مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره چهارم شماره چهارم. ص ۲۷۸-۲۶۶، ۱۳۹۶

مدل‌سازی بیماری سرطان پستان با استفاده از روش‌های مبتنی بر داده‌کاوی

پروانه دهقان[۱]، مائده مقربی[۲]، ایمان ذباح[۳]*، کامران لایقی[۴]، علی ماروسی[۵]

• **دریافت مقاله:** ۹۶/۴/۲۳ • **پذیرش مقاله:** ۹۶/۹/۵

مقدمه: سرطان سینه رایج‌ترین شکل سرطان در زنان است. اهمیت تشخیص سرطان سینه به عنوان یکی از موضوعات مهم در علم پزشکی مطرح می‌شود. تشخیص خوش‌خیم یا بدخیم بودن سرطان علاوه بر کاهش هزینه‌ها در جهت‌گیری نوع درمان از اهمیت فوق‌العاده ای برخوردار است. هدف از این پژوهش ارائه مدل‌هایی بر اساس داده‌کاوی است که قابلیت پیش‌بینی بیماری سرطان سینه را داشته باشند.

روش: این مطالعه از نوع توصیفی-تحلیلی می‌باشد. پایگاه داده آن شامل ۶۸۳ رکورد مستقل شامل ۹ متغیر موجود در پایگاه داده یادگیری ماشین UCI می‌باشد. در این مقاله، از شبکه‌های عصبی مصنوعی پرسپترون، بیزین و شبکه عصبی LVQ برای کلاس بندی سرطان سینه به دوکلاس خوش‌خیم و بدخیم استفاده شده است. از ۸۰٪ داده‌ها جهت آموزش و از ۲۰٪ باقی‌مانده جهت آزمون استفاده شد.

نتایج: پس از پیش‌پردازش داده‌ها شبکه‌های عصبی متفاوت با معماری‌های مختلف مورد بررسی قرار گرفتند. در بهترین حالت خوش‌خیم یا بدخیم بودن سرطان را در شبکه‌های عصبی پرسپترون چند لایه و شبکه عصبی LVQ و بیزین با میانگین ده بار تست به ترتیب با دقت‌های ۹۷/۵٪ و ۹۷/۶٪ و ۹۸/۳٪ پیش‌بینی شد. بررسی‌های مطالعه نشان داد که شبکه عصبی بیزین در تشخیص بیماری موفق‌تر است.

نتیجه‌گیری: سرطان پستان یکی از شایع‌ترین سرطان‌ها در بین زنان می‌باشد. تشخیص به موقع بیماری ضمن کاهش هزینه‌ها، شانس درمان موفقیت‌آمیز بیمار را افزایش می‌دهد. در این مطالعه ضمن تشخیص بیماری به کمک روش‌های داده‌کاوی، توانست با استفاده از شبکه عصبی بیزین به دقت بالایی در تشخیص بیماری دست یابد.

کلید واژه‌ها: سرطان پستان، شبکه عصبی مصنوعی، شبکه عصبی پرسپترون، LVQ، داده کاوی

• **ارجاع:** دهقان پروانه، مقربی مائده، ذباح ایمان، لایقی کامران، ماروسی علی. مدل‌سازی بیماری سرطان پستان با استفاده از روش‌های مبتنی بر داده‌کاوی. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۶؛ ۴(۴): ۲۷۸-۲۶۶.

۱. متخصص رادیوانکولوژی، استادیار ، دانشگاه علوم پزشکی تربت حیدریه، تربت حیدریه، ایران
۲. کارشناس کامپیوتر، کمیته تحقیقات دانشجویی، دانشگاه علوم پزشکی تربت حیدریه، تربت حیدریه، ایران
۳. دانشجوی دکتری کامپیوتر، دانشکده برق و کامپیوتر، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران
۴. دکتری کامپیوتر، استادیار گروه کامپیوتر، دانشکده برق و کامپیوتر، واحد تهران شمال، دانشگاه آزاد اسلامی، تهران، ایران
۵. دکتری کامپیوتر ، استادیار، گروه برق و کامپیوتر، دانشگاه تربت حیدریه، تربت حیدریه، خراسان رضوی، ایران

*** نویسنده مسئول:** خراسان رضوی، تربت حیدریه، دانشگاه آزاد اسلامی تربت حیدریه

• **شماره تماس:** ۰۹۱۵۹۳۱۱۰۵۰ • **Email:** imanz.zabbah@iaut.ac.ir

مقدمه

سرطان پستان (Breast cancer) یک تومور بدخیم است که سلول‌های بافت سینه به علت اختلالات ژنتیکی مانند جهش، افزایش کروموزومی، حذف، بازسازمانی، جابه‌جاشدگی و تکرارشدگی کروموزومی بدون هیچ کنترلی شروع به تقسیم شدن کرده و به وجود می‌آیند [۱]. امروزه سرطان پستان، مشکل بهداشتی عمده برای زنان در سراسر جهان محسوب می‌شود [۲]. این نوع سرطان در ایالات متحده یکی از رایج‌ترین سرطان‌هایی است که زنان به آن مبتلا می‌شوند [۳]. در ایران این سرطان اولین نوع سرطان تشخیص داده شده در میان زنان است که ۲۴/۴٪ از همه انواع بدخیمی‌ها را به خود اختصاص می‌دهد [۴]، همچنین میزان بروز سرطان پستان در زنان ایرانی ۲۲ در ۱۰۰ هزار نفر و میزان شیوع آن ۱۲۰ در هر ۱۰۰ هزار نفر می‌باشد [۵]. اگرچه سرطان پستان اخیراً از جنبه‌های مختلفی موضوع تحقیقات گسترده در اغلب مراکز پژوهشی سرطان در جهان قرار گرفته، ولی با این حال تحقیقات همچنان ادامه دارد [۶]. سن مرگ‌ومیر ناشی از سرطان پستان در ایران بین ۴۰ تا ۴۹ سالگی است، درحالی که این سن در کشورهای پیشرفته بین ۵۵ تا ۶۰ است؛ بنابراین در ایران حداقل ده سال پایین‌تر از کشورهای پیشرفته است و با توجه به نقش محوری زنان در این سن در خانواده و جامعه، مرگ‌ومیر و ناتوانی ناشی از بیماری صدمات جبران‌ناپذیری به جامعه و خانواده وارد می‌کند [۷] اگر چه در ده سال گذشته شیوع این بیماری بسیار بالا بوده، ولی میزان مرگ ومیر بر اثر این بیماری کاهش یافته است [۸]. سرطان پستان بر اساس نوع خطر آن به دو دسته خوش‌خیم و بدخیم طبقه‌بندی می‌شود. تومورهای خوش‌خیم به طور غیرطبیعی رشد می‌کنند، ولی به‌ندرت باعث مرگ فرد می‌شوند در عین حال برخی از این توده‌ها نیز می‌توانند خطر ابتلا به سرطان پستان را افزایش دهند [۹]. در سال‌های اخیر استفاده از روش‌های داده‌کاوی به منظور تشخیص بیماری‌ها، مورد توجه بسیاری از محققین قرار گرفته است. در بین روش‌های مختلف، شبکه‌های عصبی مصنوعی به دلیل ماهیت ساختاری خود از محبوبیت ویژه‌ای برخوردارند. Werner و همکاران برای تشخیص سرطان پستان بر روی مجموعه داده ویسکانسین از الگوریتم ژنتیک استفاده کردند و پس از مقایسه خروجی با خروجی واقعی به دقت ۹۶/۳۲٪ رسیدند [۱۰].

