مجله محاسبات نرم، جلد ۱۰، شماره ۱ (بهار و تابستان ۱۴۰۰)، صفحه ۴۸ـ۶۹
https://dx.doi.org/10.22052/scj.2022.243330.1039

مجله محاسبات نرم
SOFT COMPUTING JOURNAL
تارنمای مجله: scj.kashanu.ac.ir

دانشگاه کاشان
University of Kashan

پیش‌بینی زمان بقا عاری از بیماری تهاجمی در بیماران مبتلا به سرطان پستان با به‌کارگیری روش‌های یادگیری ماشین نیمه نظارتی مبتنی بر گراف✧

رمضان تیموری یانسری[۱]، مربی، میترا میرزارضایی[۱*]، استادیار، مهدی صادقی[۲]، دانشیار، بابک نجار اعرابی[۳]، استاد

[۱] گروه مهندسی کامپیوتر، واحد علوم و تحقیقات، دانشگاه آزاد اسلامی، تهران، ایران.

[۲] پژوهشگاه ملی مهندسی ژنتیک و زیست‌فناوری، تهران، ایران.

[۳] دانشکده مهندسی برق و کامپیوتر، دانشکدگان فنی، دانشگاه تهران، تهران، ایران.

اطلاعات مقاله

تاریخچه مقاله:

دریافت ۱۱ آذر ماه ۱۴۰۰

پذیرش ۰۸ فروردین ماه ۱۴۰۱

کلمات کلیدی:

تاموکسی‌فن

تحلیل بقا

زمان بقا عاری از بیماری تهاجمی

سرطان پستان

مدل شتاب‌دار زمان شکست

مدل خطرهای متناسب کاکس

یادگیری ماشین

یادگیری نیمه نظارتی مبتنی بر گراف

چکیده

سرطان پستان در حال حاضر شایع‌ترین سرطان تشخیص‌داده شده و علت اصلی مرگ و میر ناشی از سرطان در زنان در سراسر جهان است. در سال‌های اخیر در حوزه مطالعات سرطان پستان و روند درمان این بیماری، تحلیل زمان بقای بیماران مبتلا، بسیار مورد توجه بوده است. انتخاب مدل مناسب برای تحلیل زمان بقا چالش اصلی در تحلیل بقا این بیماران است. در این پژوهش کاربردی به کمک روش های یادگیری ماشین نیمه نظارتی مبتنی بر گراف، مدلی برای تحلیل بقا بیماران مبتلا به سرطان پستان پیشنهاد شده است. اطلاعات بالینی و فارماکوژنومیکی، به همراه نتایج مصرف داروی تاموکسی‌فن در فرایند درمان سرطان تهاجمی مربوط به ۳۸۳۳ بیمار مبتلا به سرطان پستان که در بازه ۵ سال تحت پیگیری بوده‌اند، مورد استفاده قرار گرفته است. همچنین با شبیه‌سازی مدل‌ها در نرم‌افزار متلب، عملکرد مدل پیشنهادی در تخمین زمان بقای عاری از بیماری تهاجمی و سایر پارامترهای بقا با مدل های رایج تحلیل بقا، مورد ارزیابی قرار گرفته است. نتایج نشان می‌دهد که با به‌کارگیری مدل پیشنهادی تحلیل بقا در پیش‌بینی زمان بقا عاری از سرطان پستان تهاجمی و همچنین استفاده ترکیبی از ویژگی‌های بالینی و فارماکوژنومیکی، دقت پیش‌بینی ۱۴ درصد بیشتر از زمانی بود که فقط از ویژگی‌های بالینی استفاده شد و ۱۵ درصد بیشتر از زمانی بود که فقط ویژگی‌های فارماکوژنومیکی به کار گرفته شد. علاوه بر این، مدل پیشنهادی تحلیل بقا در پیش‌بینی زمان بقا عاری از بیماری تهاجمی و پارامتر نسبت خطر در مقایسه با مدل‌های رایج تحلیل بقا دقت بالاتری داشته است.

۱. مقدمه

سرطان پستان شایع‌ترین سرطان زنان در جهان است که ابتلا به آن در میان زنان در حال افزایش است. با وجود پیشرفت‌هایی که در تشخیص و درمان این بیماری صورت گرفته است، اما همچنان دومین علت مرگ ناشی از سرطان در بین زنان است.

✧ نوع مقاله: پژوهشی

* نویسنده مسئول

پست(های) الکترونیک: r.taimourei@srbiau.ac.ir (تیموری یانسری)

mirzarezaee@srbiau.ac.ir (میرزارضایی)

sadeghi@nigeb.ac.ir (صادقی)

araabi@ut.ac.ir (نجار اعرابی)

نحوه ارجاع به مقاله: تیموری یانسری، رمضان، میرزارضایی، میترا، صادقی، مهدی، نجار اعرابی، بابک، «پیش‌بینی زمان بقا عاری از بیماری تهاجمی در بیماران مبتلا به سرطان پستان با به‌کارگیری روش‌های یادگیری ماشین نیمه نظارتی مبتنی بر گراف»، مجله محاسبات نرم، جلد ۱۰، شماره ۱، ص۴۸ـ۶۹، بهار و تابستان ۱۴۰۰.

مطالعه بر روی این بیماری جهت پیش‌بینی موارد ابتلا و همچنین به منظور پیشگیری از بیماری، اقدامات دارویی و اصلاح سبک زندگی ضروری است [۱]. در این نوع سرطان رشد و تکثیر غیرقابل کنترل تعدادی از سلول‌های ناهنجار توموری را تشکیل می‌دهند که به صورت یک توده احساس می‌شوند. در صورت رشد تومور، از طریق غدد لنفاوی مجاور پستان و یا از طریق جریان خون، سرطان متاستاتیک[1] ایجاد می‌شود [۲]. سرطان پستان مجرایی تهاجمی[2]، سرطانی است که دیواره غدد شیری را شکافته و شروع به انتشار به بافت پستانی مجاور خارج از مجرا مانند بافت فیبروز و چربی نموده و به مرور به غدد لنفاوی و سایر نواحی بدن منتشر می‌گردد. این سرطان تهاجمی یا نفوذ کننده، شایع‌ترین نوع سرطان پستان می‌باشد [۳]. یکی از متداول‌ترین روش‌ها در درمان سرطان پستان، شناسایی گیرنده‌های هورمون‌های استروژن و پروژسترون است که عامل رشد خیلی از تومورهای پستان هستند. روش‌های مختلفی برای درمان سرطان پستان وجود دارد که رایج‌ترین روش‌های درمان سرطان پستان عبارتند از: جراحی سرطان پستان، شیمی درمانی، پرتودرمانی، هورمون درمانی، درمان هدفمند و ایمونوتراپی. افراد مبتلا به سرطان پستان ممکن است ترکیبی از چند نوع درمان را دریافت کنند. هورمون درمانی کمکی[3] روشی است که سلول‌های سرطانی را از هورمون استروژن که برخی سلول‌های سرطانی پستان برای رشد به آن نیاز دارند، محروم می‌کند. در اکثر موارد هورمون درمان کمکی با استفاده از داروی تاموکسی‌فن[4] انجام می‌شود [۴]. داروی تاموکسی‌فن به طور گسترده‌ای در درمان گیرنده مثبت استروژن سرطان پستان استفاده می‌شود و گیرنده های استروژن را هدف قرار می‌دهد. هنگامی که تاموکسی‌فن به عنوان یک روش هورمون درمانی کمکی در مراحل اولیه سرطان پستان تجویز می‌شود، باعث کاهش قابل ملاحظه میزان عود و مرگ و میر در بیماران مبتلا می‌شود. با این حال، تنوع بالا پاسخ‌های مشاهده شده میان افراد، نیازمند رویکرد شخصی به درمان تاموکسی‌فن است. تاثیرگذاری متفاوت این دارو در بیماران مختلف به ویژگی‌های جمعیتی، سابقه‌ای، فنوتایپی و ژنوتایپی بیماران برمی‌گردد. در بررسی نحوه تاثیرگذاری تاموکسی‌فن در درمان سرطان پستان، می‌توان سرعت متابولیزه شدن دارو را در دسته‌های تاثیر کند، متوسط و سریع طبقه‌بندی کرد [۵].

فاصله زمانی تا وقوع بعضی حوادث مانند مرگ و میر، از جمله داده‌هایی است که همواره مورد علاقه پژوهشگران حوزه پزشکی است. در این دسته از تحقیقات، شرایط گروهی از افراد، به جهت آنکه پس از مدتی برای هر کدام از آنها یک نقطه زمانی، به نام شکست یا وقوع حادثه رخ می‌دهد یا خیر، پیگیری می‌شود. مواردی که می‌تواند معرف شکست یا واقعه مورد نظر باشد، طول عمر یک بیمار و یا اولین زمان مراجعه یک بیمار به پزشک است [۶]. روش‌هایی که برای مطالعات داده‌های مرتبط با مرگ و میر استفاده می‌شد، به نام روش‌های تجزیه و تحلیل زمان بقا نامیده شده‌اند. تحلیل بقا[5]، به مجموعه‌ای از روش‌های آماری و محاسباتی اطلاق می‌شود که متغیرهای پاسخ آنها زمان لازم تا رخداد پیشامد هستند. در این تعریف، متغیر زمان می‌تواند، سال، ماه، هفته یا زمان شروع یک مطالعه تا زمان رخداد پیشامد مورد نظر یا سن افراد در زمان رخداد پیشامد مورد نظر باشد. همچنین، معنی پیشامد در این تعریف، مرگ، بروز بیماری، بازگشت به بیماری، بهبودی از بیماری یا هر تجربه تعریف شده‌ای است که افراد با آن مواجه‌اند، می باشد [۷]. مهمترین هدف در تحلیل بقا، محاسبه زمان بقا، به دست آوردن توزیع بقا با استفاده از داده‌های بقا و در نهایت مقایسه توزیع‌های زمان بقا بین گروه‌های مورد مطالعه می‌باشد. در فرایند تحلیل بقا نمی‌توانیم همه داده‌های موجود برای بررسی طول عمر بیمار را مورد مطالعه قرار دهیم. ممکن است افراد قبل از پایان مطالعه به علتی غیر از علت مورد بررسی از مطالعات بالینی خارج شوند که منجر به تولید داده‌های سانسور شده[6] می‌شود. در تحلیل بقا، توجه به داده‌های سانسور شده ضروری

[1] Metastatic
[2] Invasive Ductal Carcinoma (IDC)
[3] Hormone Adjuvant therapy
[4] Tamoxifen
[5] Survival Analysis
[6] Censored Data

است. زمانی که اطلاعات مربوط به زمان تا نتیجه رویداد برای همه شرکت‌کنندگان مطالعه در دسترس نباشد، با داده‌های سانسور شده مواجه هستیم. در ضمن زمانی که اطلاعات مربوط به زمان رویداد افراد مورد مطالعه به دلیل از دست دادن پیگیری یا عدم وقوع نتیجه رویداد قبل از پایان کارآزمایی در دسترس نباشد، گفته می‌شود که شرکت‌کنندگان سانسور شده‌اند [۸]. در این پژوهش زمان بقای عاری از بیماری تهاجمی[1] یا عدم عود بیماری، فاصله زمانی از تشخیص اولیه تا عود مجدد بیماری در نظر گرفته می‌شود.

روش کاپلان مایر[2] یکی از روش‌های متداول و محبوب برای تجزیه و تحلیل داده‌های زمان تا رویداد است. این روش به ما امکان می‌دهد منحنی بقای متوسط مربوط به یک جمعیت را تخمین بزنیم [۹]. یکی دیگر از مدل‌های رایج برای تحلیل داده‌های بقا، مدل خطرهای متناسب معروف به مدل کاکس Cox-PH[3]، است. فرم کلی مدل خطرهای متناسب کاکس، مخاطره از زمان t برای هر فرد از جامعه مورد مطالعه را بیان می‌کند. در این مدل، مخاطره در زمان t به صورت حاصل ضرب دو کمیت بیان می شود (رابطه (۱)). در این رابطه $h_0(t)$ تابع خطر پایه[4] نامیده می‌شود. تابع خطر پایه توضیح می‌دهد چگونه مخاطره رخداد، در هر واحد زمانی در طول زمان در سطوح ابتدایی پارامترهای اثر تغییر می‌کند.

$$h(t|x) = h_0(t)e^{\beta_1 x_1 + \cdots \beta_p x_p} \qquad (۱)$$

در رابطه فوق، دومین کمیت، عبارت نمایی است به صورت $\sum \beta_i X_i$ که این مجموع روی p متغیر توضیحی جمع بسته شده است. در این رابطه، X متغیرهایی هستند که در فرایند بقا موثر هستند و متغیرهای پیش‌بینی‌کننده نامیده می‌شوند. همچنین پارامتر β نشان‌دهنده اندازه‌ی اثر یک متغیر یا همان ضریب رگرسیون است [۱۰].

