مجله دانش و تندرستی، دوره ۱۲، شماره ۱، بهار ۱۳۹۶
دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شاهرود

DOI: 10.22100/jkh.v12i1.1613

مقاله پژوهشی

مقایسه عملکرد دو مدل پیش‌بینی متاستاز براساس تکنیک‌های داده‌کاوی در بیماران سرطان پستان

نجمه ناظری[۱]، علیرضا آتشی*[۱]، سارا دری[۲]، ابراهیم عباسی[۳]، محسن علیجانی[۱]، محسن گلی[۱]، مسعود عبدالهی بیدهندی[۱]

۱- جهاد دانشگاهی- مرکز تحقیقات سرطان پستان- پژوهشکده معتمد- گروه پژوهشی انفورماتیک پزشکی.

۲- دانشگاه علوم پزشکی شهید بهشتی- دانشکده پیراپزشکی- گروه انفورماتیک پزشکی- کمیته تحقیقات دانشجویی.

۳- دانشگاه علوم پزشکی مشهد- دانشکده پزشکی- کمیته تحقیقات دانشجویی.

تاریخ دریافت: ۱۳۹۵/۹/۱۵، **تاریخ پذیرش:** ۱۳۹۵/۱۲/۲۴

چکیده

مقدمه: *با شناسایی فرآیند متاستاز و عوامل مؤثر بر آن به بهبود و بقای طولانی مدت بیماران کمک شایانی خواهد شد. هدف از مطالعه حاضر بررسی و شناسایی عوامل تأثیرگذار در پیش‌بینی متاستاز سرطان پستان با استفاده از ابزارهای داده‌کاوی است. داده‌کاوی ابزار کشف دانش از میان انبوهی از داده است که امروزه در زمینه‌های مختلفی کاربرد پیدا کرده است. تشخیص بیماری در علم پزشکی یکی از زمینه‌های رو به رشد و پرکاربرد داده‌کاوی است.*

مواد و روش‌ها: *در این پژوهش پس از آماده‌سازی داده‌ها، ۲۰۲۵ رکورد قابل استفاده مورد تجزیه و تحلیل قرار گرفت. سپس با استفاده از الگوریتم‌های شبکه عصبی و CHAID به کشف الگوهایی که به پیش‌بینی متغیرهای تأثیرگذار بر متاستاز در بیمار کمک می‌کند، پرداخته‌ایم.*

نتایج: *دو الگوریتم فوق‌الذکر پیاده‌سازی گردید و از نظر درجه اطمینان مقایسه شدند، درجه اطمینان برای الگوریتم شبکه عصبی ۹۴/۱۴ و برای CHAID 24/92 به‌دست آمد. براساس نتایج متغیرهای سطح بیماری (Stage)، نوع عمل جراحی، نوع سرطان براساس پاتولوژی و Her2 مهمترین متغیرهای پیش‌بینی‌کننده متاستاز هستند.*

نتیجه‌گیری: *مقایسه عملکرد مدل‌ها در این پژوهش نشان می‌دهد که الگوریتم‌های CHAID و شبکه عصبی در پایگاه داده مورد استفاده، با درجه اطمینان بسیار بالا و تقریباً برابر، روش‌های مناسبی برای پیش‌بینی متاستاز در بیماران سرطان پستان می‌باشد.*

واژه‌های کلیدی: داده‌کاوی، سرطان پستان، متاستاز، پیش‌بینی.

***نویسنده مسئول:** تهران، میدان ونک، ابتدای گاندی جنوبی، پلاک ۱۴۶، طبقه دوم، **تلفن:** ۰۲۱۸۸۶۷۷۵۷۸، **نمابر:** ۰۲۱۸۸۶۷۹۴۰۲، **Email:** smatashii@gmail.com

ارجاع: ناظری نجمه، آتشی علیرضا، دری سارا، عباسی ابراهیم، علیجانی محسن، گلی محسن، عبدالهی بیدهندی مسعود. مقایسه عملکرد دو مدل پیش‌بینی متاستاز براساس تکنیک‌های داده‌کاوی در بیماران سرطان پستان. مجله دانش و تندرستی ۱۳۹۶؛۱۲(۱):۳۶-۴۲.

مقدمه

به‌طور کلی یک نفر از هر هشت زن در طول عمر خود به سرطان پستان مبتلا می‌شود. این نوع سرطان اغلب در بانوان و به‌صورت محدود در مردان مشاهده شده است. ماموگرافی، بایوپسی و آسپیراسیون سوزنی (Fine needle aspiration) سه تکنیک متداول کشف و تشخیص سرطان پستان می‌باشد (۱).

