردند. این تکنیک‌ها شامل C4.5، QUEST، CHAID، C&R، و ID3 بودند. داده‌ها مربوط به ۵۰۰ بیمار در بین سال‌های ۱۹۹۷ تا ۲۰۰۶ بودند. نتایج نشان‌دهنده‌ی عملکرد بهتر درخت تصمیم C4.5 نسبت به بقیه‌ی تکنیک‌ها بود.[۲۲]

Wang و همکاران از تکنیک‌های درخت تصمیم، K نزدیکترین همسایه و رگرسیون خطی برای پیش‌بینی میزان بقا ۵ ساله در بیماران مبتلا به سرطان پستان استفاده کردند. همچنین در این مدل از تئوری مورچگان Actor–network theory, (ANT) جهت انتخاب صفات کلیدی استفاده شده است. علاوه بر این، تکنیک‌های بیش نمونه‌برداری بیش نمونه‌سازی ترکیبی اقلیت استفاده شده است. این تکنیک‌ها برروی داده‌هایی که در بازه‌ی زمانی بین سال‌های ۱۹۷۳ تا ۲۰۰۷ از پایگاه داده‌ی جهانی برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود، پیاده‌سازی شدند. این مجموعه شامل ۹۷۳۱۲۵ رکورد و ۱۱۸ متغیر بود. نتایج نشان‌دهنده‌ی این بود که ترکیب تئوری مورچگان به‌همراه درخت تصمیم C5 و تکنیک بیش نمونه‌برداری بیش نمونه‌سازی ترکیبی اقلیت عملکرد بهتری خواهند داشت.[۲۳]

Kim و همکاران از تکنیک‌های نیمه اتوماتیک، شبکه‌ی عصبی و ماشین بردار پشتیبان برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. با توجه به اینکه جمع‌آوری داده‌های بدون برچسب راحت‌تر است، پیشنهاد استفاده از تکنیک‌های نیمه اتوماتیک برای تعیین میزان بقا در بیماران مبتلا به سرطان پستان ارایه شد. این تکنیک برروی داده‌هایی که در بازه‌ی زمانی بین سال‌های ۱۹۷۳ تا ۲۰۰۳ از پایگاه داده‌ی جهانی برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود، پیاده‌سازی شدند. داده‌های جمع‌آوری شده شامل ۵۰۵۳۶۷ و ۸۶ متغیر است. نتایج نشان‌دهنده ی عملکرد مناسب این تکنیک نسبت به شبکه‌ی عصبی و ماشین بردار پشتیبان بود.[۲۴]

Park و همکاران از تکنیک‌های شبکه‌ی عصبی، ماشین بردار پشتیبان و دسته‌بندی نیمه‌اتوماتیک برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. این تکنیک‌ها بر روی داده‌هایی که در بازه‌ی زمانی سال‌های ۱۹۷۳ تا ۲۰۰۳ از پایگاه داده‌ی جهانی برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود، پیاده‌سازی شدند. داده‌های جمع‌آوری شده شامل ۵۰۵۳۶۷ و ۸۶ متغیر است. بررسی‌ها نشان داد که دو تکنیک شبکه‌ی عصبی و ماشین بردار پشتیبان وابسته به مقادیر مختلف پارامترها هستند، این در حالی است که تکنیک دسته‌بندی نیمه اتوماتیک وابستگی به مقدار پارامترها نداشته و عملکرد نسبتا قابل قبولی داشته است [۲۵]

Wang و همکاران در مطالعه‌ی خود از تکنیک‌های پایه درخت تصمیم و رگرسیون خطی و ترکیب آن‌ها با عملیات دیگر برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. این تکنیک‌ها بر روی نزدیک به یک میلیون داده که در بازه‌ی زمانی بین سال‌های ۱۹۷۳ تا ۲۰۰۷ از پایگاه داده‌ی جهانی برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود، پیاده‌سازی شدند. نتایج نشان‌دهنده‌ی عملکرد بهتر تکنیک لجستیک رگرسیون به‌عنوان تکنیک پایه و ترکیب آن با دیگر تکنیک‌ها بود.[۲۶]

Wang و همکاران عملکرد تکنیک شبکه‌ی عصبی برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان را بررسی کردند. این تکنیک بر روی داده‌های ۵۱۹۰ بیمار از یک بیمارستان در تایوان در بازه‌ی زمانی بین سال‌های ۲۰۰۰ تا ۲۰۰۳ پیاده‌سازی شد و عملکرد خوبی از خود به‌جای گذاشت.[۲۷]

Chao و همکاران از تکنیک‌های پرکاربرد ماشین بردار پشتیبان، رگرسیون خطی و درخت تصمیم برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. این تکنیک‌ها بر روی داده‌هایی که در بازه‌ی زمانی بین سال‌های ۲۰۰۲ تا ۲۰۱۰ از یک بیمارستان در تایوان جمع‌آوری شده بود، پیاده‌سازی شدند. داده‌های جمع‌آوری شده شامل ۱۷۲۱ رکورد و هشت متغیر است. نتایج نشان‌دهنده‌ی عملکرد بهتر ماشین بردار پشتیبان نسبت به دو تکنیک دیگر در این مجموعه‌ی داده‌ای به نسبت کوچک بود.[۲۸]

Afshar و همکاران از تکنیک‌های بیزین، ماشین بردار پشتیبان، تشخیص تعاملی خودکار Chi-square automatic interaction detection, (CHAID) برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. این تکنیک‌ها بر روی داده‌هایی که در بازه‌ی زمانی بین سال‌های ۱۹۹۹ تا ۲۰۰۴ از پایگاه داده‌ی جهانی برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود،