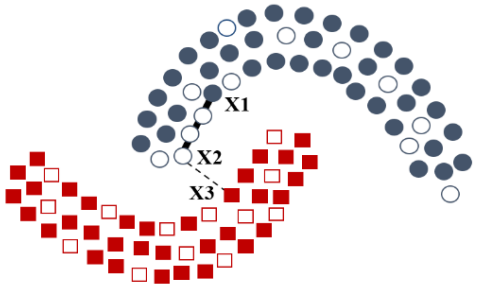
مدل شتاب‌دار زمان شکست AFT[5] مدلی برای تجزیه و تحلیل بقا است. در مدل AFT فرض این است که اثر متغیرهای کمکی نسبت به زمان بقا، افزاینده است (متناسب)، در حالی که فرض اساسی در مدل Cox-PH این است که اثر متغیرهای کمکی نسبت به خطر افزاینده است [۱۱]. در مدل AFT برای زمان بقا، فرض می‌شود که لگاریتم زمان بقا t_i به طور خطی با متغیرهای کمکی x_i مرتبط است (رابطه (۲)).

$$ln(T) = \beta_1 x_1 + \cdots + \beta_p x_p + Ln(\varepsilon) \qquad (۲)$$

در روش‌های یادگیری نظارتی[6]، نگاشتی از داده‌های ورودی به خروجی مشخص شده توسط مربی انجام می‌شود. در روش‌های یادگیری بدون نظارت[7] به دنبال یافتن الگوی شباهت میان داده‌ها و دسته‌بندی آنها درون خوشه‌ها هستیم ولی در روش‌های یادگیری نیمه نظارتی، یادگیری با استفاده از تعداد کمی داده‌های برچسب‌دار و تعداد بسیار زیادی داده‌های بدون برچسب تحقق می‌یابد [۱۲، ۱۳]. یادگیری نیمه‌نظارتی[8]، روشی میان یادگیری نظارتی و یادگیری غیرنظارتی است. در این روش بعضی از داده‌ها دارای برچسب بوده و بعضی نیز برچسب ندارند. یکی از مهمترین دلایل استفاده از یادگیری نیمه‌نظارتی در برنامه‌های کاربردی آن است که در خیلی از این برنامه‌ها عملیات برچسب‌گذاری داده‌های بدون برچسب، معمولا هزینه‌بر و زمان‌بر می‌باشد. در یادگیری نیمه‌نظارتی هدف یافتن روش‌هایی است که با استفاده از داده‌های برچسب‌دار و داده‌های بدون برچسب به حل مساله دسته‌بندی پرداخته و کارایی دسته‌بندی را نسبت به قبل افزایش دهد. به عبارتی دیگر در یادگیری نیمه‌نظارتی با استفاده همزمان از داده‌های بدون برچسب و داده‌های برچسب‌دار به دنبال بهبود دقت یادگیری هستیم [۱۴]. در روش‌های یادگیری با نظارت مطابق با فرض هموار نظارتی، برچسب دو داده بافاصله نزدیک به هم دارای مقادیر نزدیک به هم است ولی در مقابل فرض هموار نیمه‌نظارتی بیان می‌کند که برچسب دو داده در یک ناحیه چگال و بافاصله نزدیک به هم دارای مقادیر نزدیک است. به عبارتی برچسب

[1] Invasive Disease–Free Survival (IDFS)
[2] Kaplan-Meier (KM)
[3] Cox Proportional Hazard Model (Cox-PH)
[4] Baseline Hazard Function
[5] Accelerated Failure Time Model (AFT)
[6] Supervised Learning
[7] Unsupervised Learning
[8] Semi-Supervised Learning

داده‌ها بر روی گراف به صورت هموار تغییر می‌کند [۱۵]. مطابق شکل (۱) نمونه‌های برچسب‌دار X_1 و X_3 و نمونه بدون برچسب X_2 نشان داده شده است. اگرچه فاصله اقلیدسی بین نقاط X_2 و X_3 کوچک‌تر از فاصله بین X_2 و X_1 است، ولی طبق گراف همسایگی، شباهت بین X_2 و X_1 بیشتر از شباهت بین X_2 و X_3 است؛ لذا، برچسب نمونه بدون برچسب X_2 بیشتر تحت تاثیر برچسب X_1 است.



شکل (۱): تعیین برچسب نمونه جدید با استفاده از گراف همسایگی

روش‌های یادگیری نیمه‌نظارتی مبتنی بر گراف، یکی از روش‌های یادگیری نیمه‌نظارتی هستند. در میان روش‌های یادگیری نیمه‌نظارتی، در صورتی استفاده از روش‌های مبتنی بر گراف مفید خواهد بود که علاوه بر برقراری فرض هموار نیمه‌نظارتی، فرض خمینه نیز برقرار باشد. مطابق فرض خمینه، مجموعه نقاط مساله در یک فضا با بعد بالا، در حقیقت بر روی یک خمینه با بعد کم قرار دارند. جهت بررسی برقراری شرط فرض خمینه باید ساختار خمینه را بیان کرد. از روش‌های رایج در بیان ساختار خمینه، استفاده از گراف‌های همسایگی است. در گراف همسایگی، راس‌ها همان نمونه‌ها هستند و نمونه‌های نزدیک به هم روی خمینه با استفاده از یال با وزن متناسب مشخص می‌شوند [۱۴].

در تمام الگوریتم‌های نیمه‌نظارتی مبتنی بر گراف، برای برچسب‌گذاری نمونه‌های بدون برچسب مراحل زیر انجام می‌شود. همانند تمام الگوریتم‌های یادگیری ابتدا گام پیش‌پردازش داده‌ها که شامل استخراج ویژگی‌ها، کاهش بعد و حذف نویز است، انجام می‌شود. سپس با محاسبه فاصله بین نقاط، گراف همسایگی روی نقاط ساخته می‌شود و در انتها برچسب نقاط بدون برچسب با یکی از روش‌های استنتاج برچسب تعیین می‌شود [۱۵]. در روش‌های یادگیری نیمه‌نظارتی مبتنی بر گراف، از ماتریس لاپلاسین نرمال گراف برای برچسب‌زنی نمونه‌های بدون برچسب استفاده می‌شود. بدین منظور، ماتریس شباهت W از روی گراف همسایگی ساخته شده و از روی آن ماتریس قطری D ساخته می‌شود. با استفاده از ماتریس‌های W و D، لاپلاسین نرمال گراف محاسبه می‌شوند. مقادیر این ماتریس اطلاعات مناسبی در خصوص ساختار گراف می‌دهد [۱۶].

۲. مرور کارهای پیشین

در حوزه تحلیل بقا خاص سرطان پستان پژوهش‌های متنوعی انجام شده است که عمده این تحقیقات از روش‌های تحلیل بقای رایج بهره گرفته‌اند. جیانگ و همکاران، ۲۰۱۳، یک روش جدید برای پیش‌بینی بروز بقاء بیمار با استفاده از شبکه یادگیری بیزی کارا بر روی ۸۹۹ بیمار مبتلا به سرطان پستان را به کمک روش شبکه یادگیری بیزی مورد بررسی قرار دادند. نتایج تحقیق مشخص کرد که مدل پیشنهادی از مدل Cox-PH بهتر عمل کرده و قابل مقایسه با روش جنگل بقا تصادفی است [۱۷]. مطالعه بشیری و همکاران، ۲۰۱۷، در یک تحقیق مروری بهبود پیش‌بینی بقا در بیماران سرطانی با استفاده از تکنیک‌های یادگیری ماشین به کمک داده‌های بیان ژن و دقت بالا و کارایی داده‌های بیان ژن در مقایسه با داده‌های بالینی در پیش‌بینی بقا را نشان داد. همچنین در این پژوهش به کارگیری الگوریتم‌های یادگیری ماشین باعث افزایش کارایی در برای پیش‌بینی بقا از سرطان شد [۱۸]. در مطالعه اندو و همکاران، ۲۰۰۸، دو مدل رگرسیون لجستیک و شبکه عصبی مصنوعی برای پیش‌بینی ۵ ساله بیماران مبتلا به سرطان پستان طراحی شدند. حساسیت[1] و تشخیص‌پذیری[2] برای دو مدل رگرسیون لجستیک و شبکه عصبی مصنوعی بترتیب (۰/۵۹۴ و ۰/۷۰) و (۰/۶۲۱ و ۰/۷۳۲) به دست آمد. در مدل لجستیک و شبکه عصبی دقت و سطح

[1] Sensitivity
[2] Specificity

زیر منحنی[1] ROC بترتیب (۰/۶۸۸ و ۰/۷۲۵) و (۰/۷۰و۰/۷۲۵) بود. در این تحقیق که برای پیش‌بینی بقای ۵ ساله بیماران مبتلا به سرطان پستان با مقایسه ۷ مدل رگرسیون لجستیک، شبکه عصبی مصنوعی، شبکه بیز و انواع درخت تصمیم انجام شد، دقت مدل رگرسیون لجستیک از دیگر مدل‌ها بیشتر بود [۱۹]. باقریان و همکاران، ۲۰۲۱، نشان دادند که از بین تکنیک‌های داده‌کاوی که برای پیش‌بینی احتمال بقا استفاده شده است، سه تکنیک رگرسیون لجستیک، درخت تصمیم و ماشین بردار پشتیبان بیشترین کاربرد را در مقالات داشته‌اند. در بیشتر مطالعات، خطر فاکتورهای سن، گرید تومور، استیج تومور و اندازه تومور استفاده شده بودند [۲۰]. قاسمی و همکاران، ۲۰۱۸، با هدف بررسی عوامل مؤثر بر بقای بیماران مبتلا به سرطان پستان به مطالعه داده‌های کوهورت مربوط به بیماران مبتلا به سرطان پستان که در سال‌های ۱۳۸۵-۱۳۷۶ به پژوهشکده معتمد جهاد دانشگاهی در تهران مراجعه کرده بودند و در سال‌های ۱۳۹۶-۱۳۹۱ مورد پیگیری قرار گرفتند، پرداختند. در این مطالعه تعداد ۲۷۰ بیمار به صورت تصادفی انتخاب و مشخصه‌های فردی آنها ثبت شد. نتایج برازش مدل نشان داد مدل بتا وایبل پواسون برازش بهتری نسبت به مدل شفایافته ناآمیخته وایبل دارد [۲۱]. مطالعات محدودی در زمینه مقایسه مدل‌های مختلف بر داده‌های بقاء وجود دارد که در برخی به کارایی مدل‌های پارامتری در برخی دیگر به کارایی بالاتر مدل‌های نیمه پارامتری اشاره شده است. انتخاب روش مدل‌سازی و تحلیل آن به ماهیت متغیرها و شرایط حاکم بر مساله بستگی دارد. بر این اساس مدل‌های مختلف بر روی مجموعه داده‌های متفاوت کارایی یکسانی ندارند، از این رو تحلیل مجموعه داده‌های مختلف مستلزم جستجو در میان مدل‌های مختلف برای یافتن کاراترین مدل می‌باشد. منابع مطالعاتی منتشر شده نشان می‌دهد که مدل نیمه‌پارامتری Cox-PH بیشتر مورد توجه محققان و نویسندگان قرار گرفته است. این محبوبیت نشان می‌دهد که مدل مخاطرات کاکس در شرایط وجود داده‌های حذف شده، نتایج قانع‌کننده‌ای را در مقایسه با دیگر مدل‌ها فراهم می‌آورد [۱۹، ۲۲]. کیانی و همکاران، ۲۰۱۴، با استفاده از داده‌های ۸۰۹ بیمار مبتلا به سرطان پستان، با به‌کارگیری درخت از یک مدل پیش آگهی مبتنی بر داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان، استفاده کردند. نتایج این تحقیق نشان داد، ویژگی و حساسیت مدل توسعه یافته بترتیب ۵۳ و ۸۵ درصد بود. این مدل، تنها ۱۴ درصد از بیماران دچار عود مجدد را به اشتباه، مستعد عود مجدد پیش‌بینی کرده است [۲۳]. در پژوهش صادقی و همکاران، ۲۰۲۰، به کمک الگوریتم ژنتیک متغیرهای وابسته کاهش یافت و از الگوریتم نزدیک‌ترین همسایگی برای پیش‌بینی عود سرطان پستان استفاده شده است. نتایج ارزیابی مدل نشان داد که مقدار صحت برای مدل پیشنهادی ۷۷/۱۴ درصد است که نسبت به روش‌های دیگر خروجی مناسب‌تری دارد [۲۴]. در تحقیق مولر و همکاران، ۲۰۰۷، بررسی زمان بقای سرطان پستان خانوادگی برای ۴۴۲ بیمار مبتلا به سرطان پستان به کمک روش کاپلان میر نشان از تفاوت نتایج بر اساس وضعیت جهش ژن BRCA1 و BRCA2 داشته و مشخص شد بقای ۵ ساله برای بیماران دارای جهش در ژن BRCA1 برابر ۷۲٪ و در مقایسه ۹۶٪ برای بیماران دارای جهش در BRCA2 بود [۲۵]. در یکی از پژوهش‌های جدید لیانگ و همکاران، ۲۰۱۶، تجزیه و تحلیل بقا سرطان را با استفاده از روش یادگیری نیمه‌نظارتی مبتنی بر مدل‌های Cox-PH و زمان شکست شتاب‌دار AFT انجام دادند. نتایج این تحقیق نشان داد، مدل یادگیری نیمه‌نظارت می‌تواند به طور قابل توجهی پیش‌بینی عملکرد مدل‌های AFT و Cox-PH در تحلیل بقا را بهبود دهد [۲۶]. چای و همکاران، ۲۰۱۷، همچنین یک مدل یادگیری نیمه‌نظارت شده با ترکیب مدل Cox-PH و زمان شکست شتاب‌دار ارائه دادند و نتایج این تحقیق نشان داد که مدل ترکیبی پیشنهادی Cox-SP-AFT با استفاده از داده‌های سانسورشده به شیوه صرفا گام‌به‌گام می‌تواند از نمونه‌های سانسور شده بیشتر استفاده کند و زمان بقا خود را با دقت بیشتری برآورد کند [۲۷]. در ضمن خلاصه مهم‌ترین پژوهش‌ها در حوزه تحلیل بقا سرطان پستان به همراه بیان نقاط قوت و ضعف روش‌ها در پیوست (۱) آمده است.

[1] Receiver Operating Characteristic

۳. روش پیشنهادی

انتخاب مدل مناسب برای تحلیل زمان بقا چالش اصلی در تحلیل بقا بیماران مبتلا به سرطان پستان است. در این پژوهش کاربردی به کمک روش‌های یادگیری ماشین نیمه‌نظارتی مبتنی بر گراف، مدلی برای تحلیل بقا بیماران مبتلا به سرطان پستان پیشنهاد شده است. از مهم‌ترین چالش‌هایی که در اکثر تحلیل‌های بالینی با آن سروکار داریم وجود نمونه‌هایی در مجموعه‌داده‌ها است که بعضی از ویژگی‌های آنها دارای مقدار نیستند، گرچه از سایر ویژگی‌ها اطلاعات مفیدی حاصل می‌شود. بدین منظور و برای استفاده از تمام اطلاعات مجموعه داده‌ای به کمک روش‌های یادگیری ماشین نیمه‌نظارتی تلاش شده تا مقادیر این ویژگی‌ها تعیین و از آنها در پیش‌بینی زمان بقا عاری از بیماری تهاجمی استفاده شود.