سروستانی و همکاران از دو مجموعه داده ویسکانسین و بیمارستان نمازی شیراز استفاده کردند و مقایسه‌هایی بین شبکه های عصبی مصنوعی (Multi Layer Perceptron) MLP، SOM (Self Organization Map)، PNN(Probabilistic Neural Network) و RBF(Radial Basis Function Networks) برای تشخیص سرطان پستان انجام دادند و نتایج آن، حاکی از برتری شبکه‌های PNN و RBF در تشخیص این بیماری بود [۱۱].

Salama و همکاران از روش بیزین ساده و درخت تصمیم بر روی مجموعه داده بیمارستان ویسکانسین استفاده کرده و به ترتیب به دقت ۹۲/۹۷٪ و ۹۳/۱۵٪ رسیدند [۱۲]. ایران پور و همکاران از شبکه‌های RBF و Support Vector) SVM Machine) برای تشخیص سرطان پستان استفاده کردند و به دقت ۹۸/۱٪ رسیدند [۱۳].

در مطالعه‌ای دیگر نتایج پیاده‌سازی بر روی مجموعه داده بیمارستان ویسکانسین در روش بیزین، شبکه عصبی RBF و درخت تصمیم به ترتیب به دقت ۹۲/۶۱٪ ، ۹۳/۶۷٪ و ۹۲/۹۷٪ دست یافته است [۱۴].

گنجی و همکاران از الگوریتم مورچگان برای تشخیص خوش خیم یا بدخیم بودن سرطان پستان استفاده کردند و به دقت ۹۵ ٪ رسیده‌اند [۱۵].

Zhou و همکاران از ترکیب الگوریتم c4.5 و شبکه‌های عصبی استفاده کردند و توانستند دسته‌بندی بیماران مبتلا به سرطان پستان را با دقت ۹۴٪ انجام دهند [۱۶].

قیومی‌زاده برای تعیین خوش‌خیم یا بدخیم بودن سرطان پستان از ترکیب شبکه عصبی خودسازمانده (SOM) و شبکه پرسپترون چند لایه (MLP) استفاده کرده‌اند. در شبکه خودسازمانده، از روش یادگیری رقابتی برای آموزش استفاده می‌شود [۱۷].

اگرچه تحقیقات مربوط به سرطان بیشتر بالینی و بیولوژیکی است؛ اما استفاده از تحقیقات آماری به یک امر رایج تبدیل شده است. توجه بیشتر به فاکتورهای بالینی و بیولوژیکی بیماری سرطان پستان و تشخیص دقیق‌تر این بیماری حتی به اندازه ۱ ٪ باعث خواهد شد که شانس بقاء بیمار افزایش یافته و روند دوره درمان با سرعت بیشتری طی شود؛ لذا در این مطالعه که با هدف تشخیص خوش‌خیم یا بدخیم بودن بیماری سرطان پستان صورت گرفته است، با استفاده از شبکه‌های عصبی مختلف تلاش شد که دقت تشخیص بیماری را نسبت به مطالعات مشابه افزایش دهد.

روش

این مطالعه از نوع توصیفی، تحلیلی است که بر اساس متغیرهای ورودی به پیشگویی وضعیت بیماری سرطان سینه از نظر خوش

خیم یا بدخیم بودن می‌پردازد. داده‌های مورد استفاده در این پژوهش از مجموعه داده مربوط به بیماران مبتلا به سرطان پستان بیمارستان ویسکانسین، موجود در انبار داده یادگیری ماشین دانشگاه ایروین، کالیفرنیا تأمین شده است [۱۸]. بانک اطلاعاتی این منبع شامل ۶۹۹ نمونه با ۹ پارامتر می‌باشد. در ایـن مجمـوعه ۱۶ نمونه دارای ویـژگی‌های کامـل نیست.

در نهایت داده‌های این منبع را می‌توان یک ماتریس ۶۸۳*۹در نظر گرفت. از این تعداد ۶۵٪ این بیماران یعنی ۴۴۴ نفر سرطان به خوش‌خیم و ۳۵٪ یعنی ۲۳۹ نفر مبتلا به سرطان بدخیم سینه بوده اند. سطرهای این ماتریس ۹ پارامتری هستند که می‌توانند خوش‌خیم با بدخیم بودن سرطان را مشخص نمایند. جدول ۱ پارامترهای تشخیص بیماری سرطان پستان و دامنه تغییرات هر یک از آن‌ها [۲۰-۱۸] و نیز جدول ۲ آماره‌های توصیفی بیماری را نشان می‌دهد.

جدول ۱: متغیرهای ورودی در تشخیص سرطان پستان

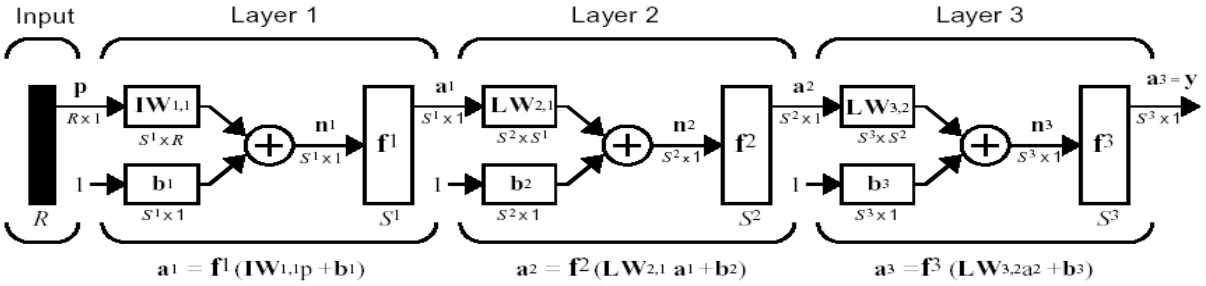
Attribute	Domain
Clump Thickness	۱-۱۰
Uniformity of Cell Size	۱-۱۰
Uniformity of Cell Shape	۱-۱۰
Marginal Adhesion	۱-۱۰
Single Epithelial Cell Size	۱-۱۰
Bare Nuclei	۱-۱۰
Bland Chromatin	۱-۱۰
Normal Nucleoli	۱-۱۰
Mitoses	۱-۱۰