از مسائل مهم در انواع سرطان انتشار سلول سرطانی به بافت‌های دیگر (متاستاز دوردست) می‌باشد. متاستاز به‌عنوان پراکندگی و گسترش سلول‌های تومور اولیه و برپایی دومین تومور و به‌دنبال آن تومورهای بعدی در بافت‌های دیگر تعریف می‌شود. وقتی که سرطان تنها به درگیری نسج پستان محدود می‌شود، میزان بقای آن بالاست اما وقتی که انتشار سلول‌ها به بافت‌های دیگر اتفاق می‌افتد میزان بقا به‌صورت معنی‌داری کاهش می‌یابد. کیفیت زندگی بیماران با متاستاز در وضعیت بدتری نسبت‌به بیماران با سرطان موضعی قرار دارد. بنابراین با شناسایی فرایند متاستاز و عوامل مؤثر بر آن به بهبود بقای طولانی مدت بیماران کمک شایانی خواهد شد (۲).

داده‌کاوی، تکنیک و ابزار کشف دانش از میان انبوهی از داده است که امروزه در زمینه‌های مختلفی کاربرد پیدا کرده است. هدف نهایی داده‌کاوی، ایجاد سیستم‌های پشتیبانی تصمیم‌گیری سازمانی است. داده‌کاوی، به استخراج اطلاعات مفید و دانش از حجم زیاد داده‌ها می‌پردازد (۳). تشخیص بیماری‌های مختلف در علم پزشکی یکی از زمینه‌های پرکاربرد داده‌کاوی محسوب می‌شود که در سال‌های اخیر مطالعات بسیاری در این حوزه و با استفاده از تکنیک‌های متفاوت داده‌کاوی انجام پذیرفته است که به برخی از آنها خواهیم پرداخت.

در مطالعه حاضر بررسی و شناسایی عوامل تاثیرگذار در پیش‌بینی متاستاز سرطان پستان با استفاده از دو الگوریتم متفاوت داده‌کاوی بر روی پایگاه داده مرکز خدمات تخصصی بیماری‌های پستان جهاد دانشگاهی، انجام شده است. سپس نتایج حاصل از این الگوریتم‌ها از نظر درجه اطمینان با یکدیگر مقایسه شده تا مناسب‌ترین الگوریتم برای پیش‌بینی متاستاز پیشنهاد شود.

پیش‌بینی استعداد ابتلا به بیماری یا متاستاز، یک نمونه از مدل‌سازی چند متغیره است. مدل پیش‌بینی برای یک بیماری به خوبی توسط پزشکان پذیرفته شده است. هدف از مدل‌سازی پیش‌بینی در طب بالینی، استخراج مدلی است که می‌تواند برای پیش‌بینی نتیجه مورد توجه استفاده شود. درنتیجه پشتیبانی از تصمیم‌گیری بالینی در پیش آگهی، تشخیص و یا برنامه‌ریزی درمان براساس اطلاعات خاص بیماران صورت می‌گیرد (۴). تکنیک‌های پیش‌بینی معمول و مورد استفاده عبارتند از: درخت تصمیم‌گیری، رگرسیون لجستیک، شبکه‌های عصبی مصنوعی، KNN و ماشین‌بردار پشتیبانی.

در سال ۲۰۱۲ مطالعه‌ای در خصوص بررسی روش‌های جدید و مؤثر کشف سرطان پستان با استفاده از شبکه‌های عصبی فازی جهت تلفیق محاسن روش شبکه عصبی با رویکردهای فازی انجام گرفته است. بدین ترتیب که با استفاده از آن سیستم تصمیم‌گیری هوشمندی را طراحی نموده است. باتوجه به آنکه افزایش تعداد داده‌های ورودی باعث رشد تصاعدی در تعداد پارامترهای شبکه عصبی فازی شده و در نتیجه افزایش و پیچیدگی محاسبات را به همراه دارد- به این پدیده نفرین ابعاد می‌گویند- شبکه عصبی فازی سلسله مراتبی (HFNN) و FGPNN برای مواجهه با این مشکل استفاده شده است. نتایج شبیه سازی نشان داده است که این روش‌ها حتی در حالت کم بودن قوانین و شاخص‌ها نیز، مؤثر می‌باشند (۵).

مطالعه‌ای در سال ۲۰۱۲ با هدف ایجاد یک مدل جدید برای پیش‌بینی متاستاز به غدد لنفاوی زیر بغل (AxLN: Axillary lymph node) در سرطان پستان اولیه انجام شد. برای این منظور از روش پیش‌بینی درخت تصمیم- درخت تصمیم تناوبی (ADTree)- استفاده شده است و پایگاه داده‌هایی از بیماران سرطان پستانی که برای آنها بیوپسی یا خارج کردن غدد زیر بغل قبل از درمان انجام شده بود از سه مؤسسه جمع‌آوری شد و آزمون‌های مدل و روایی و اعتبارسنجی خارجی انجام شد. در این مطالعه مدل درخت تصمیم، ۱۵ متغیر بالینی و آسیب‌شناسی (Clinico-pathological) را از میان ۲۴ متغیر انتخاب کرد و دقت بالایی برای پیش‌بینی متاستاز در بیماران سرطان پستان با استفاده از این متغیرها نشان داد (۶).