در این پژوهش کاربردی، ترکیبی از اطلاعات بالینی و فارماکوژنومیکی (ویژگی‌های جمعیتی، سابقه‌ای، فنوتایپی و ژنوتایپی) بیماران مبتلا به سرطان پستان که تحت درمان داروی تاموکسی‌فن با مصرف دوز بیست میلی‌گرم در روز، در بازه پنج سال مورد پیگیری قرار گرفته‌اند، استفاده شده است. به کمک مدل پیشنهادی، با داشتن اطلاعات بالینی و فارماکوژنومیکی نمونه‌های مختلف سرعت متابولیزه شدن تاموکسی‌فن در دسته‌های تاثیر کند، متوسط و سریع که نشان‌دهنده تأثیرات متفاوت آن بر بیماران است، زمان بقای عاری از بیماری تهاجمی آنها تخمین زده می‌شود. گام‌های مدل پیشنهادی برای تحلیل بقا به کمک داده‌های بالینی و فارماکوژنومیکی در شکل (۲) نشان داده شده است.

الف) پیش‌پردازش: در گام پیش‌پردازش در مدل پیشنهادی، مهم‌ترین موارد حذف داده‌های ناخواسته و نرمال‌سازی داده‌ها است. سپس مجموعه داده‌ها به سه زیر مجموعه: (۱) مجموعه آموزشی شامل ۷۰ درصد کل داده‌ها، (۲) مجموعه اعتبارسنجی شامل ۱۵ درصد و (۳) مجموعه آزمایشی به نسبت ۱۵ درصد به صورت تصادفی از مجموعه داده‌ها تقسیم می‌شود. سپس در هر دسته از داده‌های انتخاب شده، داده‌های برچسب‌دار، مقادیر گمشده و داده‌های سانسور شده با توجه به تعریف مشخص می‌شود. در کنار داده‌های دارای برچسب و داده‌های گمشده، داده‌های سانسورشده هم وجود دارند که یکی از ویژگی‌های مدل پیشنهادی استفاده از این داده‌های در فرایند تحلیل بقای داده‌ها است.

شکل (۲): مدل پیشنهادی تحلیل بقا با استفاده از داده‌های بالینی و فارماکوژنومیکی

یکی از گام‌های مهم که می‌تواند در انجام پردازش‌ها کمک شایانی به بهتر شدن نتایج کند، انتخاب درست ویژگی‌ها است. تعداد زیاد ویژگی‌ها می‌تواند تعداد محاسبات را به‌شدت افزایش دهد درصورتی‌که تغییر محسوسی در صحت ایجاد نکند [۲۸]. یکی از امتیازات پژوهش انجام شده، استفاده همزمان از ویژگی‌های بالینی و فارماکوژنومیکی است، نکته‌ای که در خیلی

از تحقیقات مشابه دیده نشده است.

در این گام با مقایسه روش‌های مختلف انتخاب ویژگی بر روی داده‌های انتخاب شده، با توجه به دقت دسته‌بندی از روش درخت تصادفی برای انتخاب بهترین ویژگی‌ها استفاده شده است. با اعمال روش انتخاب ویژگی، از بین ویژگی‌های موجود، بیست و دو ویژگی برای تحلیل بقای نهایی انتخاب شده‌اند که این ویژگی عبارتند از: سن در تشخیص اولیه سرطان پستان، سابقه سرطان، وضعیت یائسگی، حداکثر ابعاد تومور، درجه ناتینگهام بافت تومور، وضعیت استعمال سیگار در زمان ابتلا بیماری، وضعیت گیرنده استروژن، وضعیت گیرنده پروژسترون، وضعیت پرتو درمانی، وضعیت شیمی درمانی، دوز مصرف تاموکسی‌فن، بازه زمانی مصرف تاموکسی‌فن، وضعیت بازدارندگی ژن CYP2D6، وضعیت متابولیزه شدن تاموکسی‌فن با توجه به ژنوتیپ‌های آلل‌های مختلف ژن CYP2D6، وضعیت متاستاز بیماری اولیه، تعداد گره‌های مثبت از گره‌های بیوپسی غده لنفاوی پیش‌آهنگ و غدد لنفاوی زیربغل، درمان کمکی مهارکننده آروماتاز، هورمون‌درمانی پس از جراحی پستان، وجود درمان منظم برای پیشگیری از سرطان پستان، وضعیت فوت بیمار، وجود سرطان دومی یا عود مجدد محلی یا منطقه‌ای همان طرف بدن (تهاجمی یا غیرتهاجمی)، وضعیت سرطان دوم یا عود مجدد دور از پستان.

ب) دسته‌بند اول: مدل Cox-PH یک مدل موفق در دسته‌بندی داده‌ها به دسته پرخطر و کم‌خطر است. برای افزایش دقت تحلیل زمان بقا، در گام دوم به کمک این مدل، از مجموعه داده‌های آموزشی نمونه‌های ورودی بر اساس زمان بقای محاسبه شده در دو گروه نمونه‌های با خطر بالا ابتلا به سرطان و نمونه‌های با خطر پایین ابتلا به سرطان قرار می‌گیرند.

ج) دسته‌بند دوم: در گام سوم به کمک مدل AFT روی داده‌هایی که برچسب خطر بالا و خطر پایین دارند و همچنین داده‌های سانسورشده تحلیل بقا انجام شده و برای نمونه‌های انتخاب شده خطر نسبی ابتلا به سرطان پستان تهاجمی تخمین زده می‌شود. دسته‌بندی داده‌های در دسته‌های خطر بالا و خطر پایین موجب افزایش کارایی مدل AFT می‌شود. با توجه به وظیفه اصلی مدل که محاسبه و تحلیل زمان بقا به کمک داده‌های برچسب‌دار و داده‌های گمشده است، در گام چهارم با استفاده از الگوریتم‌های نیمه‌نظارتی مبتنی بر گراف، به تخمین مقدار داده‌های گمشده می‌پردازیم.

د) تخمین مقادیر گمشده : الگوریتم‌های نیمه‌نظارتی مبتنی بر گرافی که در مدل پیشنهادی استفاده شده است، الگوریتم نیمه‌نظارتی مبتنی بر گرافی با سازگاری محلی و سراسری[1] است. در این روش برای برچسب‌زنی نمونه‌های بدون برچسب، از ماتریس لاپلاسین نرمال گراف استفاده می‌شود. با استفاده از این الگوریتم ابتدا ماتریس شباهت W ساخته می‌شود. سپس از روی ماتریس شباهت، ماتریس قطری D و لاپلاسین نرمال گراف محاسبه خواهد شد. بر پایه اصل پیوستگی و فرض‌های یادگیری نیمه‌نظارتی مبتنی بر گراف برچسب دو داده در یک ناحیه چگال و بافاصله نزدیک به هم دارای مقادیر نزدیک است. به عبارتی برچسب داده‌ها بر روی گراف به‌صورت هموار تغییر می‌کند. در گراف وزن‌دار $G = (E, V)$، V مجموعه راس‌های متناظر با هر کدام از نمونه‌ها و E مجموعه یال‌های بین راس‌ها بوده و با ماتریس شباهت W نشان داده می‌شود. هر کدام از درایه‌های ماتریس W با تابع هسته گوسی مطابق رابطه (۳) به دست می‌آیند.

$$W_{ij} = exp\left(\frac{-\|X_i - X_j\|}{(2\delta^2)}\right) \qquad (۳)$$

در این رابطه x_j و x_i نمونه‌هایی دارای برچسب و بدون برچسب در مجموعه داده‌ها هستند و $\|\mathrm{X_i} - \mathrm{X_j}\|$ فاصله اقلیدسی نمونه‌ها از یکدیگر است. در ضمن δ انحراف معیار فاصله‌ها است که از روی مجموعه داده‌ها محاسبه می‌شود. مجموعه داده‌های آموزشی $X = \{X_L, X_U\}$ از ترکیب دو مجموعه، داده‌های برچسب‌دار $X_L = \{X_1, X_{|L|}\}$ و داده‌های بدون برچسب $X_U = \{X_{|L|+1}, X_{|U|}\}$ تشکیل شده است؛ مجموعه برچسب‌ها به صورت $Y_{n\times c} = \{Y_L, Y_U\}$ که n تعداد نمونه‌ها و c تعداد کلاس برچسب‌ها است، در نظر گرفته می‌شود. دقت کنید که Y_L ماتریس برچسب X_L ها و Y_U ماتریس

1 Local and Global Consistency

برچسب X_U ها است. برای برچسب‌گذاری نمونه‌های بدون برچسب $Y_{l(i,k)} = 1$ خواهد بود اگر X_i دارای برچسب کلاس kام باشد و در غیر این صورت برابر صفر خواهد بود. ماتریس Y_U در ابتدا برابر صفر است. با استفاده از الگوریتم نیمه‌نظارتی مبتنی بر گرافی با سازگاری محلی و سراسری پیش‌بینی برچسب برای مجموعه Y_U انجام می شود (شکل (۳)) [۱۶].

Algorithm Yu = SSL with Local and Global Consistency (V, E ,Y_L)
Input: Graph (V,E), labels Y_L
Output: labels Y_U

$\forall_i , D_{ii} = \sum_j W_{i,j}$
$S = D^{-1/2}WD^{-1/2}$
$Y_U = 0$
$Y^{(0)} = [Y_L, Y_U], t = 0$
Repeat
$Y^{(t+1)} = \alpha SY^t + (1-\alpha)Y^0,\ \alpha \epsilon (0,1)$
t = t+1
Until $Y^{(t)}$
Y^* denote the limit of the sequence {Y(t)}
Label each point X_i as a label $Y_i = arg\ \max_{j<=c} Y^*_{ij}$

شکل (۳): الگوریتم نیمه نظارتی مبتنی بر گراف جهت تخمین مقدار داده‌های گمشده [۱۶]

ویژگی اصلی الگوریتم یادگیری نیمه‌نظارتی مبتنی بر گرافی با سازگاری محلی و سراسری در این است که از تمام نمونه‌های برچسب‌دار و بدون برچسب در فرایند یادگیری بهره می‌گیرد. همچنین این روش یادگیری با در نظر گرفتن سازگاری محلی و سراسری، احتمال برچسب‌ها در نزدیکی هر نمونه و سایر نمونه‌هایی که به کمک گراف همسایگی در مجاورت نمونه‌ها و همسایه آن در نظر گرفته می‌شوند، قادر است برچسب نمونه‌های بدون برچسب را با دقت بالاتری محاسبه کند. در انتهای این گام، مجموعه‌داده‌های آموزشی دارای برچسب با اضافه‌کردن داده‌های تخمین زده شده (برچسب‌گذاری شده جدید)، به‌روزرسانی می‌شود. در گام پنجم، بررسی می‌شود که آیا مقدار تمام داده‌های گمشده، تعیین شده است یا خیر. اگر تمام داده‌های گمشده مقدار نداشتند، برای افزایش دقت دسته‌بندی مراحل کار از گام دوم تکرار می‌شود و از مجموعه برچسب خورده به‌روز شده حاصل در پایان این گام برای تعیین مقدار داده‌های گمشده در تکرارهای بعدی استفاده می‌شود. مراحل فوق تا زمانی که شرط بیان شده درست باشد و مدل برای تمامی داده‌های گمشده، مقدار تعیین کند، تکرار خواهد شد.

ه) دسته‌بند نهایی: اگر تمام داده‌های گمشده مقدار داشتند، در گام نهایی با خاتمه برچسب‌گذاری داده‌های بدون برچسب با الگوریتم نیمه‌نظارتی، با داشتن تمام داده‌های برچسب‌گذاری شده، به کمک مدل Cox-PH زمان بقا عاری از بیماری تهاجمی تخمین زده می‌شود. در گام نهایی همچنین خطر نسبی برای گروه‌های مختلف داده‌ای محاسبه می‌شود.

۴. داده‌ها

مدل‌های بررسی شده همزمان از اطلاعات بالینی و فارماکوژنومیک و دانش موجود در مطالعات بالینی و داده‌های فارماکوژنومیکی در تحلیل پارامترهای بقای بیماران مبتلا به سرطان پستان استفاده نمی‌کنند. در پژوهش حاضر مجموعه داده‌های بیماران مبتلا به سرطان پستان که در بازه ۵ سال تحت پیگیری بوده‌اند مورد استفاده قرار گرفته است. اطلاعات بالینی و فارماکوژنومیکی شامل نتایج مصرف دارو تاموکسی‌فن و نحوه تأثیرگذاری آن در فرایند درمان سرطان تهاجمی پستان در دسترس و مورد استفاده قرار می‌گیرد.

این مجموعه داده در مجموعه داده‌های بنیاد دانش فارماکوژنومیکس[1] به عنوان مجموعه داده کنسرسیوم بین‌المللی فارماکوژنومیک تاموکسی‌فن (ITPC)[2] در دسترس است. انواع داده‌های مورد استفاده در دسته‌های جمعیتی[3]، سابقه‌ای[4]، فنوتایپی، بالینی یا پیگیری، ژنوتایپی و فارماکوژنومیکی تقسیم‌بندی شده‌اند (جدول (۱)). جزئیات هر یک از انواع داده‌ای و شرح هر یک از ویژگی‌های آن در پیوست (۲) آمده است.