جدول ۲: آماره‌های توصیفی سرطان پستان به تفکیک ویژگی

نام ویژگی	کلاس	Means±SD
ضخامت غده	خوش‌خیم	۲/۹۶ ± ۱/۶۷۳
	بدخیم	۷/۱۹ ± ۲/۴۳۸
	مجموع	۴/۴۴ ± ۲/۸۲۱
یکنواختی اندازه سلول	خوش‌خیم	۱/۳۱ ± ۰/۸۵۶
	بدخیم	۶/۵۸ ± ۲/۷۲۴
	مجموع	۳/۱۵ ± ۳/۶۵
یکنواختی شکل سلول	خوش‌خیم	۱/۴۱ ± ۰/۹۵۷
	بدخیم	۶/۵۸ ± ۲/۵۶۹
	مجموع	۳/۲۲ ± ۲/۹۸۹
چسبندگی لبه ها	خوش‌خیم	۱/۳۵ ± ۰/۹۱۷
	بدخیم	۵/۵۹ ± ۳/۱۹۷
	مجموع	۲/۸۳ ± ۲/۸۶۵
اندازه سلول مخاطی منفرد	خوش‌خیم	۲/۱۱ ± ۰/۸۷۷
	بدخیم	۵/۳۳ ± ۲/۴۴۳
	مجموع	۳/۲۳ ± ۲/۲۲۳
نوکلئی لخت	خوش‌خیم	۱/۳۵ ± ۱/۱۷۸
	بدخیم	۷/۶۳ ± ۳/۱۱۷
	مجموع	۳/۵۴ ± ۳/۶۴۴
رنگینه ملایم	خوش‌خیم	۲/۰۸ ± ۱/۰۶۲
	بدخیم	۵/۹۷ ± ۲/۲۸۲
	مجموع	۳/۴۵ ± ۲/۴۵۰
نوکلئی نرمال	خوش‌خیم	۱/۲۶ ± ۰/۹۵۵
	بدخیم	۵/۸۶ ± ۳/۳۴۹
	مجموع	۲/۸۷ ± ۳/۰۵۳
تقسیم هسته سلول	خوش‌خیم	۱/۰۷ ± ۰/۵۱۰
	بدخیم	۲/۶۰ ± ۲/۵۶۴
	مجموع	۱/۶۰ ± ۱/۷۳۳

داده کاوی

به فرایند استخراج دانش ناشناخته، درست، و بالقوه مفید از داده، داده‌کاوی گفته می‌شود. داده‌ها اغلب حجیم؛ اما بدون ارزش می باشند. دانش نهفته در داده‌ها آن‌ها را قابل استفاده می‌کند. روش های داده‌کاوی را می‌توان به صورت بدون ناظر (Unsupervised learning) و با ناظر (Supervised learning) در نظر گرفت. داده‌کاوی در پی ساختارهایی در بین متغیرها می‌باشد که روش خوشه‌بندی (Clustering) به عنوان معمول‌ترین روش شناخته می‌شود [۲۱،۲۲]. در این پژوهش از سه نوع شبکه عصبی مختلف به منظور پیش‌بینی بیماری سرطان سینه استفاده گردید. مدل‌سازی به کمک شبکه عصبی پرسپترون چند لایه (Multilayer perceptron) در مرحله اول این پژوهش ا از ساده‌ترین و کارآمدترین ساختارهای پیشنهادی برای استفاده در مدل‌سازی به نام مدل پرسپترون چندلایه (MLP) استفاده شده است. شکل ۱ ساختار شبکه عصبی مورد استفاده در این پژوهش که از نوع ۳ لایه است را نشان می‌دهد.



$$a^3 = f^3\left(LW^{3,2}\, f^2\left(LW^{2,1} f^1\left(IW^{1,1}p + b^1\right) + b^2\right) + b^3\right) = y$$

شکل ۱: معماری شبکه عصبی پرسپترون چند لایه مورد استفاده در این مطالعه

تابع خروجی شبکه در لایه آخر با فرمول ۱ محاسبه می‌شود که در آن h و o به ترتیب نشـــان‌دهنده لایه نهان و لایه خروجی بوده و منظور از w همان وزن‌های لایه‌ها می‌باشد.

$$O_i = sgm\left(\sum_m sgm\left(\sum_l x_i w_{lm}^h\right) w_{mi}^o\right) \qquad (1)$$

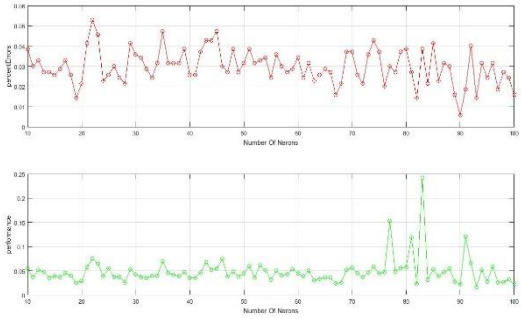
تعداد ورودی شبکه مانند مطالعات مشابه ۹ پارامتر مندرج در جدول ۲ است.

به طورکلی برای بررسی میزان موفقیت و کارایی سیستم‌های دسته‌بندی و تشخیص بیماری‌ها، از ماتریس آشفتگی (Confusion) استفاده می‌شود. تحلیل‌های این ماتریس در دسته‌بندی و تشخیص بیماران منجر به ۴ حالت مثبت حقیقی (Positive True یا PT)، منفی حقیقی (True Negative یا TN)، مثبت کاذب (False Positive یا FP) و منفی کاذب (False Negative یا FN) می‌شود. از نتایج ماتریس کانفیوژن سه شاخص حساسیت (دقت سیستم در تشخیص نوع بدخیم) و

Sgm نیز تابع سیگموئید ا ست که به صورت زیر تعریف می گردد:

$$sgm(x) = \frac{1}{1 + e^{-x}} \qquad (2)$$

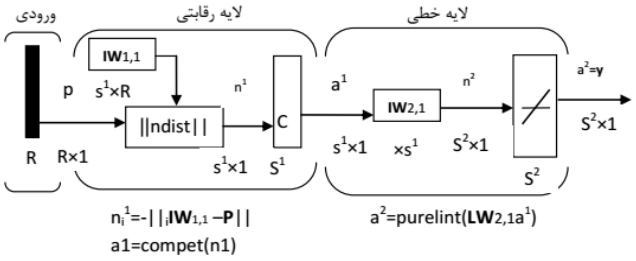
ویژگی (دقت سیستم در تشخیص نوع خوش‌خیم) و صحت (نسبت تمام مواردی که به صورت صحیح طبقه‌بندی شدند) به دست می‌آید که برای تحلیل عملکرد سیستم‌های طبقه‌بندی استفاده می‌شود. پس از بررسی حدود ۹۰ شبکه عصبی مختلف کمترین خطا ۰/۰۰۸۶و مربوط به زمانی بود که ۸۰ نرون در لایه مخفی اول وجود داشت. در تمامی مدل‌ها از ۷۰٪ داده‌ها جهت آموزش (Train) و ۱۵٪ آن‌ها جهت آزمون (Test) و ۱۵٪ باقی‌مانده جهت اعتبارسنجی (Validation) استفاده شده است. شکل۲- الف خطای مربوط به هر یک از انواع معماری شبکه و شکل۲- ب بازدهی هریک از معماری مختلف را نشان می‌دهد.