در مطالعه‌ای دیگر که جهت ارایه شاخص‌های پیش‌بینی‌کننده برای تمایز زیرگروه‌های بیماری سرطان پستان توسط ترو (Mevlut ture) انجام شد از الگوریتم‌های درخت تصمیم (C&R,CAHID, QUEST, ID3, C4.5, C5) و از تحلیل رگرسیون Cox برای محاسبه دوره بقای عاری از بیماری (Disease free survival) در بیماران سرطان پستان استفتده شد. نتایج این مطالعه نشان داد که رگرسیون Cox متداول‌ترین روش برای بررسی همزمان اثر چندین عامل بر روی بقای بیمار (Overal survival) است. اما پیش‌بینی‌کننده خوبی برای تعیین بقای بدون بیماری (DFS) نیست. درخت تصمیم از رگرسیون Cox مناسب‌تر است چرا که درخت تصمیم برای استخراج الگوها و روابط پنهان از پایگاه داده توانمندتر می‌باشد. در این مقاله هفت رویکرد مورد استفاده قرارگرفت. تمام شاخص‌های پیش‌بینی‌کننده، درجه خوبی از طبقه‌بندی را با استفاده از عوامل ریسک استاندارد نشان دادند. این پژوهش نشان داد C4.5 بهترین عملکرد را نسبت‌به سایر الگوریتم‌ها در تعیین ریسک گروه‌ها داشت. پیشنهاد این مطالعه ارزیابی داده‌ها با استفاده از روش‌های مختلف و نه فقط یک روش در مطالعات سرطان پستان و یا سایر شرایط پزشکی بود (۷).

در تحقیقی به‌منظور بهبود نسبت هزینه اثربخشی (Cost-effectiveness) غربالگری (Screening) سرطان پستان، نویسنده عملکرد الگوریتم شبکه عصبی مصنوعی را ارزیابی نمود تا خروجی (سرطان/ عدم سرطان) را برای استفاده به‌عنوان یک طبقه‌بندی‌کننده پیش‌بینی نماید. شبکه‌ها بر روی داده‌هایی از داده‌های مربوطبه سرطان، بیماری‌های زنان و زایمان و متغیرهای تغذیه مورد آزمون قرار گرفتند. شبکه مصنوعی عصبی تا ۹۴/۰۴٪ مقدار پیش‌بینی مثبت و ۹۷/۶۰٪ مقدار پیش‌بینی منفی را به‌دست آوردند. نتایج می‌تواند به‌عنوان خط مشی‌ای برای زنانی که مستعد سرطان پستان هستند مورد استفاده قرارگیرد (۸).

در ایران در مطالعه‌ای تجزیه و تحلیل داده‌ها با استفاده از قوانین همبستگی و رگرسیون لجیت رتبه‌ای، توسط حسینی و همکاران انجام گرفت. نتایج این بررسی نشان داد با درنظر گرفتن بیماران با گیرنده‌های هورمونی استروژن و پروژسترون و ژن P53 مثبت، شانس متاستازهای بیشتر در بیماران با نسبت بالای غدد لنفاوی درگیر در زیر بغل، به‌ترتیب ۴/۷۷، ۱/۴۴ و ۴/۴۵ برابر بیماران با نسبت غدد لنفاوی درگیر کمتر است. در مورد بیماران با وضعیت منفی استروژن و پروژسترون و p53 شانس متاستازهای بیشتر در بیماران با نسبت بالای غدد لنفاوی درگیر در زیر بغل، ۹۷ درصد افزایش دارد (۲).

شبکه‌های عصبی مصنوعی (Artificial neural network- ANN) یک سیستم پردازش اطلاعات است که به‌صورت ترکیبی از عناصر به شدت به هم پیوسته (نورون‌ها) به‌صورت موازی برای حل یک مشکل عمل می‌کنند. شبکه‌های عصبی می‌توانند برای استخراج الگو و تشخیص روند ازحجم عظیمی از داده‌ها مورد استفاده قرار گیرند که به‌روش دیگری قابل تشخیص نیستند. شبکه‌های عصبی مصنوعی مزیت یادگیری برای پیش‌بینی ارتباطات غیر خطی پیچیده تصادفی بین متغیرهای مستقل و وابسته را دارا می‌باشند (۹).

CHAID (Chi-Squared automatic interaction detection) یک تکنیک داده‌کاوی از نوع درخت تصمیم است که بر مبنای آزمون معنی‌داری تنظیم شده است. این روش شباهت بسیاری به‌روش تجزیه و تحلیل رگرسیون دارد و برای تعداد زیادی از متغیرهای پیش‌بینی‌کننده به‌کار می‌رود. در عمل CHAID اغلب در زمینه پیدا کردن گروه‌هایی از عوامل و پیش‌بینی چگونگی تأثیر پاسخ آنها به برخی متغیرها روی متغیرهای دیگر به‌کار می‌رود و کاربرد آن در پژوهش‌های پزشکی و روانشناسی بسیار است (۱۰).