پیاده‌سازی شدند. داده‌های جمع‌آوری شده شامل ۲۲۷۶۳ رکورد و ۶۷ متغیر بود. نتایج نشان‌دهنده‌ی دقت بیشتر روش ماشین بردار پشتیبان نسبت به دیگر تکنیک‌ها بود.[۲۹]

Zolbanian و همکاران از مدل‌های جنگل تصادفی درختان، درخت تصمیم، ماشین بردار پشتیبان و شبکه‌ی عصبی برای پیش‌بینی میزان بقا در مبتلایان به سرطان پستان استفاده کردند. مجموعه‌ی داده‌ای مورد استفاده از پایگاه SEER شامل تقریبا یک میلیون و پانصد رکورد در یک بازه‌ی زمانی تقریبا ۴۰ ساله بوده که یک مجموعه‌ی داده‌ای تقریبا بزرگ است. نتایج نشان می‌دهد که جنگل تصادفی درختان عملکرد بهتری نسبت به بقیه‌ی تکنیک‌ها در این مجموعه داده‌ای داشته است.

Garsia و همکاران در مطالعه‌ی خود از تکنیک‌های K نزدیک‌ترین همسایه، درخت تصمیم، ماشین بردار پشتیبان و درختی‌های کلاسه‌بندی برای پیش‌بینی میزان بقا در مبتلایان به سرطان پستان استفاده کردند. مجموعه داده‌ای مورد استفاده شامل ۳۹۹ رکورد بیمار زن و ۱۶ متغیر از یک مرکز ثبت سرطان در کشور پرتغال است. نتایج نشان‌دهنده‌ی عملکرد بهتر K نزدیکترین همسایه نسبت به دیگر تکنیک‌ها در این مجموعه داده‌ای کوچک بود.[۳۱]

Montazeri و همکاران در مطالعه‌ی خود از مدل‌های بیزین ساده، جنگل تصادفی درختان، ادراک چند سطحی، ادابوست، ماشین بردار پشتیبان و K نزدیکترین همسایه برای پیش‌بینی میزان بقا در مبتلایان به سرطان پستان استفاده کردند. نتایج نشان می‌دهد که جنگل تصادفی درختان عملکرد بهتری نسبت به بقیه‌ی تکنیک‌ها داشته است. مجموعه داده‌ای مورد استفاده شامل ۸۷۶ زن و ۲۴ مرد و هشت صفت است.[۳۲]

Kate و همکاران در مطالعه‌ی خود از تکنیک‌های الگوریتم CART از تکنیک‌های درخت تصمیم، بیزین ساده و لجستیک رگرسیون برای پیش‌بینی میزان بقا در مبتلایان به سرطان پستان استفاده کردند. مجموعه‌ی داده‌ای مربوط به پایگاه SEER در بازه‌ی زمانی بین سال‌های ۱۹۷۳ تا ۲۰۰۳ شامل ۱۶ متغیر بود. نتایج نشان داد درخت تصمیم نسبت به دو تکنیک دیگر عملکرد بهتری داشته است.[۳۳]

Khalkhali و همکاران در مطالعه‌ی خود از تکنیک‌های الگوریتم CART از تکنیک‌های مبتنی بر درخت تصمیم برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. مجموعه داده‌ای مورد استفاده در پژوهش براساس داده‌های کوهورت جمع‌آوری شده از مرکز تحقیقاتی و رفتاری امید که از سال‌های ۲۰۰۷ تا ۲۰۱۰ را دربرمی‌گیرد، بود. این مجموعه شامل ۵۶۹ بیمار بوده و از این مجموعه، نه فاکتور ریسک انتخاب شد. نتایج نشان‌دهنده‌ی دقت بالای روش پیشنهادی ایشان بود.[۳۴]

Varlamis و همکاران از تکنیک‌های الگوریتم CART از تکنیک‌های درخت تصمیم، جنگل تصادفی درختان، ماشین بردار پشتیبان، بیزین ساده و رگرسیون خطی برای پیش‌بینی میزان بقا در بیماران مبتلا به سرطان پستان استفاده کردند. مجموعه‌ی داده‌ای مربوط به یک مرکز ثبت سرطان در یونان در بازه‌ی زمانی بین سال‌های ۱۹۹۸ تا ۲۰۰۴ بود. نتایج نشان‌دهنده‌ی عملکرد بهتر ماشین بردار پشتیبان و لجستیک رگرسیون نسبت به سه تکنیک دیگر بود [۳۵]

Tapak و همکاران از تکنیک‌های شبکه بیزین، درخت تصمیم و رگرسیون لجستیک برای پیش‌بینی میزان بقا پنج ساله در مبتلایان به سرطان پستان استفاده کردند. همچنین در این مدل از ترکیب تئوری مورچگان و انتخاب صفت مبتنی بر همبستگی جهت انتخاب صفات کلیدی استفاده شده است. علاوه بر این تکنیک‌های بیش نمونه‌برداری نوسانگر ذخیره سازی دیجیتال و بیش نمونه‌سازی ترکیبی اقلیت حاشیه‌ای استفاده شده است. این تکنیک‌ها بر روی داده‌هایی که در بازه‌ی زمانی بین سال‌های ۲۰۰۴ تا ۲۰۰۷ از پایگاه داده‌ی جهانی برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی در کشور آمریکا جمع‌آوری شده بود، پیاده‌سازی شد. این مجموعه شامل ۲۷۰۹۸۹ رکورد و ۱۵۱ متغیر انتخاب شد. نتایج نشان‌دهنده‌ی این بود که عملکرد بهتر ترکیب تئوری مورچگان و انتخاب صفت مبتنی بر همبستگی و درخت تصمیم C4.5 به همراه تکنیک بیش نمونه‌برداری نوسانگر ذخیره‌سازی دیجیتال عملکرد بهتری خواهند داشت.[۳۶]