شکل ۲: الف بالا- خطای مربوط به تعداد ۹۰ معماری شبکه مختلف. ب پایین- بازدهی مربوط به هریک از معماری ها

مدل‌سازی با شبکه عصبی مصنوعی LVQ (Learning Vector Quantization)

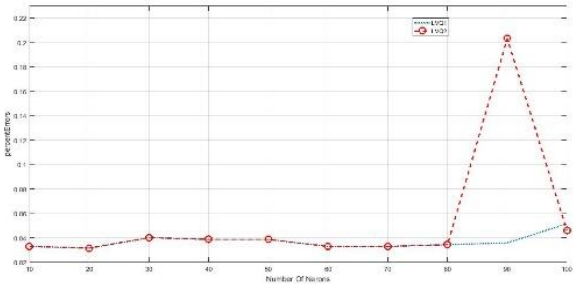
شبکه عصبی LVQ دارای دو لایه رقابتی و خطی می‌باشد. لایه رقابتی دسته‌بندی کردن بردارهای ورودی را یاد گرفته و در نهایت لایه خطی کلاس‌های لایه رقابتی را به دسته‌های هدف که توسط کاربر تعیین شده نگاشت می‌کند. شکل ۳ معماری شبکه‌های LVQ در این پژوهش را نشان می‌دهد [۲۳،۲۴]. در این شکل S^1 و S^2 به ترتیب تعداد نرون‌های لایه رقابتی و خطی و R تعداد عضوهای بردار ورودی می‌باشند. هردو لایه رقابتی و خطی دارای یک نرون به ازای هر کلاس هستند [۲۵].



شکل ۳: معماری شبکه عصبی LVQ مورد استفاده در این مطالعه

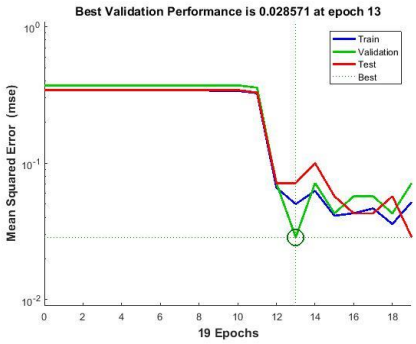
در این مطالعه از ۱۰ شبکه عصبی LVQ با معماری‌های متفاوت استفاده شد تا بتوان بهترین معماری شبکه جهت تشخیص بیماری شناسایی شود. در تمامی شبکه‌ها از ۸۰٪ داده جهت آموزش شبکه و ۱۰٪ آن‌ها جهت اعتبارسنجی و ۱۰٪ به منظور آزمون استفاده گردید. تعداد تکرارها (Epochs) برای تمامی آن‌ها ۳۰ و از دو نوع الگوریتم یادگیری LVQ1 و هم LVQ2 استفاده شده است. نمودار ۴ وضعیت خطای شبکه‌های عصبی مختلف در ازای تغییر تعداد نرون‌های لایه رقابتی را نشان می‌دهد. در این نمودار محور افقی بیانگر تعداد نرون‌ها در لایه رقابت است که عددی بین ۱۰ تا ۱۰۰ در نظر گرفته شده است و محور عمودی نمودار، خطا را نشان می‌دهد.

نمودار حاکی از آن است که جزء در مواردی خاص رفتار کلی دو الگوریتم در تشخیص سرطان خوش‌خیم از بدخیم شبیه به هم است. فقط زمانی که تعداد نرون‌ها در لایه رقابتی ۹۰ انتخاب شده است رفتار دو الگوریتم متفاوت بوده و LVQ1 عملکرد بهتری داشته است. همچنین نتایج نشان می‌دهد که الزاماً با افزایش تعداد نرون‌های لایه رقابتی و ضعیت شبکه بهتر نشده است. به عنوان مثال زمانی که تعداد نرون‌های لایه یادگیری ۳۰ نرون می‌باشد خطای شبکه حدود ۰/۰۴ و زمانی که تعداد نرون ها ۱۰۰ بوده است خطای شبکه بیشتر از ۰/۰۵ بوده است. در بهترین وضعیت کمترین خطا با مقدار ۰/۰۲۸۵۷۱ به دست آمد.



شکل ٤: نمودار خطا مربوط به تعداد نرون‌های لایه رقابتی با ۲ الگوریتم LVQ1 و LVQ2

شکل ۵ فرایند یادگیری بهترین شبکه عصبی LVQ از بین شبکه‌های عصبی فوق را نشان می‌دهد.



شکل ۵: فرایند یادگیری بهترین شبکه عصبی LVQ در تشخیص سرطان خوش‌خیم از بدخیم

مدل‌سازی به وسیله شبکه عصبی بیزین (Bayesian network)

این روش یکی از روش‌های سیستم پشتیبان تصمیم‌گیری می‌باشد که ابزار قدرت‌مندی در مدل کردن روابط علی و معلولی در قالب شبکه‌ای از احتمالات است. نکته بسیار مهم در مورد روش شبکه عصبی بیزین این است که این روش به اطلاعات دقیق و تاریخچه کامل یک واقعیت نیاز ندارد، بلکه می‌تواند با استفاده از اطلاعات ناقص و غیردقیق نیز به نتایج قانع‌کننده‌ای درزمینه تخمین وضعیت فعلی یا آینده یک سیستم دست یابد. بر این اساس سومین نوع مدل‌سازی با استفاده از شبکه‌های عصبی بیزین انجام شده است. برای توزیع وزن‌ها از رابطه ۳ استفاده شد.

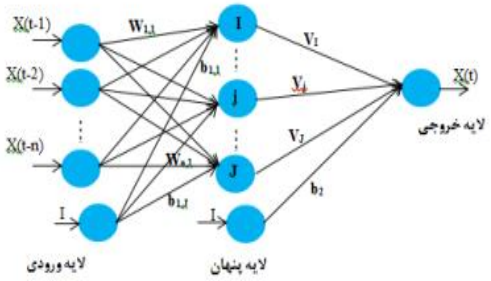
$$y_k = f_{outer}\left(\sum_{j=1}^{m} w_{kj}^{(2)} f_{inner}\left(\sum_{i=1}^{d} w_{ji}^{(1)} + w_{j0}^{(1)}\right) + w_{k0}^{(1)}\right) \quad (3)$$

$W_{ji}^{(1)}$ و $W_{ki}^{(2)}$ وزن‌ها در لایه اول و لایه دوم به ترتیب با ورودی I و لایه مخفی j می‌باشد و $w_{j0}^{(1)}$ بایاس برای واحد مخفی j است. M تعداد واحدهای مخفی، d تعداد واحد ورودی و K شاخصی برای واحد خروجی است. تابع $f_{outer}(0)$ خطی است و $f_{inner}(0)$ تابع تانژانت هایپربولیک می‌باشد. درروش بیزین برای توزیع وزن‌ها از معادله (۳) استفاده می‌شود و میزان وزن‌های شبکه با قرار دادن در رابطه ۴ به دست می‌آیند [۲۴].

$$P(W|D) = \frac{P(D|W)P(w)}{P(D)} \quad (4)$$

P(w) تابع توزیع احتمال در فضای وزنی با فقدان داده است که به صورت تابع اولیه می‌باشد. کمیت P(W|D) تابع احتمالی وزن‌ها است که به عنوان تابع توزیع احتمالی بعد از آموزش داده ها دیده می‌شود. P(D|W) تابع توزیع احتمال و P(D) تابع توزیع احتمال ثانویه است.

ساختار شبکه ع صبی بیزین که در این پژوهش مورد ا ستفاده قرار گرفته است در شکل ۶ نشان داده شد.



