مقاله پژوهشی

فصلنامهٔ بیماری‌های پستان ایران، سال هفتم، شماره دوم، تابستان ۱۳۹۳

تعیین مرحله بالینی سرطان پستان توسط الگوریتم‌های داده‌کاوی

مریم سادات محمودی: مربی، گروه مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه پیام‌نور

سید عباس محمودی*: کارشناسی ارشد کامپیوتر، دانشگاه آزاد اسلامی واحد علوم تحقیقات یزد

فاطمه حقیقی: دانشیار گروه پاتولوژی، دانشکده پزشکی دانشگاه علوم پزشکی بیرجند

سید مصطفی محمودی: استادیار آسیب‌شناسی دهان و فک و صورت، دانشگاه علوم پزشکی بیرجند

چکیده

مقدمه: سرطان پستان شایع‌ترین سرطان زنان در ایران است. مرحله بالینی بیماری (stage) تعیین‌کننده نوع درمان، پیش-آگهی و میزان بقا است. هدف اصلی از این مقاله، انتخاب یک الگوریتم مناسب است که بتواند مرحله بیماری فرد مبتلا به سرطان پستان را تعیین کند.

روش بررسی: در این مطالعه توصیفی- تحلیلی، داده‌های اندازه تومور، درگیری غدد لنفاوی و متاستاز مربوط به ۷۳۲ بیمار مبتلا به سرطان پستان در بیمارستان ولی‌عصر بیرجند مورد استفاده قرار گرفت. از این داده‌ها جهت انتخاب الگوریتمی مناسب از میان الگوریتم‌های درخت تصمیم، شبکه‌های عصبی مصنوعی، شبکه‌های بیزین و K- نزدیک‌ترین همسایه برای تشخیص مرحله بیماری فرد جدید استفاده شده است.

یافته‌ها: بررسی‌ها نشان می‌دهد که الگوریتم K- نزدیک‌ترین همسایه می‌تواند با احتمال ۹۶٪ به درستی مرحله بیماری فرد مراجعه‌کننده را با اندازه‌گیری اندازه تومور، میزان درگیر بودن غدد لنفاوی و میزان متاساز را تشخیص دهد.

نتیجه‌گیری: الگوریتم K- نزدیک‌ترین همسایه در تشخیص مرحله بالینی سرطان پستان، بیشترین دقت را دارد.

واژه‌های کلیدی: سرطان پستان، الگوریتم های داده‌کاوی، درخت تصمیم، شبکه عصبی مصنوعی.

* نشانی نویسنده پاسخگو: یزد، دانشگاه آزاد اسلامی واحد علوم تحقیقات یزد، گروه مهندسی کامپیوتر، سید عباس محمودی.
نشانی الکترونیک: sa_mahmoodi_85@yahoo.com

[DOR: 20.1001.1.17359406.1393.7.2.4.4]

مقدمه

در حال حاضر سرطان یکی از علل عمده مرگ و میر در جهان است که بر اثر عوامل مختلفی مانند: مواد جهش‌زا و مواد شیمیایی سرطان‌زا در محیط به وجود می‌آید. بر طبق تحقیقات انجام شده ممکن است بیش از ۷۵ درصد سرطان‌ها دارای منشا محیطی باشند (۱). البته آسیب‌های ژنتیکی، تغییرات ژنتیکی ایجاد شده در توالی DNA، بروز جهش در ژن‌ها نقش بسزایی در سرطان زایی دارند (۲).

سرطان پستان یکی از سرطان‌های شایع در بین زنان دنیاست و آمار جهانی نشان می‌دهد که شیوع این بیماری در حال افزایش است (۳ و ۴). در دنیا سالیانه بیش از ۸ میلیون نفر به انواع سرطان مبتلا می‌شوند که در حدود یک میلیون نفر از آن مربوط به سرطان پستان است. متاسفانه این نوع سرطان در بین زنان ایرانی نیز در حال پیشرفت است (۵). به گزارش وزارت بهداشت، درمان و آموزش پزشکی، در سال ۱۳۷۴ سرطان پستان در میان سرطان‌های شایع زنان رتبه دوم را داشته است درحالی‌که در سال ۱۳۷۵ به مقام اول صعود کرد. روند رو به رشد این بیماری در ایران همچنان ادامه دارد (۶ و ۷). به گزارش مرکز مدیریت بیماری‌های وزارت بهداشت در سال ۱۳۸۲، سرطان پستان با شیوع ۱۵/۹٪ رتبه اول را به خود اختصاص داده است (۸ و ۹). این بیماری شایع‌ترین علت مرگ و میر ناشی از سرطان در میان انواع مختلف سرطان‌ها است (۱۰).

متاسفانه سن شروع سرطان پستان در ایران همچون دیگر کشورهای در حال توسعه، پایین‌تر از کشورهای پیشرفته است. شایع‌ترین سن مرگ و میر ناشی از سرطان پستان در ایران ۴۰ تا ۴۹ سالگی است، درحالی‌که در کشورهای پیشرفته ۵۵ تا ۶۰ است. بنابراین در ایران حداقل یک دهه پایین‌تر از کشورهای پیشرفته است و با توجه به نقش محوری زنان در این سن در خانواده و جامعه، مرگ و میر و ناتوانی ناشی از بیماری صدمات جبران ناپذیری به جامعه و خانواده وارد می‌کند. محققان میزان بالای مرگ و میر زنان بر اثر سرطان پستان را ناشی از تشخیص دیرهنگام این بیماری می‌دانند و موفقیت کشورهای پیشرفته در کنترل مرگ و میر و سایر پیامدهای ناشی از بیماری را در دو گروه تشخیص به موقع (زودرس) آن دانسته‌اند زیرا بقای فرد به طور مستقیم در ارتباط با مرحله بیماری در زمان تشخیص است (۱۱).