Zhao و همکاران از تکنیک‌های جنگل تصادفی درختان، ماشین بردار پشتیبان، شبکه‌ی عصبی و تقویت گرادیان برای پیش‌بینی میزان بقا در مبتلایان به سرطان پستان استفاده کردند. مجموعه داده‌ای مستخرج از یک مطالعه‌ی کوهورت شامل ۱۸۷۴ بیمار با میانگین سنی ۶۲ سال بود. نتایج نشان‌دهنده‌ی عملکرد بهتر ماشین بردار پشتیبان و جنگل تصادفی درختان نسبت به بقیه‌ی تکنیک‌ها بود.[۳۷]

Momenyan و همکاران از تکنیک‌های مبتنی بر درخت تصمیم (C5، QUEST، CHAID، C&RT) و تکنیک لجستیک رگرسیون

استفاده کردند. مجموعه داده‌ای شامل ۵۰۰ بیمار از مرکز تحقیقات دانشگاه شهید بهشتی بود. نتایج نشان‌دهنده‌ی عملکرد بهتر تکنیک‌های QUEST و C5 و لجستیک رگرسیون نسبت به دیگر تکنیک‌ها است.[۳۸]

Nourelahi و همکاران از عملکرد تکنیک رگرسیون خطی استفاده کردند. مجموعه‌ی داده‌ای شامل ۵۶۷۳ بیمار و ۱۶ متغیر از مرکز تحقیقات شیراز بود. نتایج نشان‌دهنده‌ی عملکرد مناسب این تکنیک بود.[۳۹]در بیشتر مطالعات، ریسک فاکتورهای سن، گرید، استیج و اندازه تومور به ترتیب در ۱۷، ۱۵، ۱۵ و ۱۴ مطالعه بکار گرفته شدند (جدول ۱). با توجه به برابری تعداد دفعات استفاده از ریسک فاکتورهای گرید، استیج و اندازه تومور می‌توان گفت برای این‌که بتوان برداشت درستی از میزان بقای بیماران مبتلا به سرطان پستان به‌دست آورد، این مجموعه ریسک فاکتورها بایستی حتما مورد استفاده قرار گیرد.

کشورهای مختلف قاره‌های آسیایی، اروپایی و آمریکای جنوبی از پایگاه داده برنامه‌ی نظارت، اپیدمیولوژی و نتایج نهایی استفاده کرده‌اند. پس از این پایگاه، نظام ثبت داخلی در مطالعات مربوط به کشورهای ایران، یونان و چین مورد استفاده قرار گرفته است. مطالعات مربوط به جنوب شرق آسیا از جمله تایوان و چین بیشتر از سیستم اطلاعات بیمارستانی برای جمع‌آوری داده‌های مورد نظر استفاده کرده‌اند.

سه تکنیک لجستیک رگرسیون، درخت تصمیم و ماشین بردار پشتیبان به ترتیب در۱۲، ۱۱ و ۱۰ مطالعه مورد استفاده قرار گرفته‌اند. دلیل استفاده این سه تکنیک، بنیادی بودن و صحت بالاتر نتایج و قابلیت اعتماد بیشتر در مطالعات انجام شده بوده است. دو تکنیک بیزین ساده و یادگیری نیمه اتوماتیک نسبت به سایر تکنیک‌ها کمتر مورد استفاده قرار گرفته‌اند (سه یا کمتر از سه مطالعه) (جدول ۲).

یافته‌ها نشان می‌دهد که چهار معیار صحت، حساسیت، ویژگی، سطح زیرنمودار ROC به‌عنوان معیارهای ارزیابی استفاده شده‌اند که از بین این معیارها، صحت بیشترین فراوانی را در تکنیک‌های مختلف دارد. مقدار یکسان حساسیت و ویژگی به این دلیل است که این دو معیار همزمان در مقالات استفاده شده و به نوعی مکمل همدیگر هستند. این معیارها در مطالعات داده‌کاوی به‌عنوان معیارهای پایه و بنیادی محسوب می‌شوند.

از میان ۵۲۷ مطالعه، در نهایت ۲۰ مطالعه مورد استفاده قرار گرفت. یکی از چالش‌های محققین در انجام این پژوهش پراکنده بودن یافته‌های مرتبط با پژوهش در مطالعات مختلف بود. بدین معنا که هر مطالعه تنها بخشی از متغیرهای مد نظر محقیقین را پوشش می‌داد، این مساله باعث پیچیده شدن فرآیند جستجو و بازیابی مطالعات شده بود.

محققین از طریق بکارگیری دامنه وسیعی از واژه‌های مرتبط و مترادف‌های آنها، همچنین استفاده از عملگرهای بولی و ترانکیشن (Boolean and transaction operators) تا حدود زیادی این مشکل را برطرف کردند.