[1] https://www.pharmgkb.org/ ITPC Dataset
[2] International Tamoxifen Pharmacogenomics Consortium (ITPC)
[3] Demographic Data
[4] Background Data

جدول (۱): انواع داده‌ای مورد استفاده در تحلیل به همراه تعداد ویژگی‌های هر نوع داده

نوع داده‌ها	تعداد ویژگی	تعداد ویژگی انتخاب شده
داده‌های جمعیتی	۶ ویژگی	**۱ ویژگی**
داده‌های سابقه‌ای	۱۰ ویژگی	**۱ ویژگی**
داده‌های فنوتایپی	۴۹ ویژگی	**۱۵ ویژگی**
داده‌های بالینی (پیگیری)	۲۱ ویژگی	**۳ ویژگی**
داده‌های ژنوتایپی	۱۶ ویژگی	–
داده‌های فارماکوژنومیکی	۸ ویژگی	**۲ ویژگی**
کل ویژگی‌ها	۱۱۰ ویژگی	**۲۲ ویژگی**

۵. نحوه اعتبارسنجی

در مرحله ارزیابی، برای بررسی میزان موفقیت و اعتبارسنجی مدل پیشنهادی در پیش‌بینی زمان بقا و دسته‌بندی بیماران در دسته‌های خطر بالا و خطر پایین ابتلا به سرطان پستان، از ماتریس درهم‌ریختگی[1] استفاده می‌شود. ماتریس درهم‌ریختگی کارایی یک مدل دسته‌بندی را بر اساس مقایسه مقدار واقعی و مقدار تخمین زده شده محاسبه می‌کند. از نتایج ماتریس درهم‌ریختگی و مهم‌ترین معیار برای تعیین کارایی یک الگوریتم دسته‌بندی محاسبه دقت[2] یا نرخ دسته‌بندی[3] است. نسبت تعداد نمونه صحیح طبقه‌بندی شده توسط الگوریتم از یک کلاس مشخص به کل نمونه‌هایی که الگوریتم چه به صورت صحیح و چه به صورت غلط در آن کلاس طبقه‌بندی کرده است، دقت دسته‌بند را نشان می‌دهد. این معیار نشان می‌دهد مدل دسته‌بندی طراحی شده چند درصد از کل مجموعه رکوردهای را به‌درستی دسته‌بندی کرده است. همچنین نتایج ماتریس درهم‌ریختگی شاخص‌های صحت[4] به معنای نسبت تعداد نمونه‌هایی که در کلاس مثبت طبقه‌بندی شدند به کل نمونه‌هایی که الگوریتم کلاس آنها را به‌درستی پیش‌بینی کرده است و حساسیت[5] به معنای نسبت نمونه‌هایی که مدل در دسته مثبت طبقه‌بندی کرده به تعداد کل نمونه‌های مثبت، نیز برای تحلیل عملکرد سیستم‌های طبقه‌بندی استفاده می‌شود [۲۹، ۳۰]. اجزا ماتریس درهم ریختگی عبارتند از:

- مثبت درست TP: نشان‌دهنده تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته‌بندی نیز دسته آنها را به‌درستی مثبت تشخیص داده است.
- منفی درست TN: به معنای تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته‌بندی نیز دسته آنها را به‌درستی منفی تشخیص داده است.
- مثبت نادرست FP: به معنای تعداد رکوردهایی است که دسته واقعی آنها منفی بوده و الگوریتم دسته‌بندی دسته آنها را به اشتباه مثبت تشخیص داده است.
- منفی نادرست FN: نشان‌دهنده تعداد رکوردهایی است که دسته واقعی آنها مثبت بوده و الگوریتم دسته‌بندی دسته آنها را به اشتباه منفی تشخیص داده است.

محاسبه دقت، صحت و حساسیت طبقه‌بندی کننده‌ها بر اساس فرمول‌های زیر به دست می‌آید:

$$\boldsymbol{Accuracy} = \frac{TN + TP}{TN + TP + FN + FP}$$

$$\boldsymbol{Sensitivity} = \frac{TP}{TP + FN} \qquad (۴)$$

$$\boldsymbol{Specificity} = \frac{TP}{TP + FP}$$

همچنین یکی از مهم‌ترین پارامترها در ارزیابی مدل‌های تحلیل بقا، مقایسه نسبت خطر است که حالت کلی نسبت خطر در مدل‌های تحلیل بقا از رابطه (۵) به دست می‌آید.

$$\widehat{HR} = \frac{\hat{h}(t,\boldsymbol{X}^*)}{\hat{h}(t,\boldsymbol{X})} = \frac{\hat{h}_0(t)e^{\sum_{i=1}^{p}\hat{\beta}_i X_i^*}}{\hat{h}_0(t)e^{\sum_{i=1}^{p}\hat{\beta}_i X_i}} = exp\left[\sum_{i=1}^{p}\beta_i\,(X_i^* - X_i)\right] \qquad (۵)$$

در این رابطه، عبارت نمایی به‌صورت $\sum \beta_i X_i$ روی p متغیر توضیحی جمع بسته شده است. در این رابطه X متغیرهایی هستند که در فرایند بقا موثر هستند و متغیرهای پیش‌بینی‌کننده

[1] Confusion Matrix
[2] Accuracy
[3] Classification Rate
[4] Specificity
[5] Sensitivity

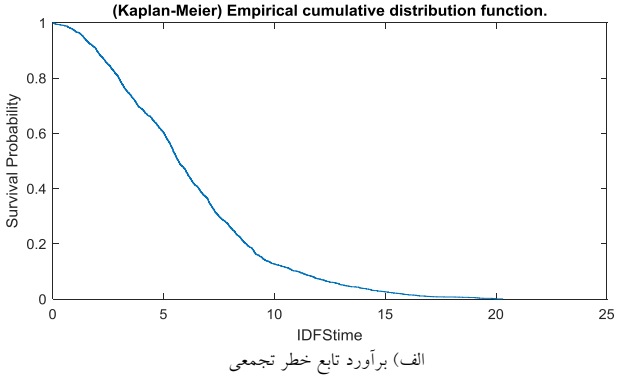
نامیده می‌شوند. پارامتر β نشان‌دهنده اندازه‌ی اثر یک متغیر یا همان ضریب رگرسیون است [۱۰].

در این پژوهش از پارامترهای مطرح شده فوق، با توجه به اهمیت معیارها، از معیار دقت و نسبت خطر در ارزیابی مدل‌های تحلیل بقا استفاده شده و نتایج ارزیابی در بخش نتایج پژوهش ارائه شده است.

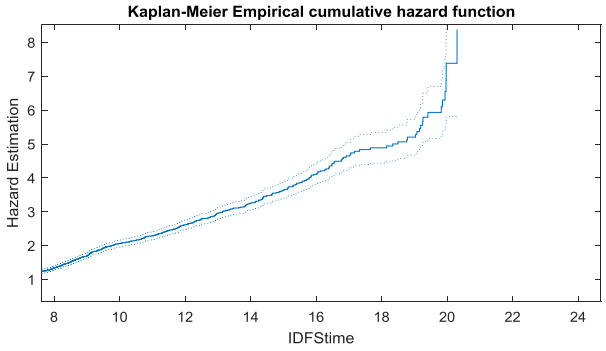
۶. نتایج

با شبیه‌سازی مدل‌ها در نرم‌افزار متلب، عملکرد مدل پیشنهادی در تخمین زمان بقای عاری از بیماری تهاجمی سرطان پستان تهاجمی و سایر پارامترهای بقا با مدل‌های گذشته تحلیل بقا، مورد ارزیابی قرار گرفته است. مجموعه داده‌های ورودی که ترکیبی از اطلاعات بالینی و فارماکوژنومیکی بیماران بود، بر روی مدل پیشنهادی اعمال شد و خروجی مدل شامل تخمین زمان بقای بیماران مبتلا، علاوه بر مقادیر پارامترهای ارزیابی نمایش داده شد. بر اساس زمان بقای تخمین زده شده توسط مدل، دسته‌بندی نهایی بیماران به دسته بیماران با خطر بالا یا پرخطر و خطر کم ابتلا انجام و نسبت خطر به دست آمده است. شکل (۴) برآورد تابع خطر تجمعی Kaplan-Meier را به همراه توزیع احتمال برازش منحنی بقا بر اساس IDFS به کمک مدل Kaplan-Meier در وضعیت‌های مختلف متابولیزه شدن تاموکسی‌فن نشان می‌دهد.

در این تحقیق اطلاعات بیماران مبتلا به سرطان پستان تهاجمی تحت درمان داروی تاموکسی‌فن با دوز بیست میلی‌گرم در روز که در بازه پنج سال مورد پیگیری بوده‌اند، استفاده شده است. بر اساس نتایج این نمودار، احتمال بقا بر اساس زمان عاری از بیماری تهاجمی برای بیماران دارای وضعیت تأثیر سریع متابولیزه شدن تاموکسی‌فن در مقایسه با سایر وضعیت‌های متابولیزه، بالاتر بوده و احتمال بقا برای بیماران دارای وضعیت متابولیزه کند و متوسط در رتبه‌های بعدی قرار دارد. شکل ۵ تخمین احتمال بقای بیماران را بر اساس زمان بقای عاری از بیماری تهاجمی برحسب ماه، در وضعیت‌های مختلف متابولیزه تاموکسی‌فن در بیماران تحت پیگیری با استفاده از مدل پیشنهادی نشان می‌دهد.



الف) برآورد تابع خطر تجمعی

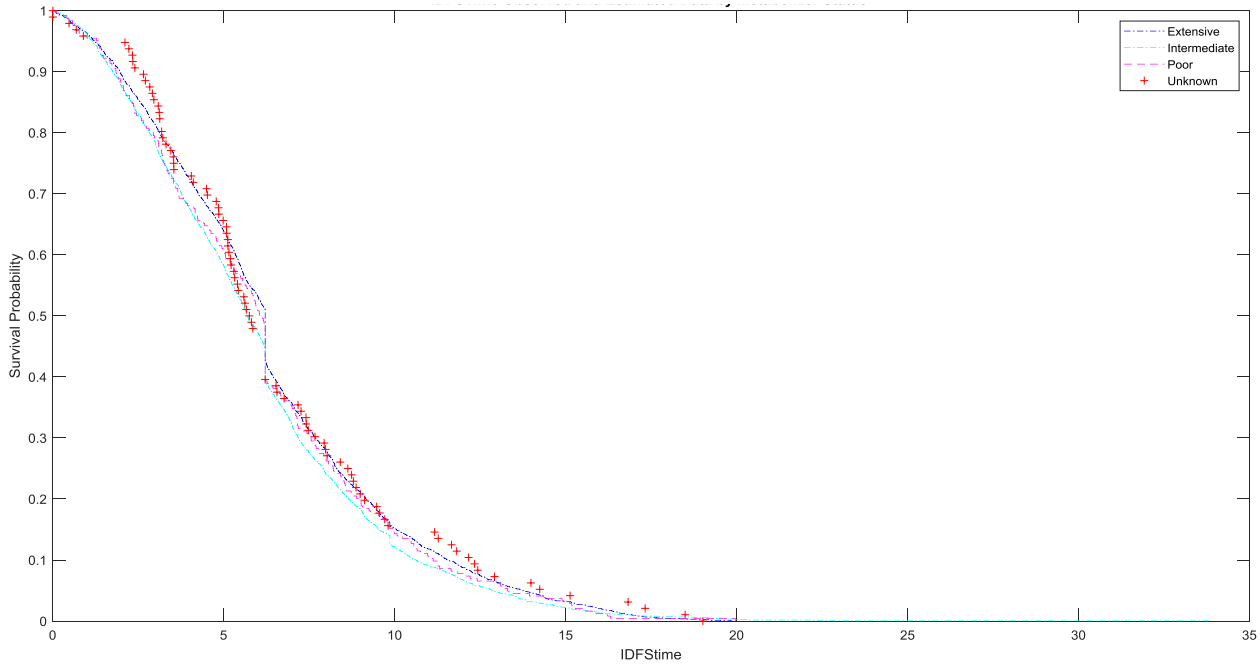


ب) توزیع احتمال برازش منحنی بقا

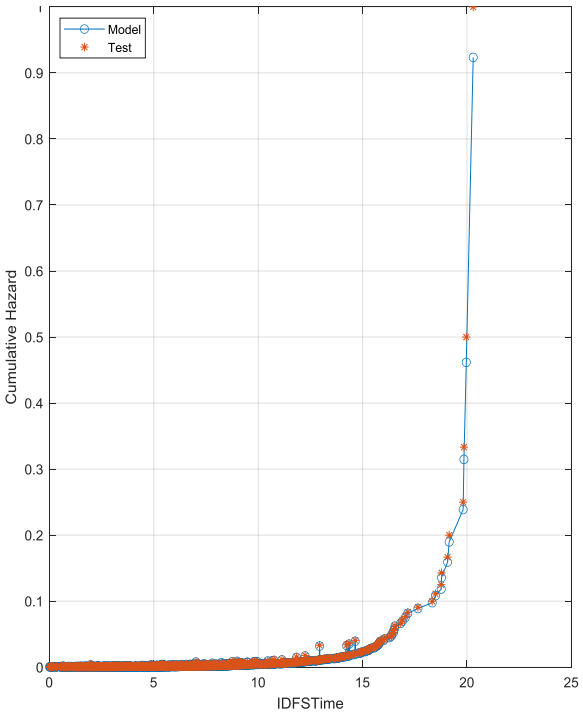
شکل (۴): الف) برآورد تابع خطر تجمعی Kaplan-Meier و ب) توزیع احتمال برازش منحنی بقا بر اساس IDFS در وضعیت‌های مختلف متابولیزه شدن تاموکسی‌فن به کمک Kaplan-Meier

این نکته جدا از تأثیرپذیری مصرف تاموکسی‌فن بر افزایش زمان بقا، وابستگی آن به اطلاعات بالینی و فارماکوژنومیکی بیماران را نیز نشان می‌دهد. چرا که بیماران با وضعیت‌های بالینی و فارماکوژنومیکی متفاوت، علی‌رغم دریافت دوز یکسان از داروی تاموکسی‌فن دارای زمان بقای متفاوتی بوده‌اند.