مدل‌های بسیاری برای پیش‌بینی متاستاز پیشنهاد شده‌اند که هر کدام دقت‌های متفاوتی دارند، دستیابی به مدلی که دقت بیشتری داشته باشد و در آن از داده‌های بومی استفاده شده باشد، مورد تلاش این پژوهش است. هدف از این پژوهش مقایسه عملکرد دو مدل شبکه‌های عصبی مصنوعی و CHAID است. این مدل‌ها می‌توانند در ساخت ابزارهای تصمیم‌یار برای پزشکان و کادر درمانی به‌کار روند. ساخت این ابزار از اهداف این پژوهش فراتر است و می‌تواند به‌عنوان پژوهش عملی دیگری تعریف و اجرا گردد.

مواد و روش‌ها

در این مطالعه، مدل‌سازی با استفاده از داده‌های مربوطبه زنان مبتلابه سرطان پستان مراجعه‌کننده به پژوهشکده سرطان پستان جهاد دانشگاهی انجام شد. در داده‌کاوی کلیه رکوردهای قابل استفاده در پژوهش که معیار ورود به داده‌کاوی را در بر دارند، مورد استفاده قرار می‌گیرند.

در پایگاه داده این مطالعه ۱۷ متغیر مستقل و یک متغیر وابسته که بروز متاستاز می‌باشد، وجود دارد و از آنجاکه روش‌شناسی مدل‌یابی معادلات ساختاری، تا حدود زیادی با برخی از جنبه‌های رگرسیون چند متغیری شباهت دارد، می‌توان از اصول تعیین حجم نمونه در تحلیل رگرسیون چند متغیری برای تعیین حجم نمونه در مدل‌یابی معادلات ساختاری استفاده نمود (۱۱). در تحلیل رگرسیون چند متغیری نسبت تعداد نمونه (مشاهدات) به متغیرهای مستقل نباید از ۵ کمتر باشد. در غیر این‌صورت نتایج حاصل از معادله رگرسیون چندان تعمیم‌پذیر نخواهد بود. از دیدگاه جیمز استیونس حتی در نظر گرفتن ۱۵ مشاهده به ازای هر متغیر پیش‌بین در تحلیل رگرسیون چندگانه با روش معمولی کمترین مجذورات استاندارد، یک قاعده سر انگشتی خوب به حساب می‌آید (۱۱). پس به‌طور کلی در روش‌شناسی مدل‌یابی معادلات ساختاری تعیین حجم نمونه می‌تواند بین ۵ تا ۱۵ مشاهده به ازای هر متغیر اندازه‌گیری شده تعیین شود. بنابراین به ازای هر متغیر مستقل و زمینه‌ای به حداکثر ۱۵ نمونه نیاز داریم و باتوجه به ۱۷ متغیر مستقل این مطالعه در مجموع برای تحلیل مناسب و مدل‌سازی حداقل ۲۵۵ نمونه خواهیم بود.

در پایگاه داده مرکز خدمات تخصصی بیماری‌های پستان جهاد دانشگاهی، پس از آماده‌سازی داده‌ها که به‌صورت مشروح به ذکر مراحل آن خواهیم پرداخت، ۲۰۲۵ رکورد قابل استفاده در داده‌کاوی مورد تجزیه و تحلیل قرار گرفت که این تعداد بسیار بیشتر از حداقل تعیین شده می‌باشد.

در مرحله data cleaning با عنایت به آنکه برخی متغیرها برای پستان سمت چپ و راست به‌طور مجزا ثبت گردیده بود، به این معنی که برای نمایش یک عامل از دو متغیر استفاده شده بود و باتوجه به آنکه قرار گرفتن توده در سینه چپ یا راست، در این تحقیق مورد توجه نبود، داده‌ها ادغام گردید و تحت عنوان یک متغیر لحاظ شد و همچنین متغیر مربوطبه سن ابتلاء به بیماری از طریق متغیرهای سال تولد و زمان مراجعه به کلینیک استخراج گردید.

باتوجه به پیش پردازش صورت گرفته در نهایت ۱۱ متغیر مستقل Stage، نوع عمل جراحی، نوع براساس پاتولوژی، سن ابتلا، وضعیت تأهل، شغل، گیرنده‌های هورمونی استروژن، گیرنده هورمونی Her2، گیرنده‌های هورمونی پروژسترون، تعداد غدد لنفاوی درگیر و ژن P53 مورد تجزیه و تحلیل قرار گرفتند. در نهایت پس از انجام کلیه مراحل مذکور ۲۰۲۵ رکورد در این پژوهش مورد تجزیه و تحلیل قرار گرفت.

مرحله بعدی کار مدل‌سازی براساس داده‌ها و دانش موجود می‌باشد. اولین گام مدل‌سازی، انتخاب تکنیک‌های مدل‌سازی که مورد استفاده قرار می‌گیرند، می‌باشد. تکنیک‌های متعددی برای استفاده در مدل‌سازی مانند استفاده از مدل‌های رگرسیونی و مدل‌های مبتنی بر داده‌کاوی و یادگیری ماشینی موجود می‌باشند و می‌توان از چندین تکنیک در مدل‌سازی بهره جست.