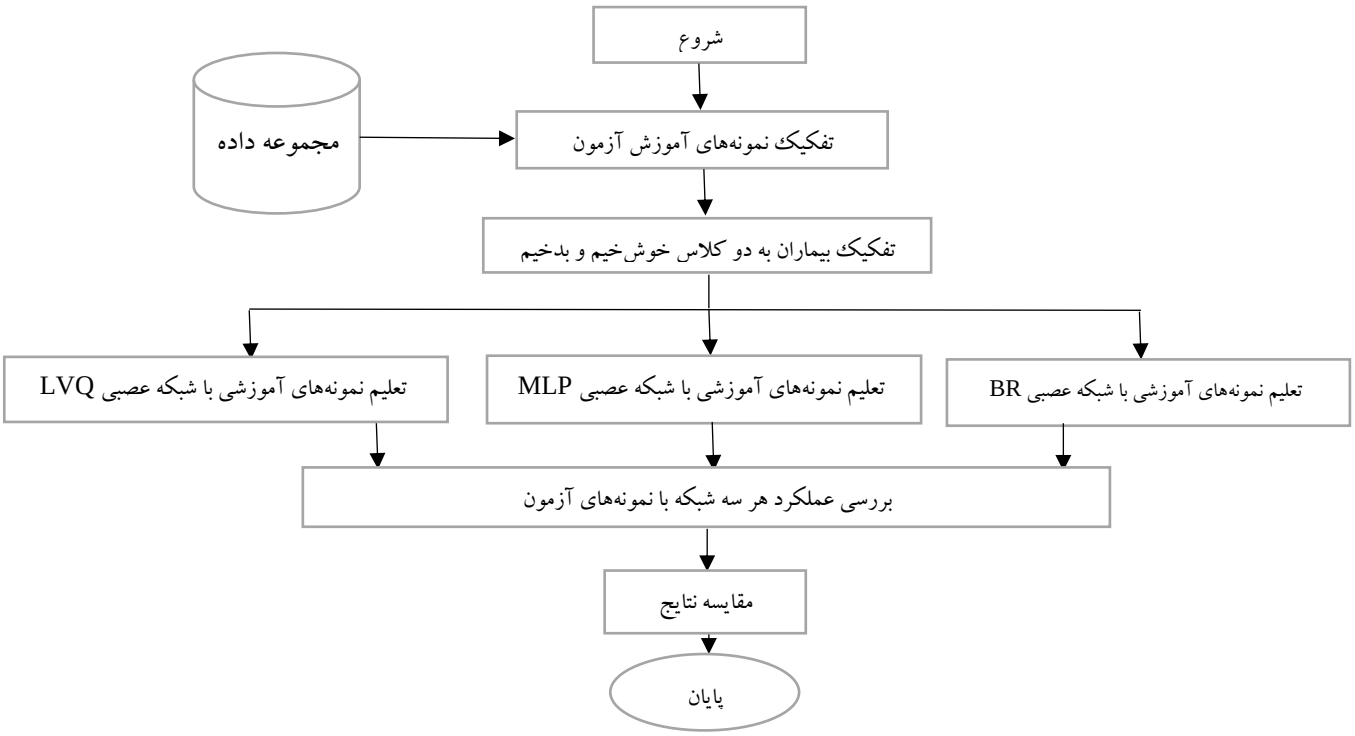
شکل ٦: معماری شبکه عصبی بیزین مورد استفاده در این پژوهش

جهت یافتن بهترین معماری از نظر تعداد نرون در لایه‌های رقابت ساختارهای متفاوتی مورد بررسی قرار گرفته است که جدول ۳ برخی از انواع معماری‌های مورد مطالعه را نشان می دهد. همان‌طور که جدول ۳ نشان می‌دهد ۴ شبکه اول جدول دو لایه مخفی دارند و با تغییر تعداد نرون‌ها تفاوت محسوسی در کاهش خطا رخ نمی‌دهد حتی زمانی که تعداد لایه‌ها از ۲ به ۳ تغییر کرده‌اند یعنی در دو ردیف آخر جدول، تنها ۰/۲٪ بهبود حاصل شده است.

جدول ۳: بررسی معماری‌های مختلف شبکه عصبی با در نظر گرفتن کلیه پارامترهای ورودی بیماری سرطان پستان

معماری شبکه عصبی	تعداد لایه مخفی	دقت تشخیص(درصد) روی مجموعه آزمون
۵-۵-۱	۲	٪۹۸
۱۰-۵-۱	۲	٪۹۸/۱
۱۵-۵-۱	۲	٪۹۸/۳
۱۵-۲۰-۱	۲	٪۹۸/۱
۱۰-۵-۵-۱	۳	٪۹۸/۲
۲۰-۱۰-۵-۱	۳	٪۹۸/۳

الگوریتم مورد استفاده در این پژوهش در شکل ۷ ارائه شده است. همانطور که الگوریتم نشان می‌دهد در این پژوهش پس از تفکیک نمونه های تست و آزمون و نیز تقسیم آن‌ها به دو کلاس خوش خیم و بد خیم به کمک ۳ نوع شبکه عصبی مختلف به مدل‌سازی سرطان پرداخته و نتایج با یکدیگر مقایسه شده است.



شکل ۷: الگوریتم مورد استفاده در این پژوهش

نتایج

در این مطالعه ۶۸۳ بیمار مبتلا به سرطان پستان که اطلاعات آن از بیمارستان ویسکانسین موجود در مخزن داده یادگیری ماشین UCI جمع آوری شده است مورد بررسی قرار گرفتند. مجموع متغیرهای بالینی بیماران شامل ۹ ریسک فاکتور بود. از این تعداد بیمار، ۸۰٪ آن‌ها (۵۴۷ نمونه) جهت آموزش شبکه و از ۱۰٪ آن‌ها جهت اعتبارسنجی و از باقی‌مانده داده‌ها جهت تست شبکه استفاده شده است. با توجه به اینکه بازه تغییرات هر یک از این ۹ ریسک فاکتور بیماری بین ۱ تا ۱۰ است توزیع فراوانی هر یک از آن‌ها در جدول ۴ نمایش داده شد.