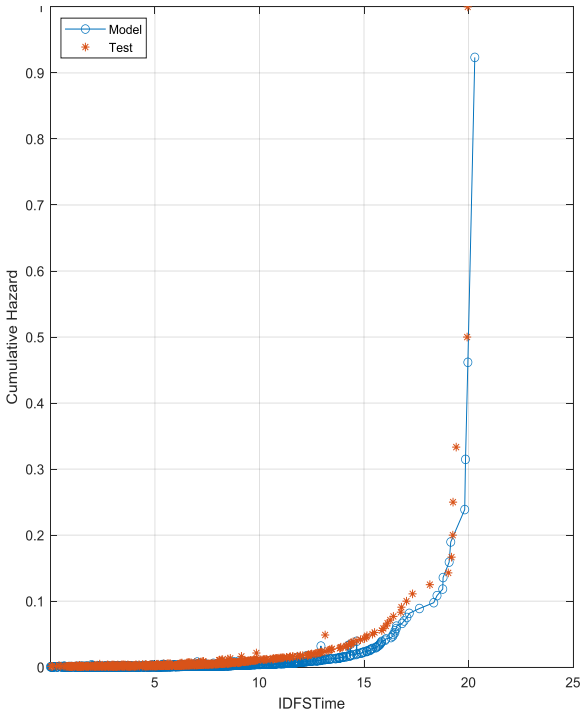
در شکل (۶) برازش منحنی توزیع خطر تجمعی برای زمان بقای بیماران مبتلا به سرطان پستان تهاجمی IDFS در وضعیت‌های مختلف متابولیزه تاموکسی‌فن با مدل پیشنهادی برای داده‌های آموزش و همچنین داده‌های آزمون مشخص شده است. مطابق شکل (۶)، دقت بالای مدل پیشنهادی در برازش منحنی توزیع خطر تجمعی بر اساس داده‌های آموزشی و همچنین داده‌های آزمون نشان داده شده است.



شکل (۵): مقایسه احتمال بقا تخمینی بیماران مبتلا به سرطان پستان تهاجمی IDFS در وضعیت‌های مختلف متابولیزه شدن تاموکسی‌فن تأثیر کند (POOR)، متوسط (INTERMEDIATE) و سریع (EXTENSIVE) در مدل پیشنهادی.



الف) برازش مدل برای داده‌های آموزشی



ب) برازش مدل برای داده‌هایی آزمون

شکل (۶): برازش منحنی توزیع خطر تجمعی برای زمان بقای بیماران مبتلا به سرطان پستان تهاجمی IDFS در وضعیت‌های مختلف متابولیزه شدن تاموکسی‌فن با مدل پیشنهادی

جدول (۲) پارامترهای ارزیابی را در تخمین زمان بقای بدون بیماری تهاجمی IDFS بر اساس وضعیت‌های مختلف متابولیسم تاموکسی‌فن با استفاده از مدل پیشنهادی و مدل‌های رایج تحلیل بقا، مدل AFT و مدل COX-PH نشان می‌دهد.

جدول (۲): مقایسه پارامترهای تخمین زمان بقا بدون بیماری در بیماران مبتلا به سرطان پستان تهاجمی بر اساس وضعیت‌های مختلف متابولیسم تاموکسی‌فن در مدل پیشنهادی با مدل‌های AFT و COX-PH

مدل بقا	زمان بقا	CI*	95%	SE**	Z-Scores
مدل تحلیل بقا COX-PH	زمان بقا بدون بیماری در بیماران مبتلا به سرطان پستان تهاجمی	۱/۴۰۴۳	۰/۷۹۲۶	۰/۱۴۵۹	۰/۳۶۷۲
		۱/۵۵۷۱	۰/۹۵۱۵	۰/۱۲۵۶	۱/۵۶۴۵
مدل تحلیل بقا AFT		۱/۳۱۸۲	۰/۸۰۵۱	۰/۱۲۵۸	۰/۲۳۶۷
		۱/۱۱۰۶	۰/۸۹۵۴	۰/۱۰۵۵	۰/۸۵۱۷
		۱/۳۱۰۱	۰/۸۹۱۴	۰/۱۱۱۳	۱/۱۱۸۱
		۱/۲۱۰۴	۰/۹۰۱۲	۰/۱۲۱۳	۰/۳۷۲۵
مدل تحلیل بقا پیشنهادی		۱/۱۲۳	۰/۸۹۳۲	۰/۰۸۹۴	۰/۲۱۷۲
		۱/۲۷۵۳	۰/۹۸۱۲	۰/۰۹۴۳	۰/۸۶۴۰
		۱/۱۸۹۴	۰/۸۶۳۱	۰/۰۹۷۸	۰/۰۹۴۳

* فاصله اطمینان ۹۵ درصد نسبت خطر

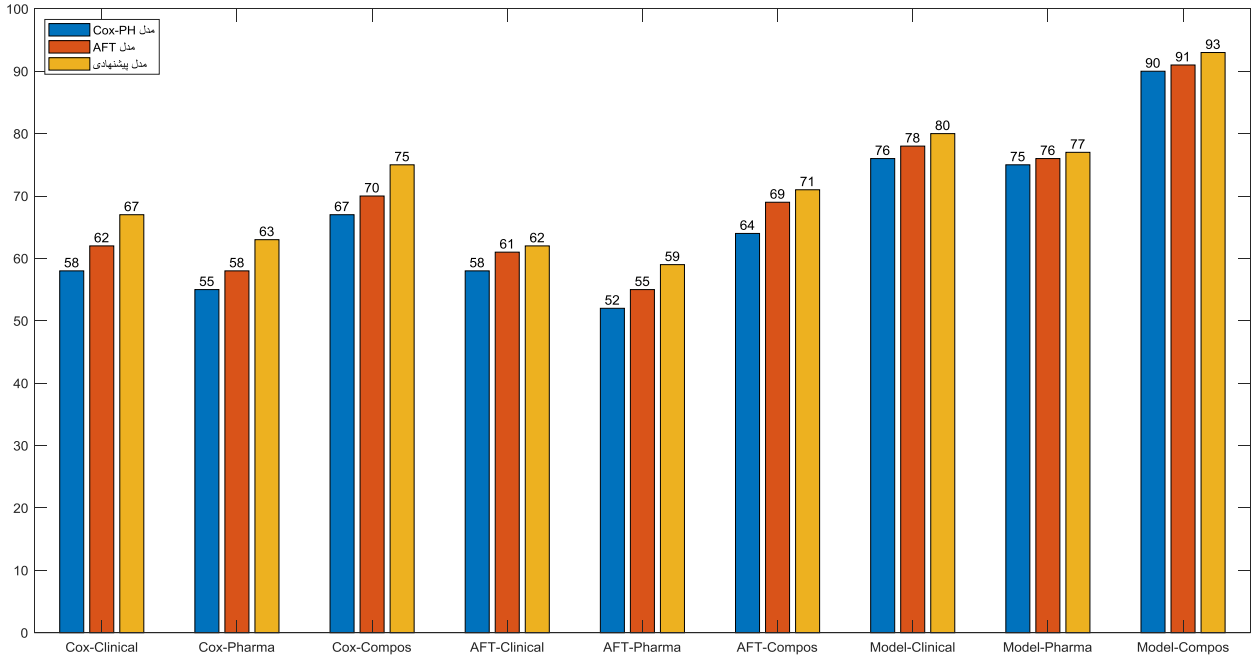
** خطای استاندارد پیش‌بینی زمان بقا

برای مقایسه دقت پارامترهای بقای تخمین زده شده در روش پیشنهادی با روش‌های دیگر، فاصله اطمینان ۹۵ درصد، بر اساس پارامترهای خروجی برآورد شده، برای هر مدل محاسبه شده است. برای انجام این کار، فاصله اطمینان (CI) ۹۵٪ نسبت خطر مطابق با رابطه (۶) محاسبه شد:

$$CI = exp(b + [-1,1] \times 1.96 \times se) \qquad (۶)$$

که در آن b تخمین ضرایب و se خطاهای استاندارد برآورد ضرایب است. مطابق با نتایج حاصل گزارش شده در جدول (۲)، خطای تخمین مدل پیشنهادی برای مجموعه داده‌های آزمایشی ۰/۰۹۷۸ است. این عدد برای مدل COX-PH برابر با ۰/۱۲۵۸ و برای مدل AFT برابر ۰/۱۲۱۳ است. مطابق با این نتایج، خطای تخمین مدل پیشنهادی برای تمام مجموعه‌های داده‌ای، در مقایسه با مدل‌های مورد بررسی به طور قابل‌توجهی کاهش پیدا کرده است.

مطابق با شکل (۷)، نتایج نشان می‌دهد که با به‌کارگیری مدل COX-PH تحلیل بقا در پیش‌بینی زمان بقا در مجموعه داده‌های آموزشی، هنگامی که ترکیبی از ویژگی‌های بالینی و فارماکوژنومیکی استفاده شد، دقت پیش‌بینی ۸ درصد بیشتر از زمانی بود که فقط از ویژگی‌های بالینی استفاده شده است.



شکل (۷) : مقایسه دقت تخمین زمان بقا بیماران مبتلا به سرطان پستان تهاجمی (بر حسب درصد) در مدل پیشنهادی در مقایسه با سایر مدل‌ها برای داده‌های بالینی و فارماکوژنومیکی و ترکیب ویژگی‌ها

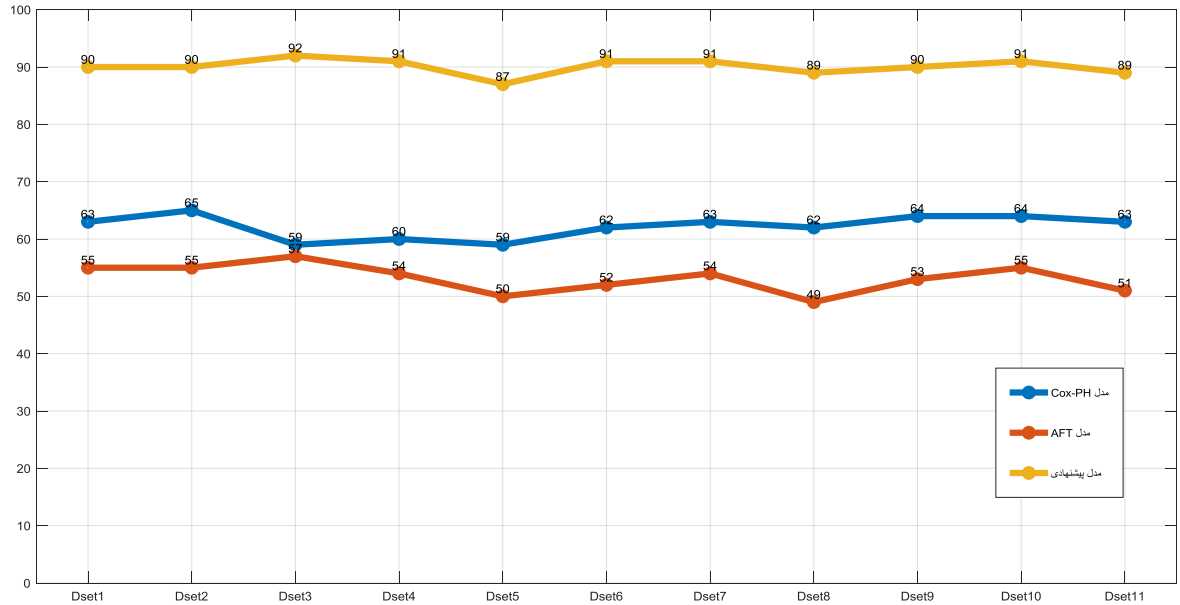
همچنین، دقت پیش‌بینی ۱۲ درصد بیشتر از زمانی بود که فقط ویژگی‌های فارماکوژنومیکی در مطالعه استفاده شده است. این عدد برای داده‌های آزمون بترتیب برابر ۹ و ۱۲ درصد است.

مطابق این شکل، با به‌کارگیری مدل AFT در پیش‌بینی زمان بقا در مجموعه داده‌های آموزشی، هنگامی که ترکیبی از ویژگی‌های بالینی و فارماکوژنومیکی استفاده شد، دقت پیش‌بینی ۹ درصد بیشتر از زمانی بود که فقط از ویژگی‌های بالینی استفاده شده است. همچنین، دقت پیش‌بینی نیز ۱۲ درصد بیشتر از زمانی بود که فقط ویژگی‌های فارماکوژنومیکی در مطالعه استفاده شده است. این عدد برای داده‌های آزمون بترتیب برابر ۶ و ۱۲ درصد است. این نتایج نشان می‌دهد که با به‌کارگیری مدل پیشنهادی تحلیل بقا در پیش‌بینی زمان بقا در مجموعه‌داده‌های آموزشی، هنگامی که ترکیبی از ویژگی‌های بالینی و فارماکوژنومیکی استفاده شد، دقت پیش‌بینی ۱۳ درصد بیشتر از زمانی بود که فقط از ویژگی‌های بالینی استفاده شده است. همچنین، دقت پیش‌بینی ۱۶ درصد بیشتر از زمانی بود که فقط ویژگی‌های فارماکوژنومیکی در مطالعه استفاده شده است. این عدد برای داده‌های آزمون بترتیب برابر ۱۴ و ۱۵ درصد است.