از عوامل خطرساز سرطان پستان، می‌توان به سابقه فامیلی، سن اولین حاملگی، شروع زودرس قاعدگی و پایان دیررس آن، چاقی، مصرف الکل و بی‌تحرکی فیزیکی اشاره کرد (۱۲). متاسفانه سرطان پستان درمان قطعی ندارد و هنوز مطالب زیادی در مورد پیشگیری از آن نمی‌دانیم (۱۳). این بیماری از نظر بالینی از یک مرحله مخفی طولانی عبور می‌کند و حدود ۱۰ تا ۸ سال طول می‌کشد تا یک سلول سرطانی به تومور قابل لمس تبدیل شود. بنابراین با شناسایی و تشخیص این توده‌ها در مراحل اولیه، می‌توان جان بیماری را از مرگ نجات داد (۱۲). از نظر علائم و نشانه‌های بیماری باید گفت، وقتی تومور کوچک است، هیچ علامتی ندارد و وقتی تومور به اندازه کافی بزرگ شد که لمس شود، درد ندارد. پس مهم‌ترین علامت، در واقع وجود توده بدون درد است. البته ممکن است علائمی غیرشایع شامل درد، تغییر قوام پستان مثل سفتی، ورم، قرمزی، بدشکلی و خارش و سوزش پوست، فرو رفتگی نوک پستان و ترشح خود به خودی وجود داشته است (۱۳و۹). بنابراین علائمی وجود دارد که می‌توان سرطان را تشخیص داد. از طرفی ممکن است بعضی از این علائم در سایر بیماری‌ها نیز وجود داشته باشد.

بنابراین پزشک برای اتخاذ یک تصمیم مناسب، باید نتیجه آزمایش‌های بیمار و تصمیم‌هایی که در گذشته برای بیماران با وضعیت مشابه گرفته است، را بررسی کند. به عبارت دیگر پزشک نیازمند دانش و تجربه خواهد بود. ولی به دلیل تعداد زیاد بیماران و آزمایش‌های متعدد هر بیمار، نیاز به یک ابزار خودکار برای کاوش در میان بیماران قبلی احساس می‌شود (۱۴). این تحقیق الگوریتمی را معرفی می‌کند که بتواند مرحله بالینی سرطان پستان را به سرعت تعیین کند.

در صورتی که این بیماری در مرحله اول (Stage 1)، یعنی زمانی که سرطان محدود به پستان است، تشخیص داده شود، ۷۵ تا ۹۰٪ از زنان، از زندگی پنج ساله سالمی برخوردار خواهند بود. چنانچه در مرحله دوم بیماری (Stage 2) که سرطان به غدد لنفاوی دست-اندازی کرده است، تشخیص داده شود، احتمال بقای

DOR: 20.1001.1.17359406.1393.7.2.4.4]

پنج ساله بیماری به ۱۶٪ کاهش می‌یابد. ضمن اینکه درمان سرطان پستان در مراحل پیشرفته با مراحل اولیه متفاوت است (۹). نکته دیگر اینکه در کلیه مطالعات انجام شده در مورد همه سرطان‌ها و سرطان پستان، برای مقایسه متغییرهای مورد بررسی، لازم است که مرحله بالینی بیماری مشخص باشد. در صورتی‌که نرم افزاری وجود داشت که به سرعت مرحله بالینی بیماری را تعیین و ثبت نماید کمک بزرگی به محققین کرده است. به همین دلیل در این مقاله برای امکان‌پذیر ساختن تعیین مرحله بالینی بیماری از هوش مصنوعی و الگوریتم‌های داده‌کاوی استفاده شده است، تا خطای احتمال ناشی از تشخیص پزشک تا حد امکان کاهش یابد. روش‌های داده‌کاوی می‌توانند در تشخیص Stage بیماری به پزشکان کمک شایانی نمایند. در نتیجه، رویکردهای جدید مانند کشف دانش از پایگاه داده (KDD[1])، که شامل تکنیک‌های داده‌کاوی هستند، روز به روز محبوبیت بیشتری یافته و تبدیل به یک ابزار تحقیقاتی مطلوب برای پژوهشگران علوم پزشکی شده‌اند.

در کارهای گذشته با استفاده از داده‌کاوی در زمینه پیش‌بینی بقا بیماران، خوش‌خیم یا بدخیم بودن و همچنین در زمینه پیش‌بینی عود مجدد سرطان پستان (۱۵) استفاده شده است. به عنوان مثال در تحقیقی که در سال ۲۰۰۷ توسط گروهی از محققین دانشگاه توکیو انجام گردید، از شبکه‌های عصبی مصنوعی در پیش‌بینی بقا بیماران مبتلا به سرطان پستان استفاده شده است (۱۶). قیومی‌زاده و همکاران از ترکیب شبکه عصبی خودسازمان‌ده (SOM[2]) و شبکه پرسپترون چند لایه (MLP[3]) برای تعیین خوش‌خیم یا بدخیم بودن سرطان پستان استفاده شده است. در شبکه خود-سازمان‌ده، از روش یادگیری رقابتی برای آموزش استفاده می‌شود و مبتنی بر مشخصه‌های خاصی از مغز انسان توسعه یافته است (۱۷).