عدم دسترسی به متن کامل برخی از مطالعات، یکی دیگر از چالشهای پیش روی محققین بود که این مورد نیز با بهره‌گیری از نظرات و خدمات تخصصی متخصصین کتابداری و اطلاع‌رسانی پزشکی کتابخانه مرکزی دانشگاه علوم پزشکی اصفهان به حداقل رسید.

جدول ۲: تکنیک‌ها به همراه تعداد دفعات، درصد فراوانی و تعداد دفعات استفاده از هر یک از معیارها

تکنیک	تعداد دفعات	درصد فراوانی	حساسیت	ویژگی	صحت	سطح زیر نمودار ROC
لجستیک رگرسیون	۱۲	٪۶۰	۸	۸	۹	۶
درخت تصمیم	۱۱	٪۵۵	۹	۹	۹	۵
ماشین بردار پشتیبان	۱۰	٪۵۰	۵	۵	۸	۶
جنگل تصادفی	۵	٪۲۵	۳	۳	۴	۳
شبکه عصبی	۵	٪۲۵	۳	۳	۵	۳
ادابوست	۴	٪۲۰	۳	۳	۳	۳

پیشنهاد می‌شود سایر محققین برای مطالعات آتی خود درباره پیش‌بینی احتمال بقای بیماران مبتلا، به سرطان‌هایی متمرکز شوند که کمتر مورد توجه قرار گرفته‌اند. از جمله سرطان غدد لنفاوی، سرطان تخمدان، گردن رحم و غیره. همچنین می‌توان درباره پیش‌بینی عود مجدد یا احتمال ابتلای افراد به انواع مختلف سرطان‌ها و به‌طور کلی انواع بیماری‌های مختلف با استفاده از تکنیک‌های داده‌کاوی نیز متمرکز شد.

توصیه می‌شود پژوهشگران برای فرموله کردن مطالعه خود از نتایج مطالعه فعلی ما برای انتخاب تکنیک‌های مناسب داده‌کاوی، ریسک فاکتورهای مناسب و منابع داده‌ای مطلوب بهره‌برداری کنند. طبق بررسی‌های انجام شده در پژوهش‌های داخلی و خارجی، غالبا روش‌های دسته‌بندی به‌صورت مجزا بررسی شده‌اند و کمتر ترکیب آن‌ها با دیگر تکنیک‌ها مورد نظر قرار گرفته است. همچنین تاثیر نوع مجموعه‌های داده‌ای و میزان توازن و عدم توازن داده‌ها کمتر در نظر گرفته شده است. این مسئله با توجه به مطالب فوق قابل توجیه است.

نتایج نشان داد که ۴۵٪ مطالعات از منبع SEER (Surveillance, Epidemiology, and End Results) استفاده کرده‌اند. SEER یکی از معتبرترین منابع در خصوص آمارهای مربوط به سرطان و به‌طور خاص مرگ و میر ناشی از آن در آمریکا است که غالبا مورد استناد قرار می‌گیرد.[۴۰و۴۲] گردآوری و انتشار میزان شیوع و بروز سرطان و میزان بقا مبتلایان به سرطان با پوشش ۳۶٪ جامعه آمریکا و تجمیع داده‌ها از بیش از ۲۰ نظام ثبت جمعیتی سرطان از ویژگی‌های برجسته به‌کارگیری SEER می‌باشد.[۴۳]

Li به حجم جمعیتی پوششی توسط این مجموعه‌ی داده‌ای که در حدود ۹۳٪ در بین سال‌های ۲۰۰۱ تا ۲۰۰۷ است، اشاره کرد.[۴۴] مطالعات مختلف، دلایل متعددی از قبیل کیفیت بالا، غربالگری نوین، پشتیبانی از نژادهای مختلف و استفاده از ابزار پردازشی قوی برای استفاده از این منبع داده ارایه کرده‌اند.[۴۵و۵۲]

روش‌های داده‌کاوی برای داده با حجم بالا مناسب‌تر هستند. همچنین تقریب و نمونه‌سازی بهتری دارند.[۵۲] بنابراین می‌توان گفت که با توجه به این‌که SEER یک مجموعه‌ی داده‌ای بزرگ با پوشش جمعیتی بالا بوده و دامنه‌ی گسترده‌ای از صفات را نیز پوشش می‌دهد و با توجه به نظر آگراوال (Agrawal)، برای داده‌کاوی گزینه‌ی مناسبی است.

نتایج مطالعه نشان داد ۲۹٪ مطالعات از نظام ثبت داخلی استفاده کرده‌اند. از مشکلات نظام ثبت داخلی، کمبود نظام ثبت بیماری است. همچنین، جمع‌آوری داده کاری بسیار زمان‌بر است. در مقابل SEER به‌راحتی قابل دسترس و رایگان است. از طرفی، نظام ثبت داخلی، داده‌های مرتبط با کشور محقق را در نظر می‌گیرد در حالی‌که SEER جهانی است. حجم داده در نظام ثبت داخلی کمتر است بنابراین ممکن است نتایج تولید شده توسط آن قابل تعمیم به کل نباشد، در حالی‌که SEER یک مجموعه‌ی داده‌ای بزرگ و با پوشش بالا است. البته این بالابودن حجم داده می‌تواند باعث پیچیدگی پیاده‌سازی این مجموعه‌ی داده‌ای شود.