جدول ٤: توزیع فراوانی ۹ ریسک فاکتور بیمار سرطان پستان

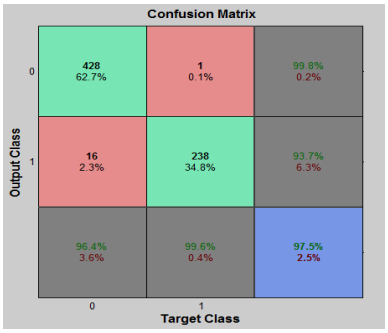
	چسبندگی لبه‌ها		نوکلئی نرمال		رنگینه ملایم		نوکلئی لخت		اندازه سلول مخاطی منفرد		تقسیم هسته سلول		یکنواختی اندازه سلول		ضخامت غده		یکنواختی شکل سلول	
	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی	درصد	فراوانی
۱	۸۰/۵	۵۶۳	۶۱/۸	۴۳۲	۲۱/۵	۱۵۰	۵۷/۵	۴۰۲	۶/۳	۴۴	۵۶/۲	۳۹۳	۵۳/۴	۳۷۳	۱۹/۹	۱۳۹	۴۹/۵	۳۴۶
۲	۵/۰	۳۵	۵/۲	۳۶	۲۲/۹	۱۶۰	۴/۳	۳۰	۵۸/۳	۳۷۶	۸/۳	۵۸	۶/۴	۴۵	۷/۲	۵۰	۸/۳	۵۸
۳	۴/۷	۳۳	۶/۰	۴۲	۲۳/۰	۱۶۱	۴/۰	۲۸	۱۰/۲	۷۱	۸/۳	۵۸	۷/۴	۵۲	۱۴/۹	۱۰۴	۷/۶	۵۳
۴	۱/۷	۱۲	۲/۶	۱۸	۵/۶	۳۹	۲/۷	۱۹	۶/۹	۴۸	۴/۷	۳۳	۵/۴	۳۸	۱۱/۳	۷۹	۶/۲	۴۳
۵	۰/۹	۶	۲/۷	۱۹	۴/۹	۳۴	۴/۳	۳۰	۵/۶	۳۹	۳/۳	۲۳	۴/۳	۳۰	۱۸/۳	۱۲۸	۴/۶	۳۲
۶	۰/۴	۳	۳/۱	۲۲	۱/۳	۹	۰/۶	۴	۵/۷	۴۰	۳/۰	۲۱	۳/۶	۲۵	۴/۷	۳۳	۴/۱	۲۹
۷	۱/۳	۹	۲/۳	۱۶	۱۰/۲	۷۱	۱/۱	۸	۱/۶	۱۱	۱/۹	۱۳	۲/۷	۱۹	۳/۳	۲۳	۴/۳	۳۰
۸	۱/۱	۸	۳/۳	۲۳	۴/۰	۲۸	۳/۰	۲۱	۰/۳	۲۱	۳/۶	۲۵	۴/۰	۲۸	۶/۳	۴۴	۳/۹	۲۷
۹	۲/۰	۱۴	۲/۱	۱۵	۱/۶	۱۱	۱/۳	۹	۰/۳	۲	۰/۶	۴	۰/۹	۶	۲/۰	۱۴	۱/۰	۷
۱۰	۰/۰	۰	۸/۶	۶۰	۲/۹	۲۰	۱۸/۹	۱۳۲	۴/۴	۳۱	۷/۹	۵۵	۹/۶	۶۷	۹/۹	۶۹	۸/۳	۵۸
مجموع	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳	۹۷/۷	۶۸۳

در مطالعات مختلف تحت عنوان داده‌کاوی پزشکی، راهکارهای متعددی جهت کشف روابط بین عوامل سرطان پستان ارائه شده است. استفاده از شبکه‌های عصبی مصنوعی جهت تشخیص نوع سرطان در مطالعات مختلف مورد بررسی قرار گرفته است. در این پژوهش سعی شد برای بهبود تشخیص بیماری از روش‌های مختلف داده‌کاوی با تکیه بر انواع شبکه‌های عصبی مصنوعی استفاده شود.

در اولین مرتبه مدل‌سازی از یک شبکه عصبی پرسپترون چند لایه با الگوریتم پس انتشار خطا استفاده شد. از این شبکه عصبی مصنوعی به عنوان یک طبقه‌بند برای کلاس‌بندی داده‌ها به دو دسته خوش‌خیم و بدخیم استفاده گردید و کلیه ۹ فاکتور مربوط به بیماری سرطان پستان به شبکه عصبی اعمال شد و آموزش شبکه با ۷۰٪ داده‌ها و تست آن با ۳۰٪ از داده‌ها انجام شد. همچنین معماری‌های مختلف شبکه عصبی برای حصول بهترین کلاس‌بندی مورد بررسی قرار گرفت. جدول ۵ برخی از بهترین معماری‌های شبکه عصبی پرسپترون چند لایه و شکل ۷ ماتریس آشفتگی مربوط به بهترین معماری از بین حدود ۱۰۰ معماری مختلف را نشان می‌دهد.

جدول ۵: تست معماری‌های مختلف شبکه MLP و درصد صحت تشخیص با میانگین ۱۰ بار تست

صحت عملکرد	معماری شبکه	تعداد لایه مخفی
٪۹۷/۵	۱۰-۱۰-۱	۲
٪۸۷/۶	۲۰-۱۰-۱	۲
٪۹۷/۱	۱۰-۱۰-۱۰-۱	۳



شکل ۸: ماتریس آشفتگی بهترین معماری شبکه پرسپترون

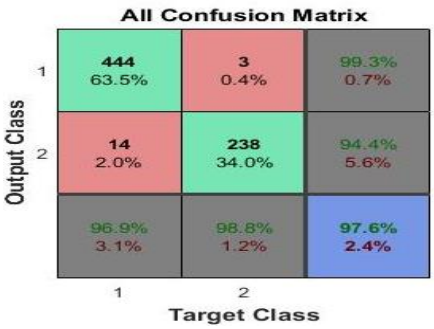
نتایج نشان می‌دهد که این شبکه عصبی توانسته است تعداد ۴۲۸ بیمار یعنی معادل ۹۶/۴٪ که مبتلا به سرطان خوش‌خیم بوده‌اند را به درستی دسته‌بندی نماید و تعداد ۱۶ نفر از این افراد اشتباهاً در کلاس سرطانی‌های بدخیم قرار گرفته است. همچنین تعداد ۲۳۸ نفر از مجموع ۲۳۹ نفر بیماری که دچار سرطان بدخیم بوده اند به درستی در کلاس بیماران سرطان بدخیم قرار گرفته‌اند؛ اما فقط یک نفر از آن‌ها یعنی معادل ۰/۱٪ از این افراد اشتباهاً در کلاس بیماران خوش‌خیم دسته بندی شده‌اند.

در مرتبه دوم مدل‌سازی، از شبکه عصبی LVQ استفاده گردید. علاوه بر بررسی شبکه‌های عصبی با معماری‌های مختلف تعداد ورودی‌های بیماری در ورودی شبکه عصبی نیز مورد بررسی قرار گرفت که نتایج حاکی از آن است که وقتی تعداد ریسک فاکتورهای تشخیص بیماری افزایش پیدا می‌کند میزان خطای شبکه عصبی کاهش می‌یابد به طوری که وقتی فقط از ویژگی اول جهت کلاس‌بندی بیماران به دو دسته خوش‌خیم و بدخیم استفاده می‌شد خطای محاسبه شده برای بهترین معماری ۰/۱۳۸۸ خواهد بود و زمانی که از ۹ ویژگی استفاده می‌شود خطای کلاس‌بندی به ۰/۰۲۴۳ کاهش می‌یابد. جدول ۶ تأثیر افزایش پارامترهای ورودی روی بیماری را نشان می‌دهد.