در این پژوهش، سعی می‌شود از میان الگوریتم‌های داده‌کاوی، الگوریتمی که دارای بیشترین دقت در تشخیص مرحله سرطان پستان است انتخاب شود.

مواد و روش‌ها

این مطالعه تحلیلی بر روی داده‌های بیماران، از یک مجموعه داده‌ی سرطان پستان که از بیماران بیمارستان ولی‌عصر بیرجند مربوط به سال‌های ۱۳۹۱ تا ۱۳۸۲جمع‌آوری شده‌اند، صورت گرفته شده است. پایگاه داده شامل ۹۳۶ نمونه بود. که تعداد ۲۰۴ نمونه دارای اطلاعات ناقص بودند که از مطالعه خارج شدند. از آنجایی که برای تعیین مرحله سرطان بالینی سرطان پستان از سیستم TNM، تومور و نودها و متاساز، استفاده می‌کند، این داده‌ها شامل سه ویژگی اندازه تومور، درگیری غدد لنفاوی و متاساز دور دست هستند. با توجه به مقادیر این سه ویژگی، داده‌ها خود به چهار مرحله تقسیم می‌شود. بنابراین بر اساس میزان این سه ویژگی، مرحله هر نمونه تشخیص داده می‌شود. زمانی که گروه‌های T، N و M یک بیمار تعیین می‌شود، می‌توان مرحله بیماری آن فرد را مشخص کرد. این مراحل یا کلاس‌ها از یک تا چهار تعریف می‌شوند (۱۸).

مرحله یک:

مرحله یک یا مرحله شروع بیماری در این مرحله، سرطان محدود به منطقه ظهور تومور، و اندازه غده هم کوچک است.

مرحله دو:

در این مرحله اندازه تومور بیشتر از ۲ سانتی‌متر نیست، بیماری پراکنده نشده و غدد لنفاوی زیر بغل را مبتلا نکرده است.

مرحله سه:

در این مرحله اندازه تومور بین ۵ تا ۲ سانتی‌متر است و ممکن است به غده‌های لنفاوی زیر بغل هم سرایت کرده باشد.

مرحله چهارم:

اندازه تومور در این مرحله بیش از ۵ سانتی‌متر است و گسترش بیشتری در غدد لنفاوی زیر بغل دارد. ضمناً به دیگر غدد لنفاوی و یا بافت‌های مجاور پستان هم سرایت کرده است. مشخصات ویژگی‌های این مجموعه داده در جدول ۱ آمده است.

[1] Knowledge Discovery in Database
[2] Self Organization Map
[3] Multi Layer Perceptron

[DOR: 20.1001.1.17359406.1393.7.2.4.4]

جدول۱: مشخصات ویژگی های مجموعه داده

مشخصات	ویژگی
اندازه تومور کمتر یا مساوی ۲ سانتی‌متر	T1
اندازه تومور بیشتر از ۲ و کمتر مساوی ۵ سانتی-متر	T2
اندازه تومور بزرگتر از ۵ سانتی‌متر	T3
درگیری لنف نود مشخص نیست	N0
۳ تا ۱ لنف نود درگیر باشد	N1
۹ تا ۴ لنف نود درگیر باشد	N2
۱۰یا بیشتر از ۱۰ لنف نود درگیر باشد	N3
متاساز مشخص نباشد	M0
متاساز وجود داشته باشد	M1

در این تحقیق از نرم‌افزار وکا(WEKA) نسخه ۳.۶ و الگوریتم‌های داده‌کاوی استفاده شده است. نرم‌افزار وکا مشتمل بر مجموعه‌ای از روش‌ها از قبیل دسته‌بندی، خوشه‌بندی و قوانین انجمنی و انتخاب ویژگی است. در دسته‌بندی، هر داده به یک کلاس از پیشین مشخص شده تخصیص می‌یابد ولی در خوشه‌بندی هیچ اطلاعی از کلاس‌های موجود درون داده‌ها وجود ندارد و به عبارتی خود خوشه‌ها نیز از داده‌ها استخراج می‌شوند (۱۹). در یادگیری با نظارت از ابتدا دسته‌ها مشخص هستند و هر یک از داده‌های آموزشی به دسته‌ای خاص نسبت داده شده است و اصطلاحاً گفته می‌شود ناظری وجود دارد که در هنگام آموزش، اطلاعاتی علاوه بر داده‌های آموزشی در اختیار یادگیرنده قرار می‌دهد. ولی در یادگیری بدون نظارت هیچ اطلاعاتی بجز داده‌های آموزشی در اختیار یادگیرنده قرار ندارد و این یادگیرنده است که بایستی در داده‌ها به دنبال ساختاری خاص بگردد (۲۰). در اینجا از عمل دسته‌بندی داده‌ها استفاده می‌شود و برای ارزیابی مدل، داده‌های جدید به طور تصادفی به کمک اعتبارسنجی متقابل(۱۰لایه)، به دو دسته آموزشی و آزمون مستقل تقسیم می‌شوند.