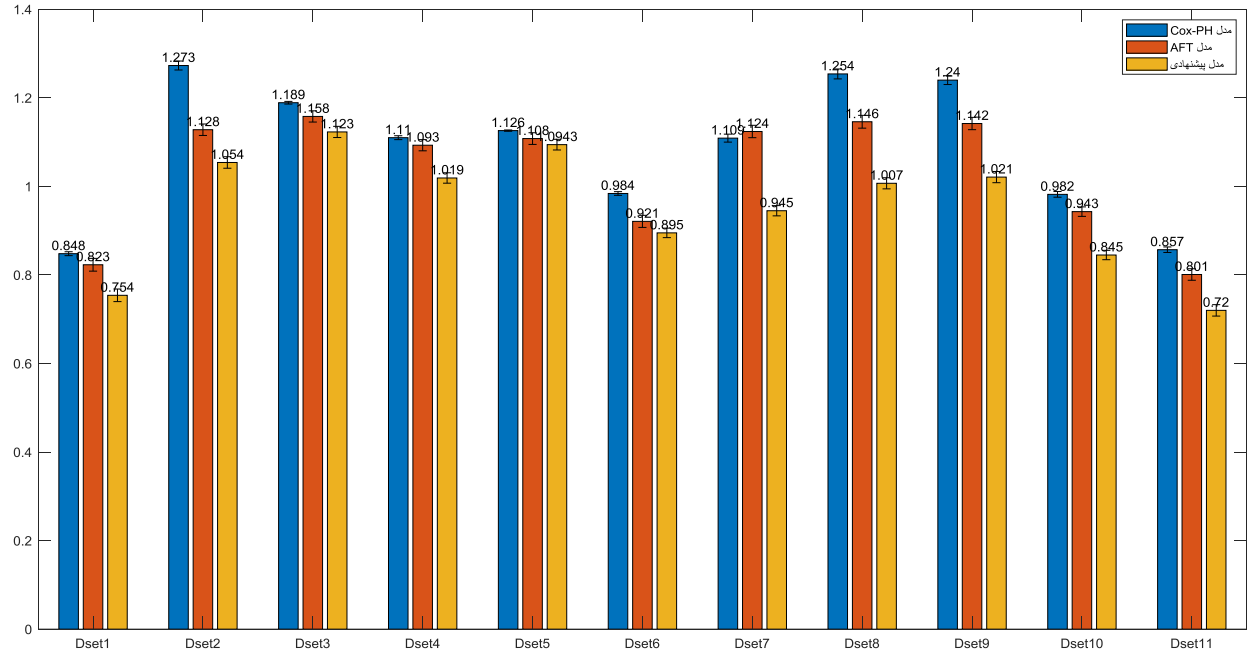
مطابق شکل (۸)، نتایج نشان می‌دهد که مدل پیشنهادی تحلیل بقا، زمان بقای عاری از بیماری سرطان پستان تهاجمی را در مقایسه با مدل‌های تحلیل بقا Cox-PH و AFT در همه سایت‌های مورد مطالعه با دقت بالاتری پیش‌بینی کرده است.

در این ارزیابی دقت تخمین مدل پیشنهادی به طور میانگین ۸۹ درصد بوده است که در مقایسه با مدل Cox-PH، بادقت ۶۲/۲ درصد و در مقایسه با مدل تحلیل بقا AFT با ۵۲ درصد، به ترتیب دارای ۲۶/۸ و ۳۷ درصد دقت بالاتری در پیش‌بینی زمان بقای عاری از بیماری سرطان پستان تهاجمی بوده است.

در شکل (۹) مقایسه نسبت خطر در بیماران مبتلا به سرطان پستان تهاجمی در مدل پیشنهادی در مقایسه با سایر مدل‌ها برای مجموعه داده‌های مختلف به همراه خطای استاندارد نمایش‌داده شده است. در این بررسی نسبت خطر در بیماران مبتلا به سرطان پستان تهاجمی در مدل پیشنهادی در مقایسه با مدل Cox-PH و مدل تحلیل بقا AFT، برای کلیه سایت‌های مورد مطالعه نشان داده شده است. نسبت خطر برای مجموعه داده‌ای آخر برای مدل Cox-PH برابر با ۰/۸۵۷ و برای مدل تحلیل بقا AFT برابر با ۰/۸۰۱ و با استفاده از مدل پیشنهادی برابر با ۰/۷۲۰ محاسبه شده است. با بررسی نسبت خطر در تمام سایت‌ها، نتایج نشان می‌دهد مدل پیشنهادی در مجموعه داده‌های مختلف با خطای کمتری، نسبت خطر را محاسبه کرده است.



شکل (۸): مقایسه پارامترهای تخمین زمان بقا بدون بیماری در بیماران مبتلا به سرطان پستان تهاجمی بر اساس وضعیت متابولیزه (بر حسب درصد) در مدل پیشنهادی با مدل‌های دیگر در مجموعه داده‌ها



شکل ۹ : مقایسه نسبت خطر در بیماران مبتلا به سرطان پستان تهاجمی* در مدل پیشنهادی در مقایسه با سایر مدل‌ها برای مجموعه‌داده‌های مختلف به همراه خطای استاندارد در مجموعه داده‌ها

* زمان بقا بدون بیماری در بیماران مبتلا به سرطان پستان تهاجمی (IDFS)، تعداد روز از تشخیص اولیه تا آخرین ارزیابی بالینی وضعیت بیماری است.

۷. بحث و نتیجه‌گیری

سرطان پستان یکی از شایع‌ترین انواع سرطان است که هر ساله باعث مرگ و میر فراوانی در بین زنان و مردان می‌شود و علیرغم پیشرفت‌های بسیاری که در مورد تشخیص زودهنگام و درمان مناسب این بیماری صورت گرفته است، کماکان سردسته علل مرگ به علت سرطان در بین زنان است. در سال‌های اخیر، مطالعات و تجزیه و تحلیل زمان بقای بیماران در تحقیقات سرطان پستان و روند درمان بسیار جالب بوده است. نتایج پژوهش حاضر نشان می‌دهد که با استفاده از داده‌های ژنتیکی و فارماکوژنومیکی بیماران مبتلا به سرطان پستان که از داروی تاموکسی‌فن در فرایند درمان استفاده کرده‌اند، می‌توان خطر نسبی را تخمین زد و همچنین ارتباط بین فنوتیپ و ژنوتیپ را با دقت بیشتری تعیین کرد. در همه مجموعه داده‌های سرطان، مقادیر داده‌های برچسب‌گذاری شده و گمشده وجود دارد، اما بسیاری از محققان از داده‌های مقادیر گمشده در فرایند تحلیل استفاده نمی‌کنند. در مدل پیشنهادی، یک مدل ترکیبی از مدل Cox-PH، مدل AFT و روش یادگیری نیمه‌نظارتی مبتنی بر گراف برای تجزیه و تحلیل بقا پیشنهاد شد. با به‌کارگیری داده‌های بالینی و فارماکوژنومیکی و با استفاده از روش یادگیری ماشین نیمه‌نظارتی، مقدار داده‌های از دست رفته در مجموعه داده‌های تعیین، سپس نرخ خطر و سایر پارامترهای بقا برآورد شده است. نتایج این تحقیق نشان داد که با استفاده از داده‌های برچسب‌گذاری شده محدود و داده‌های از دست رفته، می‌توان آنالیز بقای بیماران سرطان پستان را با دقت بیشتری انجام داد. مقایسه مدل پیشنهادی با سایر مدل‌های قبلی در تجزیه و تحلیل بقا نشان می‌دهد که با تخمین ارزش داده‌های از دست رفته و با استفاده از مدل‌های بقای پیشنهادی، مدل پیشنهادی زمان بقای عاری از بیماری تهاجمی پستان را با دقت بالاتری پیش‌بینی می‌کند. هنگامی که ترکیبی از ویژگی‌های بالینی و فارماکوژنومیکی در روند تحلیل داده‌های سرطان پستان تهاجمی استفاده شد، به کمک مدل پیشنهادی دقت طبقه‌بندی ۱۴ درصد بیشتر از زمانی بود که فقط از ویژگی‌های بالینی استفاده شد و دقت طبقه‌بندی ۱۵ درصد بیشتر از زمانی بود که فقط ویژگی‌های فارماکوژنومیکی در آزمایش استفاده شده است.

تعارض منافع: نویسندگان اعلام می‌کنند که هیچ تعارض منافعی ندارند.

مراجع

[1] Howell A., Sims A. H., Ong K. R., Harvie M. N., Evans D. G. R., and Clarke R. B., "Mechanisms of Disease: prediction and prevention of breast cancer--cellular and molecular interactions", Nat Clin Pract Oncol, 2(12):635-646, 2005, doi:10.1038/ncponc0361.

[2] Mego M., Mani S.A., and Cristofanilli M., "Molecular mechanisms of metastasis in breast cancer—clinical applications". Nature reviews Clinical oncology, 7(12): 693-701, 2010, doi:10.1038/nrclinonc.2010.171.

[3] Sharma G.N., Dave R., Sanadya J., Sharma P., and Sharma K. K., "Various types and management of breast cancer: an overview", Journal of advanced pharmaceutical technology and research, 1(2): 109-126, 2010.

[4] Dean L., Pratt V. M., Scott S. A., Pirmohamed M., Esquivel B., Kane M. S., Kattman B. L., and Malheiro A. J., "Tamoxifen Therapy and CYP2D6 Genotype", in National Center for Biotechnology Information (US), 2014.

[5] Schultink A. H. M. V., Zwart W., Linn S. C., Beijnen J. H., and Huitema A. D. R., "Effects of Pharmacogenetics on the Pharmacokinetics and Pharmacodynamics of Tamoxifen", Clinical Pharmacokinetics, 54(8): 797-810, 2015, doi:10.1007/s40262-015-0273-3.

[6] George B., Seals S., and Aban I., "Survival analysis and regression models", J. Nucl. Cardiol, 21(4): 686-694, 2014, doi:10.1007/s12350-014-9908-2.

[7] Austin P. C., Lee D. S., and Fine J. P., "Introduction to the Analysis of Survival Data in the Presence of Competing Risks", Circulation, 133(6): 601-609, 2016, doi:10.1161/CIRCULATIONAHA.115.017719.

[8] Prinja S., Gupta N., and Verma R., "Censoring in clinical trials: review of survival analysis techniques", Indian journal of community medicine: official publication of Indian Association of Preventive and Social Medicine, 35(2): 217–221, 2010, doi:10.4103/0970-0218.66859.

[9] Stel V. S., Dekker F. W., Tripepi G., Zoccali C., and Jager K. J., "Survival Analysis I: The Kaplan-Meier Method", Nephron Clinical Practice, 119(1): c83-c88, 2011, doi:10.1159/000324758.

[10] Kleinbaum D.G. and Klein M., "Survival analysis", Vol. 3, Springer, 2010.

[11] Zare, A., Hosseini M., Mahmoodi M., Mohammad K., Zeraati H., Holakouie-Naieni K., "A Comparison between Accelerated Failure-time and Cox Proportional Hazard Models in Analyzing the Survival of Gastric Cancer Patients", Iranian Journal of Public Health, 44(8): 1095-1102, 2015, doi:10.7314/APJCP.2015.16.18.8567.

[12] Zhu X., "Semi-supervised learning literature survey", 2005, http://digital.library.wisc.edu/1793/60444.

[۱۳] صادق زاده ن.، شمسی م.، رسولی کناری ع.، «حاشیه‌نویسی تصویر با استفاده از الگوریتم خوشه‌بندی نیمه‌نظارتی طیفی»، مجله محاسبات نرم، جلد ۳، شماره ۱، ص ۳۵-۲۰، ۱۳۹۳.

[14] Zhu X. and Goldberg A. B., "Introduction to Semi-Supervised Learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, 3(1): 1-130, 2009, doi:10.2200/S00196ED1V01Y200906AIM006.

[15] Subramanya A. and Talukdar P. P., "Graph-based semi-supervised learning", Synthesis Lectures on Artificial Intelligence and Machine Learning, 8(4): 1-125, 2014.

[16] Zhou D., Bousquet O., Lal T. N., Weston J., and Schölkopf B., "Learning with local and global consistency", MIT Press, pp. 321-328, 2003.

[17] Jiang X., Xue D., Brufsky A., Khan S., and Neapolitan R., "A New Method for Predicting Patient Survivorship Using Efficient Bayesian Network Learning", Cancer Informatics,13: CIN.S13053, 2014, doi:10.4137/CIN.S13053.

[18] Bashiri A., Ghazisaeedi M., Safdari R., Shahmoradi L., and Ehtesham H., "Improving the Prediction of Survival in Cancer Patients by Using Machine Learning Techniques: Experience of Gene Expression Data: A Narrative Review", Iranian Journal of Public Health, 46(2): 165-172, 2017.

[19] Endo A., Shibata T., and Tanaka H., "Comparison of seven algorithms to predict breast Cancer survival", International Journal of Biomedical Soft Computing and Human Sciences: the official journal of the Biomedical Fuzzy Systems Association, 13(2):11-16, 2008, doi:10.24466/ijbschs.13.2_11.

[20] Bagherian H., Javanmard S. H., Sharifi M., and Sattari M., "Using data mining techniques for predicting the survival rateof breast cancer patients: a review article", Tehran University Medical Journal, 79(3): 176-186, 2021.

[21] Ghasemi F., Rasekhi A., and Haghighat S., "Analyzing the Survival of Breast Cancer Patients Using Weibull and Poisson Beta-Weibull Non-Mixture Cure Models", Pejouhesh dar Pezeshki (Research in Medicine), 42(4): 236-242, 2018.

[22] Sarkar K., Chowdhury R., and Dasgupta A., "Analysis of Survival Data: Challenges and Algorithm-Based Model Selection", Journal of Clinical and Diagnostic Research: JCDR, 11(6): LC14-LC20, 2017.

[23] Kiani B. and Atashi A., "A Prognostic Model Based on Data Mining Techniques to Predict Breast Cancer Recurrence", Journal of Health and Biomedical Informatics, 1(1): 26-31, 2014.

tamoxifen therapy for breast cancer", Pharmacogenomics, 16(3): 287-296, 2015, doi:10.2217/pgs.14.171.

[36] Afshar H. L., Ahmadi M., Roudbari M., and Sadoughi F., "Prediction of breast cancer survival through knowledge discovery in databases", Global journal of health science, 7(4): 392-398, 2015, doi: 10.5539/gjhs.v7n4p392.

[37] Saadatmand S. H. A., Sadeghi A., and Mohaghghegh F., "Study on association of Single Nucleotide Polymorphism in ESRα Gene rs2234693 With Breast Cancer in Markazi Province", Arak Medical University Journal (AMUJ), 17(12) 32-38, 2015.

[38] Lei L., Wang X., Wu X.-D., Wang Z., Chen Z.-H., Zheng Y.-B., and Wang X.-J., "Association of CYP2D6*10 (c.100C>T) polymorphisms with clinical outcome of breast cancer after tamoxifen adjuvant endocrine therapy in Chinese population", American Journal of Translational Research, 8(8): 3585-3592, 2016.

[39] Charoenchokthavee W., Panomvana D., Sriuranpong V., and Areepium N., "Prevalence of CYP2D6*2, CYP2D6*4, CYP2D6*10, and CYP3A5*3 in Thai breast cancer patients undergoing tamoxifen treatment", Breast Cancer: Targets and Therapy, 8: 149-155, 2016, doi:10.2147/BCTT.S105563.