جدول ٦: میانگین ۱۰ مرتبه مدل‌سازی بر روی شبکه عصبی LVQ به منظور پیش‌بینی بیماری سرطان سینه

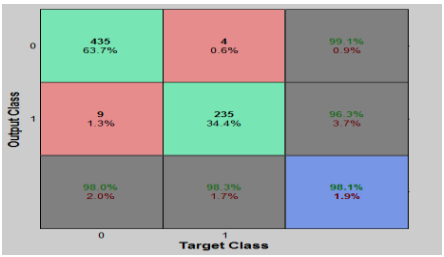
شماره مدل	ریسک فاکتورها	درصد خطا
۱	ضخامت غده	۰
۲	مجموع ریسک فاکتورهای مدل ۱ + یکنواختی اندازه سلول	۰/۱۳۸۸
۳	مجموع ریسک فاکتورهای مدل ۲ + یکنواختی شکل سلول	۰/۰۵۴۴
۴	مجموع ریسک فاکتورهای مدل ۳ + چسبندگی لبه ها	۰/۰۴۱۵
۵	مجموع ریسک فاکتورهای مدل ۴ + اندازه سلول مخاطی منفرد	۰/۰۵۴۴
۶	مجموع ریسک فاکتورهای مدل ۵ + نو کلپی لخت	۰/۰۵۰۱
۷	مجموع ریسک فاکتورهای مدل ۶ + رنگینه ملایم	۰/۰۳۲۹
۸	مجموع ریسک فاکتورهای مدل ۷ + نوکلپی نرمال	۰/۰۲۷۲
۹	مجموع ریسک فاکتورهای مدل ۸ + تقسیم هسته سلول	۰/۰۲۴۳

همچنین در این مرحله نشان داده شد که در مورد تشخیص سرطان پستان هر دو الگوریتم مورد استفاده در شبکه‌های عصبی LVQ به جزء در موارد بسیار اندک از نظر خطا رفتار مشابه هم دارند. در نهایت از بین ۱۰۰ معماری مختلف ماتریس آشفتگی مربوط به بهترین معماری محاسبه گردید که در شکل ۸ نشان داده شد.



شکل ۹: ماتریس آشفتگی بهترین معماری شبکه LVQ

در مرحله سوم مدل‌سازی از شبکه عصبی بیزین استفاده گردید که در نهایت بهترین معماری به دست آمده در شکل ۹ نمایش داده شد.



شکل ۱۰: ماتریس آشفتگی بهترین معماری شبکه بیزین

مشاهده شد که در شبکه عصبی بیزین پارامتر حساسیت ۶۳/۷٪ و پارامتر اختصاصیت ۱/۹٪، در شبکه عصبی MLP پارامتر حساسیت ۶۲/۷٪ و پارامتر اختصاصیت ۲/۵٪ و در شبکه عصبی LVQ پارامتر حساسیت ۶۳/۵٪ و پارامتر اختصاصیت ۲/۵٪ به دست آمد. جدول ۷ مقایسه بین ۳ نوع شبکه مورد استفاده در این پژوهش را نشان می‌دهد.

جدول ۷: مقایسه صحت عملکرد شبکه‌های عصبی مختلف پس از میانگین ۱۰ بار تست

نوع شبکه	صحت عملکرد
MLP	٪۹۷/۵
LVQ	٪۹۷/۶
بیزین	٪۹۸/۳

نتایج حاکی از آن است که هر سه نوع شبکه عصبی قابلیت پیش بینی بیماری سرطان پستان با قابلیت بالا را دارند و در بین این سه نوع، شبکه بیزین توانایی تقریب‌زنی بالاتری دارد.

بحث و نتیجه‌گیری

پژوهش حاضر با هدف تشخیص سرطان خوش‌خیم و بدخیم پستان با استفاده از چند تکنیک داده‌کاوی و عمدتاً بر پایه شبکه های عصبی مصنوعی انجام شد. از آن‌جایی که شبکه‌های عصبی مصنوعی به عنوان روش نوین در تشخیص بیماری‌ها مورد توجه بسیاری از محققین در سال‌های اخیر قرار گرفته است؛ لذا در این پژوهش از سه شبکه عصبی MLP، LVQ و BR برای طبقه بندی نوع سرطان پستان به دو دسته خوش‌خیم و بدخیم استفاده نمود. ابتدا شبکه‌ها با نمونه‌های آموزشی، تعلیم داده شدند و در نهایت با نمونه‌های آزمون، مورد سنجش قرار گرفتند. این بررسی به وضوح نشان از اثربخشی فن‌آوری‌های شبکه‌های عصبی در تشخیص سرطان را دارد. بسیاری از شبکه‌های عصبی در طبقه بندی با دقت سلول‌های تومور نتیجه فوق‌العاده‌ای نمایش می دهند؛ بنابراین استفاده از شبکه عصبی مصنوعی می‌تواند در کنار سایر روش‌های تشخیصی غیر تهاجمی که معمولاً مورد استفاده قرار می‌گیرند (مانند ماموگرافی و رادیوگرافی)، به عنوان یک سیستم پشتیبان تشخیص با حساسیت و ویژگی بالا، به منظور شناسایی تومورهای خوش‌خیم و بدخیم پستان مورد استفاده قرار گیرد. این نتیجه از آن جهت حائز اهمیت می‌باشد که از عوارض و آسیب‌های احتمالی روش‌های تهاجمی (نمونه برداری و عمل جراحی) برای بیمارانی که نیازی به آن‌ها ندارند، جلوگیری می‌شود [۱۷]. پیش‌بینی و تشخیص سریع تر و دقیق‌تر سرطان پستان علاوه بر کاهش هزینه‌های درمان شانس درمان را نیز افزایش می‌دهد و با توجه به اینکه سن ابتلا به سرطان در ایران ۵ تا ۱۰ سال پایین‌تر از میانگین جهانی است که ناشی از تشخیص دیرهنگام این بیماری می‌باشد؛ لذا توجه به روش‌های مبتنی بر هوش مصنوعی و داده کاوی می‌تواند به تشخیص دقیق‌تر این بیماری کمک کند. Werner و همکاران با استفاده از الگوریتم ژنتیک توانستند به دقت ٪۹۶/۳۲ برسند [۱۰]. Salama و همکاران نیز در سال ۲۰۱۲ با استفاده از درخت تصمیم و روش بیزین به ترتیب به دقت ٪۹۳/۱۵ و ٪۹۲/۹۷ دست یافتند [۱۲]. ایران‌پور و همکاران با استفاده از روش SVM و RBF به دقت ٪۹۸/۱ دست یافتند [۱۳]. Aruna و همکاران با استفاده از روش‌های RBF و درخت تصمیم و روش بیزین توانستند به ترتیب به دقت ٪۹۳/۶۷ و ٪۹۲/۹۷ و ٪۹۲/۹۱ دست یابند [۱۴]. پژوهش حاضر نیز توانسته است سرطان پستان بر روی مجموعه داده‌های بیمارستان ویسکانسین را با دقت ٪۹۸/۳ تشخیص دهد.

تشکر و قدردانی

این پژوهش با استفاده از اعتبارات پژوهشی دانشگاه علوم پزشکی تربت‌حیدریه انجام گردید.

References

1. Devita VT, Lawrence T, Rosenberg SA. Cancer: Principles & Practice of Oncology: Annual Advances in Oncology. 9th ed. Philadelphia: Lippincott Williams & Wilkins; 2012.

2. Ashkhaneh Y, Mollazadeh J, Aflakseir A, Goudarzi MA. Study of difficulty in emotion regulation as a predictor of incidence and severity of nausea and vomiting in breast cancer patients. Journal of Fundamentals of Mental Health 2015;17(3):123-8.