در این پژوهش با استفاده از الگوریتم‌های داده‌کاوی Neural Net و CHAID با استفاده از نرم‌افزار Clementine12 به کشف الگوهایی که به پیش‌بینی متغیرهای تأثیرگذار بر متاستاز در بیمار کمک می‌کند، پرداخته‌ایم. علت انتخاب این الگوریتم‌ها هم به‌دلیل استفاده رایج از آنها و نتایج مناسب در مطالعات مشابه است.

مدل‌های ایجاد شده با استفاده از محاسبه درجه اطمینان (Confidence) با یکدیگر مقایسه شدند.

نتایج

در مطالعه حاضر با استفاده از تکنیک شبکه عصبی نتایج نشان داد که متغیرهای Stage، وضعیت تأهل و نوع پاتولوژی بالاترین تأثیر را در ایجاد متاستاز داشته‌اند. متغیر متاستاز به‌عنوان خروجی و سایر متغیرها به‌عنوان متغیرهای ورودی در نظر گرفته شده‌اند.

این مدل نشان می‌دهد Stage تأثیر زیادی در ایجاد متاستاز بعدی دارد. بنابراین احتمال می‌رود انواع سرطان با Stage بالاتر بیشتر از سایرین در معرض ایجاد متاستاز قرار داشته باشند.

تحلیل درجه اطمینان این روش ۹۴/۱۴ درصد می‌باشد.

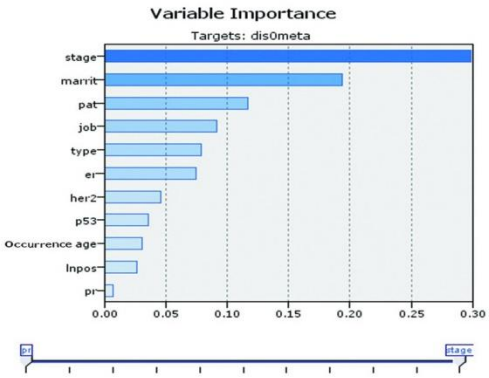
مانند دیگر درخت‌های تصمیم مزایای CHAID این است که خروجی آن به شدت بصری است و به آسانی قابل تفسیر است.

تحلیل درجه اطمینان این روش ۹۴/۲۴ درصد می‌باشد.

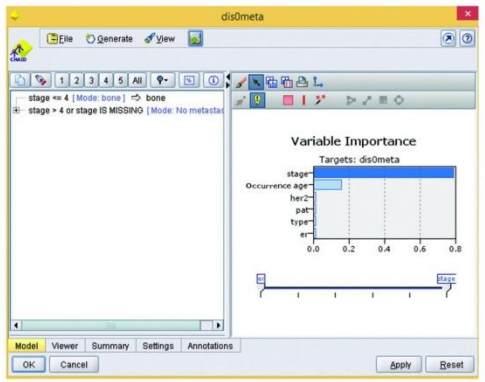
با در نظر گرفتن هر یک از الگوریتم‌ها در جدول ۱ عوامل مهمتر در پیش‌بینی متاستاز استخراج شدند.

جدول ۱- متغیرهای مهم استخراج شده از الگوریتم‌ها

CHAID	Neural Net
Stage	Stage
Occurrence age (سن ابتلا)	Marrit (وضعیت تأهل)
Her2	Pat (نوع براساس پاتولوژی)
Pat (نوع براساس پاتولوژی)	Job
Type (نوع عمل جراحی)	Type (نوع عمل جراحی)
ER	ER
Her2	-
P53	-
Occurrence age (سن ابتلا)	-
LNPos (تعداد لنف نودهای مثبت)	-
PR	



شکل ۱- تعیین متغیرها مؤثر بر متاستاز با الگوریتم شبکه عصبی



شکل ۲- تعیین متغیرهای مؤثر بر متاستاز با الگوریتم CHAID

در جدول ۲ طبق نظر متخصصان وضعیت تأهل و شغل را جدول حذف کردیم و جدول نهایی طبق الگوریتم‌های پیشنهادی به‌صورت زیر می‌باشد.