از مهم‌ترین الگوریتم‌های داده‌کاوی می‌توان به الگوریتم‌های درخت تصمیم[4]، شبکه عصبی مصنوعی[5]، شبکه‌های بیزین[6] و K- نزدیک‌ترین همسایه[7] اشاره کرد.

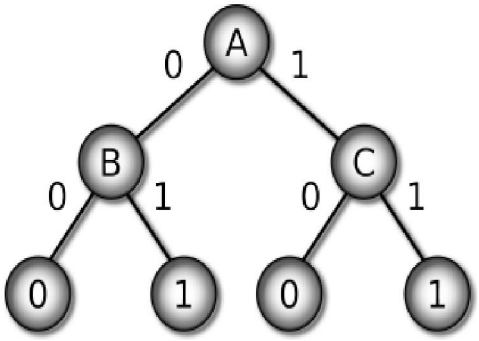
[4] Decision Tree
[5] Artificial Neural Network
[6] Bayesian Network
[7] K-nearest Neighbor

درخت تصمیم

یکی از رایج‌ترین تکنیک‌های دسته‌بندی، درخت‌های تصمیم بالا به پایین هستند. درخت تصمیم، درختی است که در آن نمونه‌ها را به نحوی دسته‌بندی می‌کند که از ریشه به سمت پایین رشد می‌کنند و در نهایت به گره‌های برگ می‌رسد. هر گره داخلی با یک ویژگی مشخص می‌شود. برگ‌های این درخت با یک کلاس و با یک دسته از جواب‌ها تعیین می‌گردند.

برای دسته‌بندی یک ورودی در این درخت، از ریشه درخت شروع می‌کند و شاخه‌ها را بر طبق مقدار ویژگی ورودی دنبال می‌نماید تا به یک برگ برسد. خروجی مقدار برگ به عنوان دسته‌عنصر ورودی در نظر گرفته می‌شود. شکل ۱، جدول درستی و درخت تصمیم مربوطه آن را نمایش می‌دهد (۲۱).

A	B	C	Class
0	0	0	0
0	0	1	0
0	1	0	1
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1



شکل ۱: جدول درستی و درخت تصمیم مربوطه آن

از مهم‌ترین الگوریتم‌های یادگیری درخت تصمیم می‌توان الگوریتم قدرتمند C4.5 نام برد. اکثر الگوریتم‌های درخت‌های تصمیم، با ساختن یک درخت از بالا به پایین

DOR: 20.1001.1.17359406.1393.7.2.4.4]

به کمک انتخاب صفات در هر لحظه و جداسازی داده‌ها با توجه به مقادیر صفات‌شان، ایجاد می‌گردند (۲۱).

شبکه عصبی مصنوعی

یک شبکه عصبی مصنوعی روشی برای پردازش اطلاعات است که از سیستم‌های عصبی زیستی الهام گرفته شده و مانند مغز پردازش اطلاعات انجام می‌گیرد. این شبکه‌ها قادر به مدل‌سازی توابع غیرخطی است که قادر به پیش‌بینی مشاهدات جدید پس از اجرای یک فرآیند یادگیری به اصطلاح از داده‌های موجود هستند (۱۵).

شبکه‌های بیزین

در مقایسه با شبکه‌های عصبی، این گروه از دسته‌بندی‌ها، میزان عضویت یک نمونه به هر کلاس را بایک احتمال نشان می‌دهد، همچنین از مفاهیم آماری مانند میانگین، انحراف معیار و یا از هیستوگرام مقادیر ویژگی، برای تولید قانون استفاده می‌نمایند (۲۲). یکی از مهم‌ترین روش‌های دسته‌بندی آماری شبکه‌های[8] NB و شبکه‌های بیزی هستند. شبکه بیزی یک مدل گرافیکی است که رابطه احتمالی بین یک مجموعه از متغیرها را بیان می‌کند. ساختار یک شبکه بیزی S، یک گراف جهت‌دار بدون دور است که گره‌ها در آن متغیرهای تصادفی را نشان می‌دهد و یال‌های آن شبکه، یک ارتباط یک به یک بین متغیرها می‌باشد (۲۳).

شبکه‌های NB، شبکه‌های بیزی خیلی ساده‌ای هستند که از گراف‌های بدون دور جهت‌دار، با تنها یک والد و چندین فرزند تشکیل شده‌اند و نودهای فرزندان را مستقل در نظر می‌گیرد. این الگوریتم احتمال شرطی هر ویژگی داده شده را با توجه به دسته مربوطه‌اش یاد می‌گیرد. سپس عمل دسته‌بندی با بکار بردن قوانین بیز برای محاسبه مقدار احتمالی دسته نتیجه نمونه داده شده، با دقت بالایی انجام می‌شود. مهم‌ترین مزیت دسته‌بندی کنندهNB، زمان کمتر محاسبه برای آموزش دیدن است (۲۴).

