پیش‌بینی ابتلا به لنف ادم با ترکیب الگوریتم‌های منتخب داده‌کاوی

آنارام یعقوبی نوتاش[1]، پیمان بیات[1*]، شهپر حقیقت[2]، علی یعقوبی نوتاش[1و3]

۱-گروه مهندسی کامپیوتر و نرم‌افزار، دانشکده فنی و مهندسی، واحد رشت، دانشگاه آزاد اسلامی، رشت، ایران.

۲- مرکز تحقیقات سرطان پستان جهاد دانشگاهی، پژوهشکده سرطان معتمد، تهران، ایران.

۳- گروه جراحی، بیمارستان سینا، دانشکده پزشکی، دانشگاه علوم پزشکی تهران، تهران، ایران.

[*] نویسنده مسئول: رشت، دانشگاه آزاد اسلامی، واحد رشت، دانشکده مهندسی کامپیوتر و نرم‌افزار، گروه مهندسی کامپیوتر و نرم‌افزار.

تلفن: ۰۱۳-۳۳۴۲۳۳۰۸

E-mail: bayat@iaurasht.ac.ir

چکیده

دریافت: ۱۴۰۰/۰۷/۲۵ ویرایش: ۱۴۰۰/۰۸/۰۲ پذیرش: ۱۴۰۰/۱۰/۲۴ آنلاین: ۱۴۰۰/۱۱/۰۱

زمینه و هدف: سرطان پستان به‌عنوان دومین عامل مرگ ناشی از سرطان در زنان است. با توجه به اهمیت پیش‌بینی این عارضه، استفاده از روش‌های داده‌کاوی می‌تواند کمک بزرگی در جلوگیری از بروز عوارض لنف ادم در بیماران باشد. هدف از این تحقیق تشخیص ابتلا به لنف ادم می‌باشد.

روش بررسی: در مطالعه کاربردی توصیفی-تحلیلی به‌صورت گذشته‌نگر حاضر، عوامل مرتبط با لنف ادم در۱۱۱۷ بیمار مبتلا به سرطان پستان بررسی و احتمال ابتلا به لنف ادم، با به‌کارگیری الگوریتم‌های یادگیری ماشین پیش‌بینی شد. به طوری‌که پس از جمع‌آوری داده‌ای (فروردین ۱۳۸۸ تا خرداد ۱۳۹۷)، احتمال ابتلا به لنف ادم برای بیمار جدید بررسی و عوامل موثر بر بیماری استخراج شد. بدون احتساب زمان جمع‌آوری داده‌های آماری، مطالعه از شهریور ماه سال ۱۳۹۸ تا اسفند ماه سال ۱۳۹۹ در مرکز توانبخشی سید خندان ادامه داشت.

یافته‌ها: نتایج الگوریتم‌ها، در روش وزن‌دهی یادگیری جمعی دارای صحت ۸۷٪ و در روش یادگیری جمعی با استخراج ویژگی‌ها دارای صحت ۹۰٪ ارزیابی شد و نهایتا براساس ارزیابی نهایی تاثیرگذارترین عوامل خطر لنف ادم استخراج شدند.

نتیجه‌گیری: یکی از مهم‌ترین عوارض در سرطان پستان، لنف ادم در اندام‌های فوقانی است، که می‌تواند کیفیت زندگی بیماران را تحت تاثیر قراردهد. وجود روشی که بتواند با دقت بالا به پزشک متخصص پیشنهاد بدهد که آیا بیمار جدید در آینده، مبتلا به لنف ادم می‌شود یا خیر و یا با چه احتمالی مبتلا می‌شود، ضروری است.

کلمات کلیدی: لنف ادم سرطان پستان، کلاس‌بندی، داده‌کاوی.