جدول ۲- رتبه هر یک از متغیرها در هر الگوریتم

متغیرها	Neural Network	CHAID
Stage	۱	۱
LNPos (تعداد لنف نودهای مثبت)	۸	–
Occurrence age (سن ابتلا)	۷	۲
Type (نوع عمل جراحی)	۳	۵
PR	۹	–
ER	۴	۶
P53	۶	–
Her2	۵	۳
Pat (نوع براساس پاتولوژی)	۲	۴

بحث

مقایسه عملکرد مدل‌های مختلف که در این پژوهش مورد آزمایش قرار گرفته‌اند، نشان می‌دهد که آلگوریتم‌های CHAID و شبکه عصبی مصنوعی روش‌های مناسبی برای پیش‌بینی متاستاز در بیماران سرطان پستان می‌باشند. در دو الگوریتم CHAID و شبکه عصبی عامل سطح بیماری (Stage) به‌عنوان نخستین عامل مؤثر مطرح است. همان‌طور که ذکر شد طبق نظر متخصصان بالینی شاغل بودن، وضعیت تأهل و شغل را از میان عوامل مؤثر حذف کردیم و در نتیجه عوامل مؤثر بر متاستاز در دو مدل با کمی تفاوت در ترتیب، نسبتاً مشابه می‌باشند. پس از متغیر سطح بیماری که به‌عنوان نخستین عامل شناسایی شد، نوع عمل جراحی، نوع براساس پاتولوژی، سن ابتلا، گیرنده‌های هورمونی استروژن، گیرنده هورمونی Her2، تعداد غدد لنفاوی درگیر، گیرنده‌های هورمونی پروژسترون و ژن P53 مثبت همگی از متغیرهای پیش‌بینی‌کننده متاستاز هستند.

همان‌طور که در نتایج پژوهشی که برای مقایسه عوامل مؤثر بر سرطان پستان با استفاده از روش های داده‌کاوی و رگرسیون Cox صورت گرفته (۷) آمده است، درخت تصمیم برای استخراج الگوها و روابط پنهان از پایگاه داده توانمندتر می‌باشد. باتوجه به اینکه آلگوریتم CHAID همان‌طور که پیش از این مطرح شد، از روش‌های درخت تصمیم به‌شمار می‌رود، نتایج مناسب و دقت بالایی که در این پژوهش با پیاده‌سازی این الگوریتم به‌دست آمده، هم راستای نتایج پژوهش مذکور می‌باشد. در پژوهشی دیگر که توسط قاسم احمد و همکاران درخصوص معرفی تعدادی از الگوریتم‌های پرکاربرد و شناخته شده داده‌کاوی در سرطان پستان انجام شده است نیز آمده است: الگوریتم‌های درخت تصمیم و ماشین‌بردار پشتیبان، در تحقیقات مختلف انجام شده، معمولاً نتایج بهتر و دقیق‌تری در زمینه دقت، حساسیت و ویژگی ارایه کرده‌اند (۱۳).

نتیجه مطالعه‌ای که در خصوص بررسی روش‌های جدید و مؤثر کشف سرطان پستان با استفاده از شبکه‌های عصبی فازی جهت تلفیق محاسن روش شبکه عصبی با رویکردهای فازی انجام گرفته است (۵)، نشان می‌دهد که روش‌های شبکه عصبی حتی در حالت کم بودن قوانین و پارامترها نیز، مؤثر می‌باشند. همچنین پژوهشی تازه‌تر از الگوریتم‌های شبکه فازی برای ایجاد مدل پیش‌بینی بقای ۵ ساله استفاده نموده است که در آن درجه اطمینان مدل ۸۵ گزارش شده است (۱۲). نتایج پژوهش حاضر نشان می‌دهد الگوریتم شبکه عصبی برای مدل پیش‌بینی متاستاز نیز با درجه اطمینان ۹۴/۱۴ به‌عنوان یک روش بسیار مناسب قابل استفاده می‌باشد. قابل ذکر است که باتوجه به مطالعات مذکور پیش‌بینی می‌شود، اگر داده‌های مورد استفاده به داده‌های فازی تبدیل شوند نتایج بهتری به‌دست خواهد آمد.

در مدل‌های پیشنهادی این پژوهش که از درجه اطمینان بسیار بالایی برخوردارند، متغیرهای بسیاری در رتبه اهمیت و تأثیر بالاتری نسبت‌به متغیرهای گیرنده‌های هورمونی استروژن و پروژسترون و ژن P53 مثبت و غدد لنفاوی درگیر قرار گرفته‌اند. چهار عامل اول در مدل نهایی پیشنهادی ما که در الگوریتم‌های مورد استفاده رتبه بالایی کسب نمودند عبارتند از Stage، نوع عمل جراحی، نوع براساس پاتولوژی و سن ابتلا. این عوامل از نظر متخصصین بالینی نیز می‌توانند تأثیر بالایی در احتمال بروز متاستاز سرطان پستان داشته باشند. این در حالی است که کلیه این عوامل در پژوهش حسینی و همکاران (۲) مغفول مانده است.