K- نزدیک‌ترین همسایه

الگوریتم k- نزدیک‌ترین همسایه (K-NN) برای دسته بندی نمونه‌ها استفاده می‌شود، و براساس این اصل است که یک نمونه با K نمونه که خصوصیات مشابه بیشتری با هم دارند، دسته‌بندی شود. روال این الگوریتم به این صورت است که هنگامی که داده جدیدی (دسته‌بندی نشده) به منظور دسته‌بندی وارد می‌شود، شروع به کار کرده و آن را با تمام داده‌های از قبل دسته‌بندی شده مقایسه می‌کند و K تا از نمونه‌های نزدیک را برای نمونه ی جدید شناسایی کرده و برچسب کلاسی که بیشترین تکرار را در میان این نمونه‌ها دارد به عنوان کلاس نتیجه آن تعیین می‌کند. بنابراین باید معیاری را برای تعیین فاصله بین نمونه‌ها مشخص کرد. این فاصله، باید فاصله بین نمونه‌های یک کلاس را مینیمم و فاصله بین نمونه‌های کلاس‌های متفاوت را ماکزیمم کند (۲۵).

در این بررسی برای مقایسه مدل‌ها و دسته‌بندی داده‌ها تکنیک داده‌کاوی مورد استفاده قرار گرفتند که برای این کار از نرم‌افزار وکا استفاده شده است. از آنجا که یک مرحله مهم در الگوریتم‌های داده‌کاوی مرحله آموزش است، بنابراین نرم افزار، داده‌ها را به دو دسته داده‌های آموزش و داده‌های آزمون تقسیم می‌کند. نرم‌افزار وکا یک مجموعه داده که توسط کاربر تعریف می‌شود را به عنوان ورودی دریافت کرده و با انتخاب روش خواسته شده سعی در دسته‌بندی داده‌ها می‌نماید. چون در هر بار اجرا، نتایج متفاوتی بدست می‌آید، می‌توان از روش 10F-CV استفاده کرد. در این روش کل مجموعه داده به ده قسمت مساوی تقسیم شده و الگوریتم ده بار اجرا می‌شود. در هر بار نه قسمت از این داده‌ها را برای آموزش استفاده می‌کند و یک قسمت باقی‌مانده را به عنوان داده آزمون برای ارزیابی کردن الگوریتم در نظر می‌گیرد. میانگین نتایجی که بدست می‌آید به عنوان نتیجه نهایی گزارش می‌شود. در مرحله آموزش، وکا تعدادی قانون را استخراج می‌نماید. سپس از این قوانین استخراج شده استفاده کرده و سه ویژگی اندازه تومور، میزان درگیر بودن لنف نود و میزان متاساز بیمار جدید را به الگوریتم وارد نموده، و مرحله سرطان را تشخیص می‌دهد. حال برای ارزیابی مدل، داده-های آزمون را به مدل وارد کرده و خروجی مدل با دسته‌های مشخص توسط مدل مقایسه می‌گردد و سپس دقت کل مدل استخراج می‌شود. در واقع برچسب شناخته شده از نمونه آزمون با نتایج دسته‌بندی مقایسه می‌شود. دقت مدل یا نرخ دسته‌بندی با توجه به جدول ۲ محاسبه می‌شود.

[8] Naïve Bayes

جدول۲: ماتریس درهم ریختگی

نوع دسته		دسته تشخیص داده شده	
		مثبت	منفی
دسته واقعی	مثبت	TP	FN
	منفی	FP	TN

دسته‌ها در مسأله تشخیص مرحله سرطان با چهار دسته مثبت و منفی و چهار عدد TP، FP، FN و TP با توجه به نوع دسته مثبت و منفی محاسبه می‌گردند.

TP، شامل نمونه‌هایی است که جز نمونه‌های دسته مثبت است و الگوریتم آن را به‌درستی در دسته مثبت تشخیص داده است.

FP، شامل نمونه‌هایی است که جز نمونه‌های دسته منفی است و الگوریتم آن را به صورت نادرستی در دسته مثبت تشخیص داده است.

FN، شامل نمونه‌هایی است که جز نمونه‌های دسته مثبت است و الگوریتم آن را به صورت نادرستی در دسته منفی تشخیص داده است.

TN، شامل نمونه‌هایی است که جز نمونه‌های دسته منفی است و الگوریتم آن را بدرستی در دسته منفی تشخیص داده است.

نرخ دسته‌بندی بیانگر دقت الگوریتم پیاده‌سازی شده در دسته‌بندی دسته‌های مختلف موجود در مسأله تشخیص مرحله سرطان است. این معیار در واقع درصد دسته‌بندی درست الگوریتم می‌باشد (۲۶). به عبارتی نرخ دسته‌بندی، تعداد نمونه‌هایی می‌باشد که به درستی دسته‌بندی شده است، و نسبت تعداد نمونه های است که به درستی تشخیص داده می شوند به کل نمونه‌ها (۲۷):

(۱)

$$Classification\ Rate = \frac{(TP + TN)}{(TP + TN + FN + FP)}$$

نرخ صحت یا میزان حساسیت که برای هر کدام از دسته‌های موجود قابل محاسبه می‌باشد، جهت تعیین دقت دسته‌بندی برای هر کدام از دسته‌ها در نظر گرفته شده است. در واقع این معیار نشان دهنده درصد موفقیت روش دسته‌بندی کننده در تشخیص نمونه‌های مربوط به هرکدام از دسته‌ها می‌باشد (۲). نرخ فراخوانی یا ویژگی که همانند معیار قبل برای هر کدام از دسته‌های موجود محاسبه می‌گردد، درصد قابلیت اعتماد به خروجی روش دسته‌بندی کننده را نشان می‌دهد (۳). در واقع احتمال پیش‌بینی درست عدم وجود وضعیت مورد نظر توسط الگوریتم‌ها است.