[40] Khalkhali H. R., Afshar H. L., Esnaashari O., and Jabbari N., "Applying Data Mining Techniques to Extract Hidden Patterns about Breast Cancer Survival in an Iranian Cohort Study", Journal of research in health sciences, 16(1): 31-35, 2016.

[41] Damkier P., Kjærsgaard A., Barker K. A., Cronin-Fenton D., Crawford A., Hellberg Y., Janssen E. A. M., Langefeld C., Ahern T. P., and Lash T. L., "CYP2C19*2 and CYP2C19*17 variants and effect of tamoxifen on breast cancer recurrence: Analysis of the International Tamoxifen Pharmacogenomics Consortium dataset", Scientific Reports, 7(1): 1-8, 2017, doi:10.1038/s41598-017-08091-x.

[42] Wang L., Qian Q., Zhang Q., Wang J., Cheng W., and Yan W., "Classification Model on Big Data in Medical Diagnosis Based on Semi-Supervised Learning", The Computer Journal, 65(2): 177-191, 2020, doi: 10.1093/comjnl/bxaa006.

[43] Afshar H. L., Jabbari N., Khalkhali H. R., and Esnaashari O., "Prediction of breast cancer survival by machine learning methods: An application of multiple imputation", Iranian Journal of Public Health, 50(3): 598-605, 2021, doi:10.18502/ijph.v50i3.5606.

[44] Sanchez-Spitman A. B., Swen J. J., Dezentjé V. O., Moes D. J. A. R., Gelderblom H., and Guchelaar H. J., "Effect of CYP2C19 genotypes on tamoxifen metabolism and early-breast cancer relapse", Scientific Reports, 11(1): 415, 2021, doi:10.1038/s41598-020-79972-x.

[45] Mulder T. A. M., de With M., Del Re M., Danesi R., Mathijssen R. H. J., and van Schaik R. H. N., "Clinical CYP2D6 Genotyping to Personalize Adjuvant Tamoxifen Treatment in ER-Positive Breast Cancer Patients: Current

[24] Sadeghi S. and Golabpour A., "An Algorithm for Predicting Recurrence of Breast Cancer Using Genetic Algorithm and Nearest Neighbor Algorithm", Journal of Health and Biomedical Informatics, 6(4): 309-319, 2020.

[25] Moller P., Evans D. G., Reis M. M., Gregory H., Anderson E., Maehle L., Lalloo F., Howell A., Apold J., Clark N., Lucassen A., and Steel C. M., "Surveillance for familial breast cancer: Differences in outcome according to BRCA mutation status", Int. J. Cancer, 121(5): 1017-20, 2007, doi:10.1002/ijc.22789.

[26] Liang Y., Chai H., Liu X.-Y., Xu Z.-B., Zhang H., and Leung K.-S., "Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with L(1/2) regularization", BMC Medical Genomics, 9(1):1-11, 2016, doi:10.1186/s12920-016-0169-6.

[27] Chai H., Li Z.-N., Meng D.-Y., Xia L.-Y., and Liang Y., "A new semi-supervised learning model combined with Cox and SP-AFT models in cancer survival analysis", Sci. Rep., 7(1): 13053-13065, 2017, doi:10.1038/s41598-017-13133-5.

[۲۸] کی‌پور ا.، براری م.، شیرازی ح.، «پیشگویی پیوند در شبکه‌های اجتماعی با استفاده از ترکیب دسته‌بندی‌کننده‌ها»، مجله محاسبات نرم، جلد ٤، شماره ۲، ص ۱۷-۲، ۱۳۹٤.

[29] Tharwat A., "Classification assessment methods", Applied Computing and Informatics, 17(1): 168-192, 2020, doi:10.1016/j.aci.2018.08.003.

[۳۰] ویسی ه.، قایدشرف ح.ر.، ابراهیمی م.، «بهبود کارایی الگوریتم‌های یادگیری ماشین در تشخیص بیماری‌های قلبی با بهینه‌سازی داده‌ها و ویژگی‌ها»، مجله محاسبات نرم، جلد ۸، شماره ۱، ص ۸۵-۷۰، ۱۳۹۸.

[31] Delen D., Walker G., and Kadam A., "Predicting breast cancer survivability: a comparison of three data mining methods", Artificial Intelligence in Medicine, 34(2): 113-127, 2005, doi:10.1016/j.artmed.2004.07.002.

[32] Kiyotani K., Mushiroda T., Sasa M., Bando Y., Sumitomo I., Hosono N., Kubo M., Nakamura Y., and Zembutsu H., "Impact of CYP2D6*10 on recurrence-free survival in breast cancer patients receiving adjuvant tamoxifen therapy", Cancer Science, 99(5): 995-999, 2008, doi:10.1111/j.1349-7006.2008.00780.x.

[33] Brauch H. and Schwab M., "Prediction of tamoxifen outcome by genetic variation of CYP2D6 in post-menopausal women with early breast cancer", British journal of clinical pharmacology, 77(4): 695-703, 2014, doi:10.1111/bcp.12229.

[34] Province M.A., et al., "CYP2D6 Genotype and Adjuvant Tamoxifen: Meta-Analysis of Heterogeneous Study Populations", Clinical Pharmacology and Therapeutics, 95(2): 216-227, 2014, doi:10.1038/clpt.2013.186.

[35] Zembutsu H., "Pharmacogenomics toward personalized

Status of a Controversy", Cancers, 13(4): 771, 2021, doi:10.3390/cancers13040771.

[46] Al-Azzam N. and Shatnawi I., "Comparing supervised and semi-supervised Machine Learning Models on Diagnosing Breast Cancer", Annals of Medicine and Surgery, 62: 53-64, 2021, doi:10.1016/j.amsu.2020.12.043.

[47] El Shawi R., Kilanava K., and Sakr S., "An Interpretable Semi-Supervised Framework for Patch-Based Classification of Breast Cancer", Research Square, 2022, doi:10.21203/rs.3.rs-1343955/v1.

[48] Teimouri-Yansari R., Mirzarezaee M., Sadeghi M., and Nadjar-Araabi B., "A New Survival Analysis Model in Adjuvant Tamoxifen-Treated Breast Cancer Patients Using Manifold-based Semi-Supervised Learning", Journal of Computational Science, 61:101645, 2022, doi:10.1016/j.jocs.2022.101645.

[49] Lin R.-H., Lin C.-S., Chuang C.-L., Kujabi B. K., and Chen Y.-C., "Breast Cancer Survival Analysis Model", Applied Sciences, 12(4): 1971, 2022, doi:10.3390/app12041971.

[50] Xiao J., Mo M., Wang Z., Zhou C., Shen J., Yuan J., He Y., and Zheng Y., "The Application and Comparison of Machine Learning Models for the Prediction of Breast Cancer Prognosis: Retrospective Cohort Study", JMIR Medical Informatics, 10(2) :e33440, 2022, doi: 10.2196/33440.

پیوست (۱): خلاصه مهم‌ترین پژوهش‌ها در حوزه تحلیل بقا سرطان پستان

مرجع	سال	عنوان تحقیق	داده‌ها	مزایا و معایب روش
[۳۱]	۲۰۰۵	پیش‌بینی سرطان پستان: مقایسه سه روش داده کاوی	بیش از ۲۰۰۰۰۰ نمونه	– استفاده از مدل ترکیبی شبکه‌های عصبی، درخت تصمیم و رگرسیون لجستیک – درخت تصمیم بهترین پیش‌بینی‌کننده بادقت ۹۳/۶٪ شبکه‌های عصبی مصنوعی بادقت ۹۱/۲٪ و مدل رگرسیون لجستیک بدترین بادقت ۸۹/۲٪
[۲۵]	۲۰۰۷	بررسی بقای سرطان پستان خانوادگی: تفاوت نتایج بر اساس وضعیت جهش ژن BRCA1 و BRCA2	۴۴۲ بیمار مبتلا به سرطان پستان	– استفاده از مدل بقای کاپلان مایر – بقای ۵ ساله برای بیماران دارای جهش در ژن BRCA1 برابر ۷۲٪ و برای بیماران دارای جهش در BRCA2 برابر با ۹۶٪
[۳۲]	۲۰۰۸	تاثیر CYP2D6*10 بر بقای بدون عود در بیماران مبتلا به سرطان پستان دریافت کننده درمان کمکی تاموکسی‌فن.	۱۷۶۴ بیمار مبتلا به سرطان پستان در کلینیک مراقبت پستان توکوشیما	– تجزیه و تحلیل بقای بدون عود توسط ژنوتیپ CYP2D6 با استفاده از روش‌های Kaplan-Meier – بررسی رابطه بین پلی‌مورفیسم ژنتیکی با استفاده از آزمون log-rank – تجزیه و تحلیل خطر متناسب کاکس برای شناسایی عوامل بالینی مهم پیش آگهی
[۳۳]	۲۰۱۴	پیش‌بینی پیامد تاموکسیفن با تنوع ژنتیکی CYP2D6 در زنان یائسه مبتلا به سرطان پستان اولیه	–	– بررسی تاثیر پلی‌مورفیسم‌های ژنتیکی CYP2D6 بر نتیجه درمان بیماران مبتلا به سرطان پستان مثبت گیرنده استروژن پس از یائسگی (ER) – بررسی پیش‌بینی‌کننده بالقوه CYP2D6 بر پیامد تاموکسی‌فن – ارتباط بین ژنوتیپ برای CYP2D6 از نظر بالینی در زنان یائسه

ادامه پیوست (۱): خلاصه مهم‌ترین پژوهش‌ها در حوزه تحلیل بقا سرطان پستان

مرجع	سال	عنوان تحقیق	داده‌ها	مزایا و معایب روش
[۳۴]	۲۰۱۴	ژنوتیپ CYP2D6 و درمان کمکی تاموکسی‌فن: فراتحلیل مطالعه ناهمگن جمعیت ها	ITPC - کنسرسیوم بین‌المللی تاموکسی‌فن فارماکوژنومیکس ۴۹۷۳ بیمار	– استفاده از مدل اثرات تصادفی یا اجزاء خطا در فراتحلیل آماری و بررسی سازگاری نتایج در بین مجموعه‌های داده‌ای – CYP2D6 یک پیش‌بینی‌کننده قوی در کند کردن حالت تهاجمی بیماری است
[۱۷]	۲۰۱۴	یک روش جدید برای پیش بینی بروز بقاء بیمار با استفاده از شبکه یادگیری بیزی کارا	۸۹۹ بیمار مبتلا به سرطان پستان از پایگاه داده TCGA	– استفاده از شبکه یادگیری بیزی – مدل پیشنهادی از مدل خطر متناسب Cox بهترعمل کرده و قابل مقایسه با روش جنگل بقا تصادفی است
[۳۵]	۲۰۱۵	فارماکوژنومیکس در جهت شخصی سازی: درمان تاموکسی‌فن برای سرطان پستان	–	– CYP2D6 به عنوان یک آنزیم کلیدی برای تولید یکی از متابولیت‌های قوی تاموکسی‌فن و اندوکسی‌فن است.
[۳۶]	۲۰۱۵	پیش بینی بقاء سرطان پستان از طریق کشف دانش در پایگاه‌های داده	۲۲۷۶۳ بیمار	– استفاده از ماشین بردار پشتیبان و شبکه بیزین – SVM دارای بالاترین عملکرد با دقت (۹۶/۷٪)
[۳۷]	۲۰۱۵	بررسی همراهی پلی‌مورفیسم تک نوکلئوتیدی در ژن گیرنده آلفای استروژن با سرطان پستان	۲۹۲ نمونه مبتلا به سرطان پستان و سالم در استان مرکزی	– استفاده از آزمون کای اسکوئر آماری – برابری ژنوتیپ‌های CT و TT نسبت به CC در جایگاه rs2234693:C/T به ترتیب افزایش خطر ۱/۵ و ۵/۵
[۳۸]	۲۰۱۶	ارتباط پلی‌مورفیسم CYP2D6 با نتیجه بالینی سرطان پستان پس از درمان کمکی تاموکسی‌فن در جمعیت چینی	پلی‌مورفیسم ۷۲ بیمار تحت درمان تاموکسی‌فن	– استفاده از مدل تحلیل بقا کوکس و کاپلان مایر – نتایج نشان داد ژنوتیپ T/T یک عامل پیش‌آگهی منفی در تشخیص زودهنگام سرطان پستان است
[۳۹]	۲۰۱۶	شیوع CYP2D6*2، CYP2D6*4 CYP2D6*10و CYP3A5*3 در بیماران مبتلا به سرطان پستان تایلندی تحت درمان با تاموکسی‌فن	۱۳۴ بیمار مبتلا به سرطان پستان تایلندی	– استفاده از روش زنجیره‌ای پلیمراز بلادرنگ – ناقص بودن عملکرد حداقل یک آلل ۹۸ درصد بیماران، تحت درمان تاموکسی‌فن
[۴۰]	۲۰۱۶	بکارگیری تکنیک های داده کاوی برای استخراج الگوهای پنهان در مورد بقای سرطان پستان در یک مطالعه گروهی ایرانی	پایگاه داده سرطان پستان حاوی اطلاعات ۵۶۹ بیمار در سال‌های ۲۰۰۷-۲۰۱۰	– استخراج قوانین پنهان مفیدی از مجموعه داده‌هایی با اندازه نسبتا کوچک – استفاده از ساختار درخت تصمیم