در پژوهش حاضر نسبت‌به مطالعات پیشین، انواع متنوع‌تری از الگوریتم‌های داده‌کاوی مورد آزمایش قرار گرفته و نتایج حاصل از آنها با یکدیگر مقایسه شده است و برای دو الگوریتم شبکه عصبی مصنوعی و CHAID که بر روی تعداد بسیار بالایی از رکوردها اجرا شد، درجه اطمینان بیش از ۹۴ به‌دست آمد که می‌توان آنها را به‌عنوان الگوریتم‌های بسیار مناسبی برای پیش‌بینی متاستاز سرطان پستان پیشنهاد نمود. در سال‌های اخیر درخصوص پیش‌بینی عود مجدد سرطان پستان با استفاده از تکنیک‌های داده‌کاوی در کشور پژوهش‌های متنوع‌تری صورت گرفته است که از میان آنها می‌توان به پیش‌بینی عود مجدد سرطان پستان به کمک سه تکنیک داده‌کاوی (۱۴) و ایجاد یک مدل پیش‌آگهی مبتنی‌بر داده‌کاوی برای پیش‌بینی عود مجدد سرطان پستان (۱۵) اشاره نمود و این در حالی است که درخصوص پیش‌بینی متاستاز دوردست تاکنون مطالعات کمتری صورت گرفته است. در پژوهشی دیگر از سه الگوریتم درخت تصمیم، ماشین‌بردار پشتیبان و شبکه عصبی مصنوعی برای پیش‌بینی عود مجدد استفاده شد که نتایج به‌دست آمده حاکی از دقت بالاتر الگوریتم شبکه عصبی نسبت‌به درخت تصمیم است (۱۴).

از مزایای دیگر این مطالعه می‌توان به استفاده از داده‌های بومی واقعی اشاره کرد. در این پژوهش از داده‌های ثبت شده در مرکز خدمات تخصصی بیماری‌های پستان جهاد دانشگاهی استفاده شد که منجر به طراحی و ارایه مدل بومی برای پیش‌بینی متاستاز سرطان پستان شده است که قابل استفاده در مرکز مورد مطالعه با خصوصیات مختص به آن است.

از محدودیت‌های این پژوهش می‌توان به این نکته اشاره کرد که در داده‌کاوی مطالعه بر روی یک پایگاه داده صورت می‌پذیرد که قابلیت تعمیم آن به سایر پایگاه داده‌ها امکان‌پذیر نیست. با اجرای الگوریتم‌های پیشنهادی بر روی پایگاه داده های متفاوت در محدوده‌ی جغرافیایی موردنظر، می‌توان به نتایج قابل تعمیم دست یافت. همچنین ممکن است با اضافه شدن بیماران به مرور یکی از این مدل‌ها نتیجه

بهتری داشته باشه، بنابراین پیشنهاد می‌گردد که پس از مدتی این مطالعه تکرار شود. وجود داده‌های گم شده نسبتاً زیاد از دیگر محدودیت‌های این پژوهش بوده که با استفاده از مدیریت داده‌های گم شده توسط نرم‌افزار تأثیرهای احتمالی آن بر نتایج تاحدودی تعدیل شده است.

مدل‌های پیش‌بینی که با استفاده از دو الگوریتم CHAID و Neural Net به‌دست آمده است، می‌توانند به‌عنوان مدل‌هایی با درجه اطمینان بالا مورد استفاده قرار بگیرند. در این مطالعه درجه اطمینان هر دو مدل بالاتر از ۹۴ به‌دست آمد، لیکن مقایسه این دو مدل می‌تواند با استفاده از داده‌های آتی در پژوهش‌های آینده مورد توجه قرار گیرد. باتوجه به محدودیت تعمیم‌پذیری ذکر شده و اهمیت پیش‌بینی متاستاز سرطان پستان، اجرای مدل‌های پیشنهادی پژوهش ما بر روی سایر پایگاه داده‌ها در نقاط مختلف شهر تهران و یا در ابعاد وسیع‌تر در سطح کشور، جهت تعمیم نتایج به‌دست آمده می‌تواند صورت پذیرد.

در این پژوهش الگوریتم‌های داده‌کاوی شبکه عصبی مصنوعی و CHAID برای پیش‌بینی متاستاز سرطان پستان، از نظر درجه اطمینان با یکدیگر مقایسه شدند. مقایسه عملکرد مدل‌های مختلف که در این پژوهش مورد آزمایش قرار گرفته‌اند، نشان می‌دهد که الگوریتم‌های CHAID و شبکه عصبی مصنوعی روش‌های مناسبی برای پیش‌بینی متاستاز در بیماران سرطان پستان می‌باشند.