طبق بررسی‌هایی که در این تحقیق انجام شده است، شش معیار Age، Grade، Stage، Tumor Size، Marital Status و Histology بیشترین کاربرد را به‌عنوان ریسک فاکتور در مقالات داشته‌اند. در میان این معیارها، Age بیشترین تکرار را داشته است. یکی از مهمترین ریسک فاکتورها برای پیش‌بینی احتمال بقای بیمار، سن بیمار می‌باشد.[۵۲و۵۳] میزان شیوع سرطان پستان در سن زیر ۳۰ سال بسیار پایین می‌باشد. پس از ۳۰ سال تا سن ۸۰ سالگی این میزان افزایش می‌یابد. سن عامل تاثیرگذاری در بقاء بیمار می‌باشد.

Singletary در مطالعه خود سن ابتلا به سرطان، مرحله بروز سرطان و میزان پیشرفت سرطان را سه عامل مهم در پیش‌بینی میزان بقا اعلام می‌کند.[۵۴] مطالعات مختلف، ریسک فاکتورهای متعددی را در ابتلا به سرطان پستان ذکر کرده‌اند. Momenimovahed در مطالعه خود فاکتورهای دموگرافیک، تولید مثل، هورمونی، وراثت، سبک زندگی و فاکتورهای مربوط به پستان را از جمله مهمترین فاکتورهای موثر در ابتلای فرد به سرطان پستان ذکر کرده است.[۵۵] ولی Bellaachia در مطالعه خود اشاره دارد که فاکتورهای موثر در تشخیص میزان بقا بیماران مبتلا به سرطان اندازه تومور، سن در هنگام تشخیص تومور، اندازه تومور و مرحله رشد تومور می‌باشند.[۵۶]

به‌نظر می‌رسد انتخاب صحیح فاکتورهای تاثیرگذار در پیش‌بینی میزان بقاء بیماران نقش برجسته و مهمی دارد. در مطالعات مختلف از روش‌های متفاوتی برای انتخاب این فاکتورها استفاده کرده‌اند.

در مطالعاتی که از SEER در مطالعه خود بهره برده‌اند عمدتا تمامی فاکتورهایی که اطلاعات آن در این پایگاه داده وجود دارد را انتخاب کرده‌اند.

در برخی مطالعات دیگر، از نظر متخصصین سرطان برای انتخاب ریسک فاکتورها بهره برده‌اند. پیشنهاد می‌شود که ترکیبی از نظر متخصصین سرطان و تکنیک‌های علمی برای شناسایی و انتخاب ریسک فاکتورهای مورد نیاز برای پیش‌بینی احتمال بقا استفاده شوند.

طبق بررسی‌هایی که در این تحقیق انجام شده است، سه تکنیک رگراسیون لجستیک، درخت تصمیم و ماشین بردار پشتیبان بیشترین تکرار را داشته‌اند. تکنیک Support Vector Machine (SVM) کارایی مناسب خود را در بسیاری از تکنیک‌های تشخیص الگو نشان داده است.[۵۷]

این تکنیک توانایی مناسبی در تعمیم دادن داده‌های تست دیده نشده دارد. بنابراین می‌تواند در حوزه‌ی تشخیص میزان بقا که داده‌های تست دیده نشده اهمیت دارند، عملکرد مناسبی داشته باشد.[۵۸] درخت تصمیم یک روش استاندارد و به آسانی قابل فهم است. همین سادگی روش منجر به کاربرد روزافزون آن در حوزه‌ی سرطان پستان شده است که با عملکرد مناسب همراه بوده است.[۵۹] تعامل با دیگر الگوریتم‌های خوشه‌بندی از خصوصیات دیگر درخت تصمیم است که آن‌را نسبت به دیگر روش‌ها متمایز می‌کند.[۶۰] رگرسیون خطی از تکنیک‌های آماری و داده‌کاوی مهم در حوزه‌ی سرطان پستان است که بیشترین کاربرد را با توجه به بنیادی بودن روش در حوزه‌ی تشخیص میزان بقا در سرطان پستان داشته است.[۶۱]

لازم به ذکر است که از بین تکنیک‌های فوق، ماشین بردار پشتیبان که پیچیدگی بیشتری نسبت به دو تکنیک دیگر دارد، عملکرد بهتری نیز نسبت به آنها داشته است. طبق بررسی‌ها، دو معیار حساسیت و ویژگی همیشه با هم بکار برده شده‌اند. در بین معیارها، صحت بیشترین تکرار را داشته است و پس از آن دو معیار حساسیت و ویژگی بیشترین تکرار را داشته اند. این داده های پزشکی هیچ ساختار رسمی ندارند همین امر داده‌کاوی را دچار چالش در این حوزه می‌کند و معیارهای مناسبی باید در این زمینه انتخاب شوند.[۶۲] این معیارها می‌توانند پیش‌بینی میزان بقا را به خوبی نشان دهند.[۵۳]

معیار حساسیت متناظر با معیار بازیابی است که از معیارهای بنیادی در داده‌کاوی است.[۶۳] یکی از رایج‌ترین معیارهای ارزیابی کیفیت در داده‌کاوی، صحت است.[۶۴] تطابق معیارهای فوق‌الذکر با خصوصیات داده‌های پزشکی که ناهمگن بودن، یکی از آن‌ها است، از دلایل دیگر انتخاب این معیارها به‌عنوان معیارهای ارزیابی در این قسمت است.[۵۲]

نتیجه‌گیری: استفاده همزمان از چند تکنیک داده‌کاوی می‌تواند نتیجه بهتری داشته باشد. تکنیک‌های درخت تصمیم، رگرسیون لجستیک و ماشین بردار پشتیبان جز تکنیک‌های بنیادی داده‌کاوی هستند. این تکنیک‌ها در حوزه‌ی سلامت به نسبت عملکرد بهتری به‌عنوان تکنیک پایه داشته‌اند و بقیه‌ی تکنیک‌ها معمولا در کنار آن‌ها استفاده می‌شوند. استفاده از این تکنیک‌ها می‌تواند زمینه مناسبی برای متخصصین بالینی جهت ارزیابی اثربخشی روش‌های مختلف درمانی و تاثیرگذاری هر یک از این شیوه‌ها بر طول عمر بیماران و میزان بقای آن‌ها فراهم آورد.