$$Precision = \frac{TP}{TP \mid FP} \qquad (2)$$

$$Recall = \frac{TP}{TP + FN} \qquad (3)$$

یافته‌ها

همچنان که گفته شد این تحقیق بر روی داده‌های بیماران، از یک مجموعه داده‌ی سرطان پستان که از بیمارانبیمارستان ولی عصر بیرجند جمع‌آوری شده‌اند، صورت گرفته شده است، که شامل ۷۳۲ بیمار می‌باشد. بر روی هر بیمار سه ویژگی اندازه تومور، میزان درگیر بودن غدد لنفاوی و میزان متاساز اندازه‌گیری شده است. جدول ۳ مشخصات این مجموعه داده را نمایش می‌دهد.

جدول ۳ : فراوانی بیماران مورد بررسی برحسب نوع ویژگی

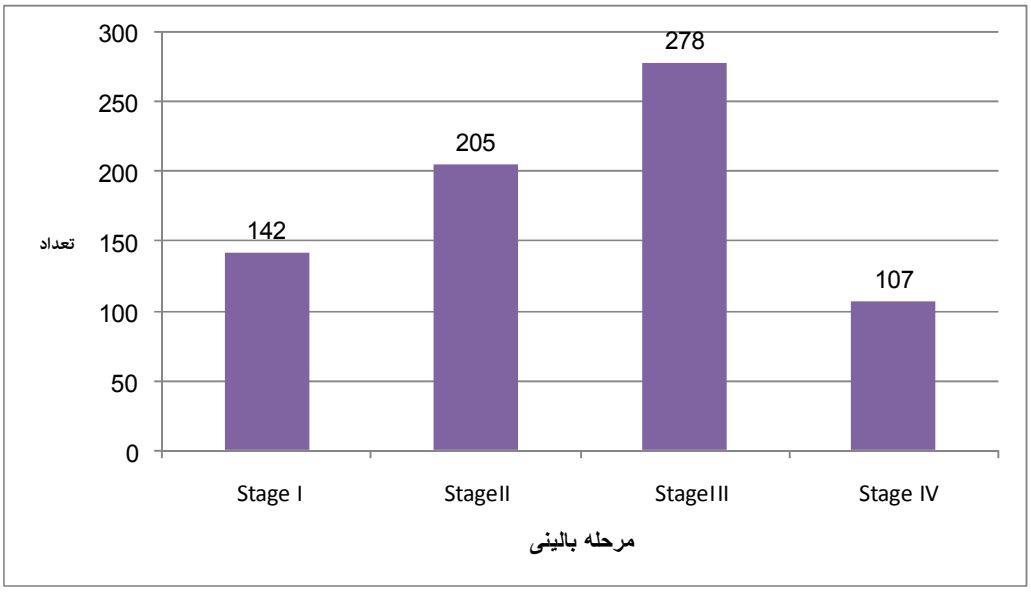
ویژگی	فراوانی	درصد
T1	۳۲۶	۴۴.۵۴
T2	۲۷۵	۳۷.۵۷
T3	۱۳۱	۱۷.۸۹
N0	۲۹۲	۳۹.۹۰
N1	۲۰۱	۲۷.۴۶
N2	۱۳۳	۱۸.۱۶
N3	۱۰۶	۱۴.۴۸
M0	۶۲۵	۸۵.۳۸
M1	۱۰۷	۱۴.۶۲

باتوجه به داده‌های جدول فوق Stage بیماران بر اساس سیستم TMN محاسبه گردیده است، که ۲۷۸ نفر در Stage III بوده است (نمودار ۱).

جدول ۴ نرخ دسته بندی، و نرخ صحت و نرخ فراخوانی را برای چهار الگوریتم داده‌کاوی نشان می‌دهد.

بر اساس نتایج بدست آمده از روش‌های داده‌کاوی می‌توان مدلی ارائه نمود که بتواند در تصمیم‌گیری در تشخیص مرحله سرطان به پزشک کمک شایانی نماید. در صورتی که از داده‌های بیشتری برای آموزش استفاده شود، نتایج بهتری حاصل می‌شود.

DOR: 20.1001.1.17359406.1393.7.2.4.4]



نمودار۱: فراوانی بیماران مورد بررسی برحسب مرحله بالینی

جدول۴: مقایسه نتایج بدست آمده با روش‌های داده‌کاوی

نام‌الگوریتم	دقت	حساسیت	ویژگی
شبکه‌های عصبی	٪۹۵.۱	٪۹۵.۳	٪۹۵.۲
شبکه بیزی	٪۹۵.۹	٪۹۵.۹	٪۹۵.۷
k- نزدیک‌ترین همسایه (K-NN)	٪۹۶	٪۹۶.۱	٪۹۶
درخت تصمیم (C4.5)	٪۹۳.۴	٪۹۳.۴	٪۹۳.۸

بحث

در این مطالعه از ۴ الگوریتم مختلف (درخت تصمیم، شبکه عصبی مصنوعی، K- نزدیک‌ترین همسایه و شبکه بیزین) برای تعیین مرحله بالینی سرطان پستان مورد استفاده قرار گرفت. اگر مرحله بالینی بیماری به راحتی تعیین و ثبت شود برای فالوآپ بیماری و برای انجام تحقیقات مفید است (۳). داده‌کاوی قادر است از میان حجم زیادی از داده‌ها دانش نهفته را کشف نماید.