مقدمه

سرطان پستان شایع‌ترین نوع بدخیمی و مهمترین عامل مرگ ناشی از سرطان در سراسر دنیاست و ۳۲٪ موارد سرطان زنان را تشکیل می‌دهد، به طوری‌که از هر نه زن یک نفر دچار این سرطان می‌شود.[1] در مطالعاتی دیگر سرطان پستان را شایع‌ترین سرطان زنان می‌داند و دومین سرطان در میان دیگر انواع سرطان‌ها تخمین زده‌اند و در هر سال تعداد موارد سرطان پستان افزایش می‌یابد.[2]

تقریبا دو نفر از پنج نفر در سطح جهان، سرطان پستان را تجربه می‌کنند و پیش‌بینی می‌شود که تعداد بیماران مبتلا تا سال ۲۰۳۰ به ۲/۱ میلیون نفر افزایش یابد.[3و4] بنابراین تشخیص زود هنگام و افزایش نرخ بقا مورد نیاز است.[5] درمان‌های متفاوتی نظیر ماستکتومی، اشعه درمانی و شیمی درمانی برای درمان این سرطان به‌کار می‌روند این درمان‌ها و عوامل خطر دیگر نظیر چاقی، پرفشاری خون و کاهش فعالیت، سبب اشکال در آناتومی و در نتیجه عملکرد سیستم لنفاوی شده و فرد را در معرض خطر ابتلا به لنف ادم (تورم دست و پا) قرار می‌دهد.[6و7]

هیچ‌یک از درمان‌ها و روش‌های دارویی و غیر دارویی موجود باعث درمان قطعی لنف ادم نمی‌گردند و فقط شدت آن را کاهش می‌دهند. به همین علت امروزه مطالعه عوامل خطر ایجاد لنف ادم بیش‌تر مورد بحث و بررسی قرار گرفته‌اند تا بتوان با شناخت آن‌ها و انجام مداخله موثر و به‌موقع از ایجاد این عارضه جلوگیری کرد.[۸] لنف ادم با تجمع غیر طبیعی لنفاوی در فضاهای بینابینی مشخص می‌شود و منجر به تورم مداوم در بازو، شانه، گردن، پستان یا ناحیه آسیب‌دیده یا هر ترکیبی از این‌ها می‌شود.[۹]

لنف ادم به یک مسئله با تاثیر بالا تبدیل شده است که کیفیت زندگی را در بازماندگان سرطان پستان به‌شدت کاهش می‌دهد.[۱۰]

با درمان زود هنگام و مداوم، پیشرفت بیماری به‌طور مشخصی آهسته شده و آسیب‌های بافتی به حداقل می‌رسد. هرچه درمان سریع‌تر آغاز شود شانس بهبودی بالاتر است. به همین خاطر وجود روشی که بتواند احتمال بروز این عارضه را پیش‌بینی کند، ضروری و حایز اهمیت است.

روش بررسی

به‌دلیل اهمیت موضوع در زمینه پزشکی تحقیقات بسیاری انجام شده است. از آن جمله می‌توان به چند نمونه اشاره داشت.

Hemmati و همکاران عوامل موثر در ایجاد لنف ادم پس از درمان اولیه کار سینوم مهاجم سینه را با آنالیز تک متغیری و چند متغیری با به‌کارگیری مدل رگرسیون لجستیک (Logistic regression) مورد بررسی قرار دادند.

در آنالیز تک متغیری، چاقی براساس شاخص توده بدنی به‌صورت معناداری بر ایجاد لنف ادم موثر بود. میانگین غدد لنفاوی درگیر در بیماران واجد و فاقد لنف ادم تفاوت معناداری داشت.[۱۱]

Azizi و همکاران به بررسی اپیدمیولوژیک خصوصیات بیماران مبتلا به لنف ادم به‌دنبال سرطان پستان پرداختند که در این پژوهش فراوانی بیماران در هر دسته بر حسب درصد مشخص شد. در نهایت ۷۸/۵٪ بیماران سن بالای ٤٥ سال داشتند، ۵۳٪ تحصیلات زیر ۱۲ سال، ۵۹٪ نمایه بدنی بالای ۳۰، ۷۷/۲٪ فعالیت فیزیکی متوسط و ٤/۷٪ فعالیت فیزیکی بالا داشتند. در ۷۱/۱٪ بیماران عمل جراحی از نوع ماستکتومی تعدیل شده انجام شده بود. در ۷۰/۵٪ بیماران تعداد بیش از ۱۰ گره لنفاوی خارج شده بود.[۸]