چنانچه از خروجی این تکنیک‌ها برای تامین ورودی داده‌های مورد نیاز یک سیستم پشتیبان تصمیم استفاده شود متخصصان بالینی می‌توانند در زمان ارایه خدمت به بیماران مبتلا به سرطان سینه از طریق وارد کردن مقادیر ریسک فاکتورهای مرتبط با بیمار، سن و وضعیت فیزیکی وی از طریق خروجی‌هایی که سیستم پشتیبان تصمیم ارایه می‌کند، بهینه‌ترین تصمیم برای انتخاب بهترین روش درمانی و به‌تبع آن افزایش میزان بقای بیماران را فراهم آوردند.

از سوی دیگر متخصصین داده‌کاوی نیز می‌توانند از طریق مرور نتایج این مطالعه و کسب اطلاع از ملزومات بهینه برای بکارگیری تکنیک‌های داده‌کاوی، ایده‌آل‌ترین تکنیک و ریسک فاکتورهای متناسب با آن را جهت پیش‌بینی احتمال بقای بیماران مبتلا به سرطان پستان انتخاب کنند.

References

1. DeVita Jr VT, Lawrence T, Rosenberg SA. Cancer: principles & practice of oncology: annual advances in oncology: Lippincott Williams & Wilkins; 2012.
2. Pilevarzadeh M, Amirshahi M, Afsargharehbagh R, Rafiemanesh H, Hashemi S-M, Balouchi A. Global prevalence of depression among breast cancer patients: a systematic review and meta-analysis. *Breast Cancer Res Treat* 2019;176(3):519-33.
3. Lindsay WD, Ahern CA, Tobias JS, Berlind CG, Chinniah C, Gabriel PE, et al. Automated data extraction and ensemble

methods for predictive modeling of breast cancer outcomes after radiation therapy. *Med Phys* 2019;46(2):1054-63.
4. Momenyan S, Baghestani AR, Momenyan N, Naseri P, Akbari ME. Survival prediction of patients with breast cancer: comparisons of decision tree and logistic regression analysis. *Int J Cancer Manag* 2018;11(7).
5. Matini AH, Jafarian-Dehkordi N, Bahmani B, Sharifi M, Jahantigh D, Mazoochi T. Association of ApaI and TaqI polymorphisms in VDR Gene with Breast Cancer. *Asian Pac J Cancer Prev* 2020;21(9):2667-72.
6. Sabouri S, Esmaily H, Shahid Sales S, Emadi M. Determining related factors to survival of colorectal cancer patients using cox regression. *Med J Mashhad Univ Med Sci* 2018;61(4):1083-92.
7. Naghitorabi M, Asl JM, Sadeghi HMM, Rabbani M, Jafarian-Dehkordi A, Javanmard HS. Quantitative evaluation of DNMT3B promoter methylation in breast cancer patients using differential high resolution melting analysis. *Res Pharm Sci* 2013;8(3):167.
8. Abachizadeh K, Moradi-Kouchi A, Ghanbari-Motlagh A, Kousha A, Shekarriz-Foumani R, Erfani A. Breast cancer in Iran: Levels, variations and correlates. *Community Health (Salamat ijtimai)* 2018;5(1):11-21.
9. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* 2018;68(6):394-424.
10. Organization WH. Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018. *International Agency for Research on Cancer Geneva: World Health Organization* 2018.
11. Haghighat S. Evaluation of survival and related factors in breast cancer patients referred to the Breast Disease Center 2013;6:28-36.
12. Sabouri S, Esmaily H, Shahid Sales S, Emadi M. Determining related factors to survival of colorectal cancer patients using cox regression. *Med J Mashhad Univ Med Sci* 2018;61(4):1083-92.
13. Forouzanfar MH, Foreman KJ, Delossantos AM, Lozano R, Lopez AD, Murray CJ, et al. Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. *Lancet* 2011;378(9801):1461-84.
14. Sehhati MR, Dehnavi AM, Rabbani H, Javanmard SH. Using protein interaction database and support vector machines to improve gene signatures for prediction of breast cancer recurrence. *J Med Signals Sens* 2013;3(2):87.
15. Goli S, Mahjub H, Faradmal J, Mashayekhi H, Soltanian A-R. Survival prediction and feature selection in patients with breast cancer using support vector regression. *Comput Math Methods Med* 2016;2016.
16. Sattari M, Bagherian H. Application of data mining techniques in predicting the probability of survival of breast cancer patients, Second Conference on Information Technology and Health Promotion Centered on Cancer. 2019.
17. Van Belle V, Pelckmans K, Van Huffel S, Suykens JA. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artif Intell Med* 2011;53(2):107-18.
18. 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.
19. Liu Y-Q, Wang C, Zhang L, editors. Decision tree based predictive models for breast cancer survivability on imbalanced data. 2009 3rd international conference on bioinformatics and biomedical engineering; 2009: IEEE.
20. Moher D, Altman DG, Liberati A, Tetzlaff J. PRISMA Statement. *Epidemiology* 2011;22(1):128.
21. Thongkam J, Xu G, Zhang Y, Huang F. Toward breast cancer survivability prediction models through improving training space. *Expert Syst Appl* 2009;36(10):12200-9.
22. Ture M, Tokatli F, Kurt I. Using Kaplan–Meier analysis together with decision tree methods (C&RT, CHAID, QUEST, C4. 5 and ID3) in determining recurrence-free survival of breast cancer patients. *Expert Syst Appl* 2009;36(2):2017-26.
23. Wang K-J, Makond B, Chen K-H, Wang K-M. A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. *Appl Soft Comput* 2014;20:15-24.
24. Kim J, Shin H. Breast cancer survivability prediction using labeled, unlabeled, and pseudo-labeled patient data. *J Am Med Inform Assoc* 2013;20(4):613-8.
25. Park K, Ali A, Kim D, An Y, Kim M, Shin H. Robust predictive model for evaluating breast cancer survivability. *Eng Appl Artif Intell* 2013;26(9):2194-205.
26. Wang K-J, Makond B, Wang K-M. An improved survivability prognosis of breast cancer by using sampling and feature selection technique to solve imbalanced patient classification data. *BMC Med Inform Decis Mak* 2013;13(1):1-14.
27. Wang T-N, Cheng C-H, Chiu H-W, editors. Predicting post-treatment survivability of patients with breast cancer using Artificial Neural Network methods. 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2013:1290-3.
28. Chao C-M, Yu Y-W, Cheng B-W, Kuo Y-L. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *J Med Syst* 2014;38(10):1-7.
29. Afshar HL, Ahmadi M, Roudbari M, Sadoughi F. Prediction of breast cancer survival through knowledge discovery in databases. *Glob J Health Sci* 2015;7(4):392.
30. Zolbanin HM, Delen D, Zadeh AH. Predicting overall survivability in comorbidity of cancers: A data mining approach. *Decis Support Syst* 2015;74:150-61.
31. García-Laencina PJ, Abreu PH, Abreu MH, Afonoso N. Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. *Comput Biol Med* 2015;59:125-33.
32. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. *Technol Health Care* 2016;24(1):31-42
33. Kate RJ, Nadig R. Stage-specific predictive models for breast cancer survivability. *Int J Med Inform* 2017;97:304-11.
34. Khalkhali HR, Afshar HL, Esnaashar O, Jabbari N. Applying data mining techniques to extract hidden patterns about breast cancer survival in an Iranian cohort study. *J Res Health Sci* 2016;16(1):31.
35. Varlamis I, Apostolakis I, Sifaki-Pistolla D, Dey N, Georgoulias V, Lionis C. Application of data mining techniques and data analysis methods to measure cancer morbidity and mortality data in a regional cancer registry: The case of the island of Crete, Greece. *Comput Methods Programs Biomed* 2017;145:73-83.

36. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. *Clin Epidemiol Glob Health* 2019;7(3):293-9.
37. Zhao M, Tang Y, Kim H, Hasegawa K. Machine learning with k-means dimensional reduction for predicting survival outcomes in patients with breast cancer. *Cancer Inform* 2018;17:1176935118810215.
38. Momenyan S, Baghestani AR, Momenyan N, Naseri P, Akbari ME. Survival prediction of patients with breast cancer: comparisons of decision tree and logistic regression analysis. *Int J Cancer Manag* 2018;11(7):1-12.
39. Nourelahi M, Zamani A, Talei A, Tahmasebi S. A Model to Predict Breast Cancer Survivability Using Logistic Regression. *Middle East J Cancer* 2019;10(2):132-8.
40. Catalano PJ, Ayanian JZ, Weeks JC, Kahn KL, Landrum MB, Zaslavsky AM, et al. Representativeness of participants in the Cancer Care Outcomes Research and Surveillance (CanCORS) Consortium relative to the Surveillance, Epidemiology and End Results (SEER) program. *Med Care* 2013;51(2):e9.
41. Cheng L, Swartz MD, Zhao H, Kapadia AS, Lai D, Rowan PJ, Buchholz TA, Giordano SH. Hazard of recurrence among women after primary breast cancer treatment--a 10-year follow-up using data from SEER-Medicare. *Cancer Epidemiol Biomarkers Prev* 2012;21(5):800-9.
42. Jagsi R, Abrahamse P, Hawley ST, Graff JJ, Hamilton AS, Katz SJ. Underascertainment of radiotherapy receipt in Surveillance, Epidemiology, and End Results registry data. *Cancer* 2012;118(2):333-41.
43. List of SEER Registries About SEER. http://seer.cancer.gov/registries/list.html. Accessed 3 May 2012
44. Li J, Djenaba JA, Soman A, Rim SH, Master VA. Recent trends in prostate cancer incidence by age, cancer stage, and grade, the United States, 2001–2007. *Prostate Cancer* 2012;2012.
45. Hankey BF, Ries LA, Edwards BK. The surveillance, epidemiology, and end results program: a national resource. *Cancer Epidemiol Biomarkers Prev* 1999;8(12):1117-21.
46. Clegg LX, Li FP, Hankey BF, Chu K, Edwards BK. Cancer survival among US whites and minorities: a SEER (Surveillance, Epidemiology, and End Results) Program population-based study. *Arch Intern Med* 2002;162(17):1985-93.
47. Hayat MJ, Howlader N, Reichman ME, Edwards BK. Cancer statistics, trends, and multiple primary cancer analyses from the Surveillance, Epidemiology, and End Results (SEER) Program. *Oncologist* 2007;12(1):20-37.
48. Goss PE, Strasser-Weippl K, Lee-Bychkovsky BL, Fan L, Li J, Chavarri-Guerra Y, et al. Challenges to effective cancer control in China, India, and Russia. *Lancet Oncol* 2014;15(5):489-538.
49. Srivastava S, Smart C, Marciniak TA, Derrick L. Accessing NCI's SEER cancer data base with SeerQuery and CD-ROM. *Public Health Rep* 1992;107(4):397.
50. Anderson WF, Chatterjee N, Ershler WB, Brawley OW. Estrogen receptor breast cancer phenotypes in the Surveillance, Epidemiology, and End Results database. *Breast Cancer Res Treat* 2002;76(1):27-36.
51. Agrawal R, Psaila G, editors. Active Data Mining. KDD; 1995:3-8.
52. Delen D. Analysis of cancer data: a data mining approach. *Expert Syst* 2009;26(1):100-12.
53. Dubey AK, Gupta U, Jain S. Breast cancer statistics and prediction methodology: a systematic review and analysis. *Asian Pac J Cancer Prev* 2015;16(10):4237-45.
54. Singletary SE. Rating the risk factors for breast cancer. *Ann Surg* 2003;237(4):474.
55. Momenimovahed Z, Salehiniya H. Epidemiological characteristics of and risk factors for breast cancer in the world. *Breast Cancer: Targets Therapy* 2019;11:151.
56. Abdelghani B, Guven E, editors. Predicting breast cancer survivability using data mining techniques2006: Siam International Conference on Data Mining.2006;58(13):10-10.
57. Byun H, Lee S-W. A survey on pattern recognition applications of support vector machines. *Intern J Pattern Recognit Artif Intell* 2003;17(03):459-86.
58. Huang M-W, Chen C-W, Lin W-C, Ke S-W, Tsai C-F. SVM and SVM ensembles in breast cancer prediction. *PloS one* 2017;12(1):e0161501.
59. Venkatesan E, Velmurugan T. Performance analysis of decision tree algorithms for breast cancer classification. *Indian J Sci Technol* 2015;8(29):1-8.
60. Rohini K, Suseendran G. Aggregated K means clustering and decision tree algorithm for spirometry data. *Indian J Sci Technol* 2016;9(44):1-6.
61. Graja O, Azam M, Bouguila N, editors. Breast cancer diagnosis using quality control charts and logistic regression. 2018 9th International Symposium on Signal, Image, Video and Communications (ISIVC); 2018: IEEE.
62. Geng L, Hamilton HJ. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)* 2006;38(3):9-es.
63. Cios KJ, Moore GW. Uniqueness of medical data mining. *Artif Intell Med* 2002;26(1-2):1-24.
64. Maciejewski T, Stefanowski J, editors. Local neighbourhood extension of SMOTE for mining imbalanced data. 2011 IEEE symposium on computational intelligence and data mining (CIDM); 2011: IEEE.