نرم‌افزارهای فراوانی برای داده‌کاوی و یادگیری ماشین در حوزه‌های مختلف داده‌ها موجود می‌باشند. هریک از آنها با توجه به نوع اصلی داده‌هایی که مورد کاوش قرار می‌دهند، روی الگوریتم‌های خاصی متمرکز شده‌اند. مقایسه دقیق و علمی این ابزارها باید از جنبه‌های متفاوت و متعددی مانند تنوع انواع و فرمت داده‌های ورودی، حجم ممکن برای پردازش داده‌ها، الگوریتم‌های پیاده سازی شده، روش‌های ارزیابی نتایج، روش‌های مصور سازی، روش‌های پیش‌پردازش داده‌ها، واسطهای کاربر پسند، قیمت و در دسترس بودن نرم‌افزار صورت گیرد. از آن میان، نرم‌افزار Weka با داشتن امکانات بسیار گسترده، امکان مقایسه خروجی روش‌های مختلف با هم، راهنمای خوب، واسط گرافیگی کارآ، سازگاری با سایر برنامه‌های ویندوزی و از همه مهم‌تر رایگان بودن نرم افزار معرفی می‌شود. این نرم‌افزار و نرم‌افزارهای مشابه داده‌کاوی شامل الگوریتم‌های هوش مصنوع می‌باشند که می‌توان از آنها برای استخراج اطلاعات نهفته در داده‌ها استفاده کرد. در این مطالعه برای تعیین مرحله بالینی از چهار الگوریتم درخت تصمیم، شبکه عصبی مصنوعی، K-

[DOR: 20.1001.1.17359406.1393.7.2.4.4]

نزدیک‌ترین همسایه و شبکه بیزین استفاده شد. زمانی که گروه‌های T، N و M یک بیمار تعیین می‌شود، می‌توان مرحله بیماری آن فرد را مشخص کرد.

نتیجه‌گیری

نتایج نشان می‌دهد که هر چهار روش، روش‌های کارایی هستند (دقت از ۹۳/۴ تا ۹۶ درصد) اما دقت تشخیص مرحله سرطان در الگوریتم K- نزدیک‌ترین همسایه از همه بیشتر و در الگوریتم درخت تصمیم C4.5 از همه پایین‌تر است. در مطالعه گنجی و همکاران دقت دسته‌بندی بیماران دیابتیک ۷۵٪ و در بیماران مبتلا بهٔ سرطان پستان (خوش خیم یا بدخیم) ۹۵٪ بوده است که مشابه مطالعه حاضر می‌باشد. در مطالعه گنجی از الگوریتم کلونی مورچه‌ها استفاده شده بود (۲۸). در مطالعه Nguyen و همکاران که به منظور تعیین Stage بیماران مبتلا به سرطان ریه انجام شد، میزان صحت اندازه تومور (T) به میزان ۷۲٪ درگیری غدد لنفاوی (N) ۷۸٪ و در ارتباط با متاستاز (M) ۹۴٪ بوده است، میزان صحت مطالعه ما بیشتر از این مطالعه بوده است (۲۹). همچنین در مطالعه Zhou و همکاران که از ترکیب الگوریتم C4.5 و شبکه‌های عصبی استفاده شده است، دقت دسته‌بندی بیماران مبتلا به سرطان پستان ۹۴٪ تعیین شده است (۳۰).

افزایش دقت در تعیین مرحله بالینی سرطان پستان به بزرگتر بودن پایگاه داده‌ها بستگی دارد. بنابراین در تحقیقات آتی می‌توان با استفاده از پایگاه داده‌های بزرگ‌تر، تعداد رکوردها را افزایش داد و دقت الگوریتم را افزایش داد. از این روش می‌توان همچنین برای انواع دیگر سرطان‌ها و در مراکز انکولوژی استفاده نمود.

References

1. Moller P, Wallin H, Knudesen LE. Oxidative stress associated psychological stress and life-style factor. Chem Bio Interact 1996; 102: 1-36.
2. Namiki M. Antioxidant /antimutagenes in foods. Crit Rev. food SciNutr 1990; 29: 273-300.
3. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics 2002. CA Cancer J Clin 2005; 55(2): 74-108.
4. Wilson CM, Tobin S, Young RC. The exploding worldwide cancer burden: the impact of cancer on women. Int J Gynecol Cancer 2004; 14: 1-11.
5. Mousavi SM, Montazeri A, Mohagheghi MA, MousaviJarrahi A, Harirchi I, Najafi M, Ebrahimi M. Breast cancer in Iran: an epidemiological review. Breast J 2007; 13: 383-91.
6. Iran Ministry of Health & Medical Education. Health deputy, Family Health office. Country plan of adult health and women prevalent cancers. 1 st ed. Tehran: Ministry of Health& Medical Education; 2002.
7. Iran Ministry of Health & Medical Education. Health deputy, Family Health office, adult health and women office. Primary report of breast cancer screening 1st ed. Tehran: Ministry of Health & Medical Education 2000; 18-36.
8. Iran Ministry of Health & Medical Education. Diseases management center, cancer office. Country report of cancer cases. Tehran: KelkZarrin Press 2004; 16.
9. Sadeghnezhad F, NiknamiSh, Ghaffari M, Effect of health education methods on promoting breast self examination (BSE), Journal of Birjand University of Medical Sciences 2009; 15(4): 38-48.
10. Keramatee K, Ghorbanian M, Abbasnia V, PazirehN, Alipour H, Effect of Flunixin as a Cox Inhibitor on Prevention and Cure of Breast Cancer in Female Wistar Rat, the Horizon of Medical Sciences 2010; 15 (4): 24-32.