3. Harrington S, Michener LA, Kendig T, Miale S, George SZ. Patient-reported upper extremity outcome measures used in breast cancer survivors: a systematic review. Arch Phys Med Rehabil 2014;95(1):153-62.

4. Sharifian A, Pourhoseingholi MA, Emadedin M, Rostami Nejad M, Ashtari S, Hajizadeh N, et al. Burden of Breast Cancer in Iranian Women is Increasing. Asian Pac J Cancer Prev 2015;16(12):5049-52.

5. Hatefnia E, Niknami S, Mahmudi M, Lamyian M. The effects of "theory of planned behavior" based education on the promotion of mammography performance in employed women. J Birjand Univ Med Sci 2010;17(1):50-8. Persian

6. Matsumoto A, Jinno H, Ando T, Fujii T, Nakamura T, Saito J, et al. Biological markers of invasive breast cancer. Japanese Journal of Clinical Oncology 2015;46(2):99-105.
7. Akbari ME, Khayamzadeh M, Khoshnevis S, Nafisi N, Akbari A. Five and ten years survival in breast cancer patients mastectomies vs. breast conserving surgeries personal experience. Iranian Journal of Cancer Prevention 2008;1(2):53-6.
8. Abaspur Kazerouni I, Haddadnia J. A Novel Smart System for Mammographic Image Classification Based on Breast Density. Iranian Journal of Breast Diseases 2013;6(1):15-22. Persian
9. West D, Mangiameli P, Rampal R, West V. Ensemble strategies for a medical diagnostic decision support system: A breast cancer diagnosis application. European Journal of Operational Research 2005;162(2):532-51.
10. Werner JC, Fogarty TC, editors. Genetic programming applied to severe diseases diagnosis. Proceedings Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2001); 2001.
11. Sarvestani AS, Safavi AA, Parandeh NM, Salehi M. Predicting breast cancer survivability using data mining techniques. 2nd International Conference on Software Technology and Engineering; 2010 Oct 3-5; San Juan, PR, USA, USA: IEEE; 2010. p. 2–227.
12. Salama GI, Abdelhalim M, Zeid MA-e. Breast cancer diagnosis on three different datasets using multi-classifiers. Breast Cancer (WDBC). 2012;32(569):2.
13. Iranpour M, Almassi S, Analoui M, editors. Breast cancer detection from fna using svm and rbf classifier. First Joint Congress on Fuzzy and Intelligent Systems; 2007 Aug 29-31; Mashhad: Ferdowsi University of Mashhad, Iran
14. Aruna S, Rajagopalan S, Nandakishore L. Knowledge based analysis of various statistical tools in detecting breast cancer. Computer Science & Information Technology 2011;2:37-45.
15. Ganji MF, Abadeh MS. Parallel fuzzy rule learning using an ACO-based algorithm for medical data mining. Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA); IEEE; 2010 Sep 23-26; p. 573–81.
16. Zhou ZH, Jiang Y. Medical diagnosis with C4. 5 rule preceded by artificial neural network ensemble. IEEE Transactions on information Technology in Biomedicine 2003;7(1):37-42.
17. Ghiomizadeh H. Clustering and Diagnosis of breast cancer via thermal images using a combination of SVM and SOM neural network. Iranian Journal of Breast Diseases 2013;5(4):13–22. Persian
18. UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set [cited 2017 May 4]. Available from: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
19. American Cancer Society Information and Resources about for Cancer: Breast, Colon, Lung, Prostate, Skin. [cited 2017 May 4]. Available from: https://www.cancer.org/
20. Ghorbaninejad A. proposing novel activation functions for complex-valued neural networks and their applications on real-valued classification problems Electronic Industries 2012 ;3 (9): 61- 80. Persian
21. Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques, (The Morgan Kaufmann Series in Data Management Systems). 3th ed. USA: Morgan Kaufmann; 2011.
22. Mitra S, Hayashi Y. Neuro-fuzzy rule generation: survey in soft computing framework. IEEE Trans Neural Netw 2000;11(3):748-68.
23. Marwala T. Bayesian training of neural networks using genetic programming. Pattern Recognition Letters 2007;28(12):1452-8.
24. Daaraaee M, Vahidi J, Alipour A. A method based on an evolutionary algorithm to achieve an efficient artificial neural network model for prediction of breast tumors status. J Mazandaran Univ Med Sci 2015;25(130):100-15. Persian

Original Article

ournal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 4(4): 266-278

Modeling Breast Cancer Using Data Mining Methods

Dehghan Parvaneh[1], Mogharabi Maedeh[2], Zabbah Iman[3]*, Layeghi Kamran[4], Maroosi Ali[5]

• **Received:** 14 Jul, 2017 • **Accepted:** 26 Nov, 2017

Introduction: Breast cancer is the most common form of cancer in women. Breast cancer detection is considered as one of the most important issues in medical science. Diagnosis of benign or malignant type of cancer reduces costs and also is important in deciding about the treatment strategy. The aim of this study was to provide data mining based models that have the predictability of breast cancer detection.

Methods: This study was descriptive-analytic. Its database included 683 independent records containing nine clinical variables in the UCI machine learning. Multilayer Perceptron artificial neural network, Bayesian Neural Network and LVQ neural network were used for classification of breast cancer to benign and malignant types. In this study, 80% of data were used for network training and 20% were used for testing.

Results: After pre-processing the data, different neural networks with different architectures were used to detect breast cancer. In the best condition, we could predict benign or malignant cancer in the MLP neural networks, LVQ and Bayesian Neural Networks with an average of ten tests with an accuracy of 97.5% and 97.6% and 98.3% respectively. Our investigations showed that Bayesian neural network had a better performance.

Conclusion: Breast cancer is one of the most common cancers among women. Early diagnosis of disease reduces healthcare costs and increases patient survival chance. In this study, using data mining techniques in diagnosis, the researchers were able to use Bayesian neural network to achieve high accuracy in diagnosis.

Keywords: Breast Cancer, Neural Networks, LVQ, Data Mining

• **Citation:** Dehghan P, Mogharabi M, Zabbah I, Layeghi K, Maroosi A. Modeling Breast Cancer Using Data Mining Methods. Journal of Health and Biomedical Informatics 2018; 4(4): 266-278.

1 Radiologist, Assistant Professor, Torbat Heydarieh University of Medical Science, Torbat Heydarieh, Iran.
2. B.Sc. Computer, Student Research Committee, Torbat Heydarieh University of Medical Sciences, Torbat Heydarieh, Iran
3. PhD Student, School of Electrical and Computer, Tehran North Branch, Islamic Azad University, Tehran, Iran.
4. Computer Ph.D., Assistant Professor, Department of Computer, North Tehran Branch, Islamic Azad University, Tehran, Iran.
5. Computer Ph.D., Electrical and Computer Engineering Dept., University of Torbat Heydarieh,Iran.

***Correspondence**: Khorasan Razavi, Torbat Heydarieh, Islamic Azad University of Torbat Heydarieh, Computer Dept.
• **Tel:** 09159311050 • **Email:** imanzabbah@iautorbat.ac.ir