ادامه پیوست (۱): خلاصه مهم‌ترین پژوهش‌ها در حوزه تحلیل بقا سرطان پستان

مرجع	سال	عنوان تحقیق	داده‌ها	مزایا و معایب روش
[۲۶]	۲۰۱۶	تجزیه و تحلیل بقا سرطان با استفاده از روش نیمه نظارت بر یادگیری مبتنی بر مدل‌های Cox و AFT با تنظیم L(1/2)	eal microarray datasets	– استفاده از یادگیری نیمه نظارتی مبتنی بر مدل‌های کوکس و زمان شکست شتاب‌دار – بهبود پیش‌بینی عملکرد مدل‌های AFM و Cox در تحلیل بقا
[۴۱]	۲۰۱۷	تأثیر تاموکسی‌فن بر عود سرطان پستان: تجزیه و تحلیل مجموعه داده های کنسرسیوم فارماکوژنومیکس بین المللی تاموکسی‌فن	ITPC - کنسرسیوم بین‌المللی تاموکسی فن فارماکوژنومیکس	– نتایج تحقیق نشان داد درمان با تاموکسی‌فن خطر عود سرطان پستان را برای حداقل ۱۵ سال کاهش می‌دهد.
[۱۸]	۲۰۱۷	بهبود پیش‌بینی بقا در بیماران سرطانی با استفاده از تکنیک‌های یادگیری ماشین	به کمک داده‌های بیان ژن	– افزایش کارایی پیش‌بینی بقا از سرطان با استفاده از شبکه‌های عصبی مصنوعی – دقت بالا و کارایی داده‌های بیان ژن در مقایسه با داده‌های بالینی
[۲۷]	۲۰۱۷	یک مدل یادگیری نیمه نظارت شده همراه با مدل‌های Cox و SP-AFT در تجزیه‌وتحلیل بقای سرطان	۵۹۰ نمونه بیمار سرطانی	– استفاده بیشتر از نمونه‌های سانسور شده در مدل ترکیبی پیشنهادی Cox-SP-AFT – برآورد زمان بقا بادقت بیشتری
[۴۲]	۲۰۲۰	مدل طبقه‌بندی کلان‌داده در تشخیص پزشکی بر پایه یادگیری نیمه نظارتی	بیماری عروق کرونر قلب (CHD)	– استفاده از الگوریتم نیمه نظارتی بهبودیافته Self-training و Cooperative Training – ارائه یک مدل طبقه‌بندی مبتنی بر الگوریتم یادگیری نیمه نظارت شده با استفاده از داده‌های برچسب‌دار و بدون برچسب برای پردازش کلان‌داده‌ها – عملکرد مناسب مدل طبقه‌بندی داده‌های تشخیص پزشکی پیشنهادی بر اساس الگوریتم یادگیری نیمه نظارت شده
[۴۳]	۲۰۲۱	پیش‌بینی بقای سرطان پستان با روش‌های یادگیری ماشین	داده های ۸۵۶ بیمار مرکز تحقیقات و درمان امید ارومیه، ۲۰۰۶ تا ۲۰۱۲	– استفاده از الگوریتم‌های C5 و هرس افزایشی مکرر برای پیش‌بینی و استخراج قوانین بالینی – بهبود عملکرد الگوریتم C5 در تمامی معیارهای ارزیابی
[۴۴]	۲۰۲۱	اثر ژنوتیپ های CYP2C19 بر متابولیسم تاموکسی‌فن و عود زودرس سرطان پستان	ITPC - کنسرسیوم بین‌المللی تاموکسی فن فارماکوژنومیکس	– استفاده ازآزمون آنالیز واریانس (ANOVA) برای ارزیابی تفاوت بین گروه‌ها – استفاده ازتجزیه و تحلیل رگرسیون خطی چندگانه برای بررسی سهم ژنوتیپ‌های CYP2C19 – هیچ ارتباطی بین فعالیت تاموکسی‌فن و ژنوتیپ‌های CYP2D6 و CYP2C19 دیده نشد.
[۴۵]	۲۰۲۱	ژنوتیپ بالینی CYP2D6 برای شخصی‌سازی درمان کمکی تاموکسی‌فن در بیماران مبتلا به سرطان پستان با ER مثبت: بحث به روز– مقاله مروری	–	– مرور آخرین پیشرفت‌ها در تحلیل رابطه ژنوتیپ CYP2D6 با غلظت پلاسمایی اندوکسی‌فن و پیامد بالینی مرتبط با تاموکسی‌فن. – تا انجام کارآزمایی‌های کنترل تصادفی بزرگ، اختلاف در مورد ارتباط ژنوتیپ CYP2D6 و پیامد بالینی مرتبط با تاموکسی‌فن وجود دارد

ادامه پیوست (۱): خلاصه مهم‌ترین پژوهش‌ها در حوزه تحلیل بقا سرطان پستان

مرجع	سال	عنوان تحقیق	داده‌ها	مزایا و معایب روش
[۱۹]	۲۰۲۱	استفاده از تکنیک های داده کاوی برای پیش بینی میزان بقای بیماران سرطان پستان: مقاله مروری.	–	– برای پیش‌بینی احتمال بقا، سه تکنیک داده‌کاوی رگرسیون لجستیک، درخت تصمیم و ماشین بردار پشتیبان بیشترین کاربرد را در مقالات داشته اند. – در بیشتر مطالعات، خطر فاکتورهای سن، گرید تومور، استیج تومور و اندازه تومور استفاده شده بودند
[۴۶]	۲۰۲۱	مقایسه مدل‌های یادگیری ماشینی نظارتی و نیمه نظارتی در تشخیص سرطان پستان	داده‌های تشخیص سرطان ویسکانسین	– استفاده از نه الگوریتم طبقه‌بندی یادگیری ماشین نظارتی و نیمه نظارتی – دقت بالایی (۹۰٪–۹۸٪) روش‌های نیمه نظارتی – مدل KNN برای SL و رگرسیون لجستیک برای SSL به بالاترین دقت ۹۸٪ دست‌یافت – با استفاده از یک نمونه کوچک برچسب‌دار و قدرت محاسباتی کم، SSL به طور کامل قادر به جایگزینی الگوریتم‌های SL در تشخیص نوع تومور است.
[۴۷]	۲۰۲۲	یک چارچوب نیمه نظارت شده قابل تفسیر برای طبقه‌بندی سرطان پستان	اطلاعات ۱۶۲ زن مؤسسه سرطان نیوجرسی	– استفاده از یک چارچوب پنج‌مرحله‌ای شامل تقویت داده‌ها، انتخاب ویژگی، برچسب‌گذاری داده‌های آموزشی به روش نیمه نظارتی، مدل‌سازی شبکه عصبی عمیق و تفسیر به کمک شبکه عصبی – بهبود عملکرد مدل شبکه عصبی هنگام ترکیب داده‌های بدون برچسب با برچسب اولیه
[۴۸]	۲۰۲۲	یک مدل جدید تجزیه و تحلیل بقا در بیماران مبتلا به سرطان پستان تحت درمان با تاموکسی‌فن کمکی با استفاده از یادگیری نیمه نظارتی مبتنی بر منیفولد	ITPC - کنسرسیوم بین‌المللی تاموکسی فن فارماکوژنومیکس	– ارائه یک روش جدید تجزیه و تحلیل بقا با ترکیب مدل Cox-PH، مدل AFT و مدل یادگیری نیمه نظارت شده مبتنی بر منیفولد برای پیش‌بینی مؤثر تحلیل بقا. – دقت بیشتری مدل پیشنهادی در مقایسه با حالتی که فقط از مدل Cox-PH و زمانی که فقط از مدل AFT استفاده می‌شد. – مدل پیشنهادی به دلیل استفاده از دانش موجود در داده ها و پایداری مدل در برابر نویز، دقت بالایی در تخمین برچسب داده های بدون برچسب دارد.
[۴۹]	۲۰۲۲	مدل آنالیز بقای سرطان پستان	اطلاعات ۲۰۸۹ بیمار طی سال ۲۰۰۵ تا ۲۰۱۰ از بیمارستان‌های منطقه‌ای تایوان.	– تحلیل بقا و متاستاز با استفاده از مدل‌های خطر کاپلان – مایر، آزمون لگاریتمی و مدل‌های خطر متناسب کاکس – بررسی اثرات تک‌متغیره و چندمتغیره فاکتورهای پیش‌آگهی بالینی بر بقای بیماران مبتلا به سرطان پستان. – تجزیه و تحلیل بقای مدل خطر کاپلان–مایر، نشان داد که خطر سرطان پستان تحت درمان با جراحی کمتر از کسانی است که جراحی دریافت نکرده‌اند – رتبه‌بندی روش‌های درمانی بر اساس بقا: جراحی، هورمون‌درمانی، شیمی‌درمانی و پرتودرمانی
[۵۰]	۲۰۲۲	کاربرد و مقایسه مدل‌های یادگیری ماشینی برای پیش‌بینی پیش‌آگهی سرطان پستان: مطالعه گروهی گذشته‌نگر	اطلاعات ۲۲۱۷۶ بیمار مبتلا به سرطان پستان	– معرفی روش جنگل بقا تصادفی به‌عنوان یک رویکرد مؤثر برای ساخت مدل‌های پیش‌بینی پیش‌آگهی در زمینه تجزیه و تحلیل بقا – توانایی تشخیص بهتر مدل جنگل بقا تصادفی نسبت به سایر مدل‌ها

پیوست (۲): انواع داده‌ای به همراه شرح هر یک از ویژگی‌ها

جزئیات هر یک از انواع داده‌ای مورد استفاده درمجموعه‌های داده‌ای و شرح هر یک از ویژگی های آن در جداول زیرآمده است.

جدول (۱): داده‌های جمعیتی

Element	مولفه
Gender	جنسیت: زن و مرد
Race (OMB)	نژاد: بر اساس تعریف دفتر مدیریت و بودجه
Race (Self-Reported)	نژاد: بر اساس گزارش شخصی
Ethnicity (OMB)	قومیت: بر اساس تعریف دفتر مدیریت و بودجه
Ethnicity (Self-Reported)	قومیت: بر اساس گزارش شخصی
Age	سن به سال در تشخیص اولیه سرطان پستان

جدول (۲): داده‌های سابقه‌ای

Element	مولفه
Height	قد به cm
Weight	وزن به kg
Has the patient died?	آیا بیمار فوت کرده است؟
Prior History of Cancer	سابقه سرطان؟ بله و خیر و نامشخص
Sites of Prior Cancer	محل سرطان قبلی
Prior Use of Hormone	سابقه استفاده از هورمون‌درمانی جایگزین
BRCA1	جهش در ژن؟ بله، خیر
BRCA2	جهش در ژن؟ بله، خیر

جدول (۳): داده‌های فارماکوژنومیکی

Element	مولفه
CYP2D6 Allele 1	اولین آلل ژن CYP2D6
CYP2D6 Allele 2	دومین آلل ژن CYP2D6
CYP2D6_Genotype_PharmGKB	ژنوتایپ آلل ۲/ ژنوتایپ آلل ۱
Weak_Drug_PharmGKB	بازدارندگی ضعیف ژن CYP2D6
Potent_Drug_PharmGKB	بازدارندگی قوی ژن CYP2D6
Metabolizer_from_Drug_CYP2D6	وضعیت متابولیزه‌کننده بر اساس امتیاز ژنوتایپ

جدول (۴): داده‌های ژنوتایپی

Element	مولفه
CYP2D6 *4 genotype-rs3892097	ژنوتایپ آلل شماره ۴ ژن CYP2D6
Rs3892097 genotyping source	منبع استخراج DNA آلل
CYP2D6 *4/*10 genotype: - rs1065852	ژنوتایپ آلل شماره ۴/۱۰ ژن CYP2D6
Genotyping source rs1065852	منبع استخراج DNA آلل
Other_CYP2D6_Genotyping	سایر ژنوتایپ های CYP2D6

جدول (۵): داده‌های فنوتایپی

مولفه	Element
وضعیت استعمال سیگار در زمان ابتلا بیماری	Smoking Status at Time of Diagnosis
وضعیت یائسگی در زمان ابتلا بیماری	Menopause Status at Diagnosis
تعداد کل گره ها مثبت مورد بررسی از بیوپسی	Number of Positive Nodes
تعداد گره های بیوپسی غده لنفاوی پیش آهنگ (نگهبان)	Nodes examined from sentinel
متاستاز بیماری اولیه (بله–خیر–ارزیابی نشده–نامشخص)	Metastatic Disease at Primary Disease
حداکثر ابعاد تومورmm	Maximum Dimension of Tumor
وضعیت گیرنده استروژن	Estrogen Receptor
وضعیت گیرنده پروژسترون	Progesterone Receptor
زمان بین عمل جراحی سرطان پستان قطعی و شروع درمان کمکی	Time between definitive breast cancer surgery and start of adjuvant therapy?
بازه زمانی در نظرگرفته شده مصرف تاموکسی‌فن	Intended Tamoxifen Duration

جدول ۶ : داده‌های زمان‌های بقا در نتایج بالینی

مولفه	Element
زمان از تشخیص تا عود مجدد ی یا منطقه ای همان طرف بدن	Time from diagnosis to recurrence
زمان از تشخیص تا عود مجدد دوراز پستان	Time from diagnosis to distant recurrence
زمان از تشخیص تا عود مجدد پستان مقابل	Time from diagnosis to contralateral BS
زمان از تشخیص تا دومین سرطان تهاجمی	Time from diagnosis to second cancer
محل سرطان اولیه دوم	Site of second primary cancer
زمان از تشخیص بیماری تاآخرین ارزیابی	Followup_time
تعداد روز های بین تشخیص بیماری تا فوت بیمار	Time from diagnosis until death
زمان بقا در صورت زنده ماندن	Survival_Time_alive
زمان بقای بدون بیماری تهاجمی – نتیجه بالینی۱	IDFS (invasive disease–free survival)
زمان بقای عاری ازسرطان پستان -نتیجه بالینی ۲	BCFI (Breast-Cancer Free Interval)

جدول ۷ : داده‌های بالینی (پیگیری)

مولفه	Element
زمان بقای عاری از بیماری؟	Disease-free survival time
زمان از تشخیص تا عود مجدد دوراز پستان سرطان دوم	Time from diagnosis to distant recurrence?
محل سرطان اولیه دوم ؟	Site of second primary cancer
وضعیت ماموگرامی سالانه	Annual mammograms status?
وضعیت آزمایش فیزیکی سالانه پس از جراحی	Annual Physical Exam after Surgery?
تعداد روز های بین تشخیص بیماری تا فوت بیمار	Time from diagnosis until death?
اولین رخداد بیماری	First_Disease_Event