TUMJ Tehran University Medical Journal, June 2021; Vol. 79, No. 3: 176-186 Review Article

Using data mining techniques for predicting the survival rate of breast cancer patients: a *review article*

Hossein Bagherian.Ph.D.[1]
Shaghayegh Haghjooy Javanmard.Ph.D.[2]
Mehran Sharifi.Ph.D.[3]
Mohammad Sattari.Ph.D.[1*]

1- Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.
2- Applied Physiology Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran.
3- Department of Hematology and Oncology, School of Medicine, Isfahan University of Medical Sciences, Isfahan, Iran.

*Corresponding author: Technology Research Center, Isfahan University of Health Information, Isfahan, Iran.
Tel: +98-21-37925152
E-mail: msattarimng.mui@gmail.com

Abstract

Received: 14 Feb. 2021 Revised: 21 Feb. 2021 Accepted: 15 May. 2021 Available online: 22 May. 2021

This review was conducted between December 2018 and March 2019 at Isfahan University of Medical Sciences. A review of various studies revealed what data mining techniques to predict the probability of survival, what risk factors for these predictions, what criteria for evaluating data mining techniques, and finally what data sources for it have been used to predict the survival of breast cancer patients. This review is based on the Prism statement consisting of published studies in the field of predicting the survival of breast cancer patients using data mining techniques from 2005 to 2018 in databases such as Medline, Science Direct, Web of Science, Embase data and Scopus. After searching in these databases, 527 articles were retrieved. After removing duplicates and evaluating the articles, 21 articles were used. The three techniques of logistic regression, decision tree, and support vector machine have been most used in articles. Age, tumor grade, tumor stage, and tumor size are used more than other risk factors. Among the criteria, the accuracy criterion was used in more studies. Most of the studies used the Surveillance, Epidemiology, and End Results Program (SEER) dataset. Typically, in the field of survival probability prediction, data mining techniques in the field of classification are given more attention due to their adaptation to this field. Accordingly, data mining techniques such as decision tree techniques, logistic regression, and support vector machine were used in more studies than other techniques. The use of these techniques can provide a good basis for clinicians to evaluate the effectiveness of different treatments and the impact of each of these methods on patients' longevity and survival. If the output of these techniques is used to provide the data input required by a decision support system, clinicians can provide risk factors related to the patient, the patient's age, and the patient's physical condition when providing services to breast cancer patients. Through the outputs provided by the decision support system, they provided the most optimal decision to choose the best treatment method and consequently increase patient survival.

Keywords: breast neoplasm, data mining, survival.