۱۱. لمیعیان م، حیدرنیا ع، احمدی ف، فقیه زاده س، آگیلاروفایی م. رفتار کنترل سرطان پستان از نگاه زنان: یک پژوهش کیفی. مجله علمی دانشگاه علوم پزشکی بیرجند، پاییز۱۳۸۷؛ ۱۵(۳):۸۸-۱۰۲.

DOR: 20.1001.1.17359406.1393.7.2.4.4]

12. Fentiman IS. Fixed and modifiable risk factors for breast cancer. Int J Clin Prat 2001; 55(8): 527-30.

13. Herdy V. Education: a key factor in fighting breast cancer. New York: Inter press services 1998; 1.

14. Andres C, Pena R, Sipper M. Designing Breast Cancer Diagnostic Systems via aHybrid Fuzzy-Genetic Methodology, IEEE International Fuzzy Systems Conference 1999; 1: 135-9.

15. Ashlaghi A, Pour Ebrahimi A, Ebrahimi M, Ahmad L. Using data mining techniques for prediction breast cancer recurrence, Iranian Journal of Breast Disease 2013; 5(4): 23-34.

16. Endo A, Shibata T, Tanaka H. Comparison of seven algorithms to predict breast cancer survival, Biomedical Soft Computing and Human Sciences 2008; 13(2): 6-11.

۱۷. قیومی‌زاده حسین، درودگرمقدم علی، حدادنیا جواد، محمدزاده محمد، رحمانی سریاست امید. خوشه‌بندی و غربالگری سرطان پستان براساس تصاویر حرارتی با استفاده از ترکیب شبکه عصبی SOM و MLP. فصلنامه بیماری‌های پستان ایران، ۱۳۹۱؛ ۵ (۲و۳):۸۳-۷۰.

18. Ganji MF, Abadeh MS. Using Fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease. Iranian conference Electrical Engineering, ICEE 2010.

۱۹. صنیعی آباده محمد، محمودی سینا، طاهرپور محدثه، داده کاوی کاربردی، چاپ اول تهران: انتشارات نیاز دانش، ۱۳۹۱.

۲۰. قیومی‌زاده حسین، درودگر مقدم، حدادنیا جواد، محمدزاده محمد، رحمانی سریاست امید. خوشه‌بندی و شناسایی کردن سرطان پستان توسط تصاویر حرارتی به کمک ترکیب شبکه عصبی SOM و SVM. فصلنامه بیماری‌های پستان ایران،۱۳۹۱؛ ۵ (۴): ۲۲-۱۳.

21. Aguilar-Ruiz JS, Riquelme JC, Toro M. Evolutionary Learning of Hierarchical Decision Rules, in IEEE Transactions on Systems 2003; 33(2): 324-34.

22. Jain R, Abraham A. A Comparative Study of Fuzzy Classification Methods on Breast Cancer Data, Australas Phys Eng Sci Med 2004; 27(4): 213-8.

23. Kotsiantis SB. Supervised machin leaming: a reviw of classification techniques, informatica 31, 2007; 249-68.

24. John GH, Langley P. Estimating continuous distributions in bayesian classifiers. In Proceedings of the Eleventh Conference on Unccertanity in Artificial Intelligence 1995; 338-345.

25. Jain R, Mazumdar J. A Genetic Algorithm based Nearest Neighbor Classification to Breast Cancer Diagnosis, Australasian Physical & Engineering Sciences in Medicine 2003; 26(1):6-11.

26. Rosai J. Surgical pathology, Elsevier Inc. 9th edition Piladerphia 2004; 380-92.

۲۷. قاسم‌احمد لیلا. مروری بر ۷ الگوریتم برتر داده‌کاوی در پیش‌بینی بقا، تشخیص و عود بیماران مبتلا به سرطان پستان. فصلنامه بیماری‌های پستان ایران، ۱۳۹۲؛ ۶ (۱): ۶۱-۵۲.

28. Ganji MF, Abadeh MS. Parallel Fuzzy Rule Learning Using an ACO-Based Algorithm for Medical Data Mining, IEEE Fifth. International conference on Bio-InnspiredCompting: theories and Applications 2010; 573-81.

29. Nguyen AN, Lawley MJ, Hansen DP, Bowman RV, Clarke BE, Duhig EE, Colquist S. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. J Am Med Inform Assoc 2010; 17(4):440-5.

30. Zhou Zh, Jiang Y. Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans InfTechnol Biomed 2003; 7(1): 37-42.

[DOR: 20.1001.1.17359406.1393.7.2.4.4]