References

1. Azimian F, Tabrizi Gh, Jalali M. Breast cancer diagnosis using data mining techniques. 4th Iran Data Mining Conference;2010.
2. Hosseini SM, Hassannejad R, Khademolghorani Sh, Tabatabaeian M, Mokarian F. Identification of patterns of breast cancer metastasis among women referred to Isfahan seyedoshohada center, Iran, between 1999 and 2009 by association rules and ordinal logistic regression. Journal of Health System Research 2011;6:746-62.
3. Han J, Kamber M, Pei J. Data mining: concepts and techniques.3rd ed. San Diego: Academic Press;2001.
4. Heckerling PS, Gerber BS, Tape TG, Wigton RS. Use of genetic algorithms for neural networks to predict community-acquired pneumonia. Artif Intell Med 2004;30:71-84.
5. Naghibi S, Teshnehlab M, Shoorehdeli MA. Breast cancer classification based on advanced multidimensional fuzzy neural network. J Med Syst 2012;36:2713-20. doi: 10.1007/s10916-011-9747-5
6. Takada M, Sugimoto M, Naito Y, Moon HG, Han W, Noh DY, et al. Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model. BMC Med Inform Decis Mak 2012;12:54. doi: 10.1186/1472-6947-12-54
7. Mevlut Ture A, Fusun Tokatli B, Imran Kurt Omurlu. The comparisons of prognostic indexes using data mining techniques and Cox regression analysis in the breast cancer data. Expert Systems with Applications 2008;36:8247-54. doi: 10.1016/j.eswa.2008.10.014
8. Ronco AL. Use of artificial neural networks in modeling associations of discriminant factors: towards an intelligent selective breast cancer screening. Artif Intell Med 1999;16:299-309.
9. Magidson J, Vermunt JK. An Extension of the CHAID tree-based segmentation algorithm to multiple dependent variables. In: Weihs C, Gaul W, editors. Classification the ubiquitous challenge. Studies in classification, data analysis, and knowledge organization. Berlin: Springer;2005.p.176-7. doi: 10.1007/3-540-28084-7_18
10. Pournik O, Dorri S, Zabolinezhad H, Alavian SM, Eslami S. A diagnostic model for cirrhosis in patients with nonalcoholic fatty liver disease: an artificial neural network approach. Med J Islam Repub Iran 2014;28:116.
11. Hooman H. Structural equation modeling using Lisrel. Organization for Researching and Composing University Textbooks in the Humanities (SAMT) 2004.
12. Wang TN, Cheng CH, Chiu HW. Predicting post-treatment survivability of patients with breast cancer using artificial neural network methods. Conf Proc IEEE Eng Med Biol Soc 2013;1290-3. doi: 10.1109/EMBC.2013.6609744
13. Ghasem Ahmad L. Review top 7 algorithms in data mining for prediction survivability, diagnosis and recurrence of breast cancer. Iranian Quarterly Journal of Breast Disease 2013;6:52- 61.
14. Ghasem Ahmad L .Using data mining techniques for prediction breast cancer recurrence. Iranian Journal of Breast Diseases 2013;5: 23-34.[Persian].
15. Kiani B, Atashi A. A prognostic model based on data mining techniques to predict breast cancer recurrence. Journal of Health and Biomedical Informatics 2014;1:26-31.[Persian].

Journal of Knowledge & Health 2017; Vol 12, No 1
Shahroud University of Medical Sciences

DOI: 10.22100/jkh.v12i1.1613

Original Article

Performance Evaluation of Two Prediction Models for Breast Cancer Metastasis Based on Data Mining Techniques, A Comparison Study

Najme Nazeri (M.Sc.)[1], Ali Reza Atashi (M.Sc.)[1*], Sara Dorri (M.Sc.)[2], Ebrahim Abbasi (M.D.)[3], Mohsen Alijani (M.D.)[1], Mohsen Goli (M.Sc.)[1], Masoud Abdollahi[1] (M.Sc.)

1- Dept. Medical Informatics, Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran.
2- Dept. of Medical Informatics, School of Allied Medical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
3- Student Research Committee, School of Medical Sciences, Mashhad University of Medical Sciences, Mashhad, Iran.

Received: 5 December 2016**, Accepted:** 14 March 2017

Abstract:

Introduction: *Defining the metastasis processes and what are the most effecting factors on improve the survival of patients and hopefully treating them. We aim to investigate and defining the factors predict breast cancer metastasis using data mining techniques. Data mining is the technique and tool of knowledge discovery from the big data which is spreading rapidly in several areas of research and business nowadays. In medicine, diagnosis of diseases is one of the fruitful and highly spreading filed of data mining.*

Methods: *There were 2025 usable records in ACECR breast disease center's data base after data preparation. In the study here we try to uncover the patterns that would help the prediction of metastasis factors using CHAID and Artificial Neural Network.*

Results: *In this study, two mentioned algorithms were implemented and compared in terms of degree of confidence. Confidence was 94.14 for artificial neural network and 94.24 for CHAID algorithm. We found the tumor stage, surgery type and pathology results, as the most important variables in metastasis prediction.*

Conclusion: *Comparing the algorithms execution results and their confidence, both artificial neural network and CHAID are convenient prediction models for breast cancer metastasis.*

Keywords: Data mining, Breast Cancer, Metastasis, Prediction.

Conflict of Interest: No

***Corresponding author:** AR. Atashi**, Email:** smatashii@gmail.com*

Citation: Nazeri N, Atashi AR, Dorri S, Abbasi E, Alijani M, Goli M, Abdollahi M. Performance evaluation of two prediction models for breast cancer metastasis based on data mining techniques, a comparison study. Journal of Knowledge & Health 2017;12(1):36-42.