Haghighat و همکاران از پرونده‌های ٤۱۰ بیمار سرطان پستان از سه مرکز در تهران و مشهد که در ۱۲۳ مورد آن‌ها لنف ادم توسعه یافته بود، استفاده کردند. در این مقاله فراوانی نسبی لنف ادم در جمعیت تحت مطالعه ۳۰٪ بود. میانگین سنی گروه مورد ۵۰/٦ سال و در گروه شاهد ۸٤/٤ سال بود.[۱۲]

Ghasem و همکاران با به‌کارگیری تکنیک‌های داده‌کاوی به پیش‌بینی عود سرطان پستان پرداختند. براساس نتایج حاصله، متغیرهای موثر بر عود بیماری مثل میزان درگیری غدد لنفی، اندازه تومور و التهاب غدد لنفی که توسط متخصصین نیز جزو عوامل خطر عود محسوب می‌شوند، استخراج شدند.[۱۳]

Sharifkhani و همکاران از سه روش شناخته شده در داده‌کاوی به نام‌های درخت تصمیم CHAID و C5.0 و شبکه عصبی مصنوعی، برای پیش‌بینی احتمال ابتلا به پوکی استخوان استفاده کردند. در این پژوهش اطلاعات مربوط به ٦۷۱ بیمار در چهار بخش اطلاعات فردی، سبک زندگی، اطلاعات بیماری‌ها و نتایج دستگاه DEXA بررسی شد و ویژگی‌های تاثیرگذار بر این بیماری شناسایی شدند. از روش‌های مورد استفاده در این مطالعه یعنی الگوریتم‌های درخت تصمیم C5.0 و شبکه عصبی مصنوعی در تعیین مدل نهایی الگوریتم یادگیری جامع این پژوهش استفاده شده است.[۱٤]

Safdari و همکاران با به‌کارگیری مدل‌های درخت‌های تصمیم و شبکه عصبی، اطلاعات مربوط به ۳۵۱ بیمار قلبی–عروقی را در سال ۱۳۹۱ گردآوری کردند. این اطلاعات با به‌کارگیری جدول مورگان (Morgan table) از بین پرونده بیماران مراجعه‌کننده به بیمارستان شهید رجایی تهران به‌دست آمد. هدف اصلی این مطالعه پیش‌گویی احتمال ابتلا افراد به آنفارکتوس قلبی با به‌کارگیری درخت تصمیم براساس ریسک فاکتورهای موثر بر ابتلاست. از این مطالعات می‌توان الگوریتم‌های نهایی با دقت بالا در پیش بینی ابتلا به بیماری‌های خاص را استخراج کرد. بنابراین از مدل هی درخت تصمیم و شبکه عصبی می‌توان در مدل نهایی یادگیری جمعی استفاده کرد.[۱۵]

Fazeli و همکاران از روش‌های شبکه‌های عصبی و ماشین بردار پشتیبان استفاده کردند. پس از پیش پردازش داده‌ها، ۹۳۳ داده شامل ۲۵ متغیر به‌عنوان ورودی مدل‌ها انتخاب شدند. عوامل مرتبط با لنف ادم در بیماران مبتلا به سرطان پستان پیش‌بینی شد و احتمال ابتلا به لنف ادم در بیماران سرطان پستان پیش‌بینی گردید.

حساسیت و دقت الگوریتم‌ها برای پیش‌بینی احتمال ادم لنفاوی به‌ترتیب ۸۲/۸۷٪ و ۷۷/۴۹٪ توسط ماشین بردار پشتیبان در مقابل ۷۹/۳۷٪ و ۷۲/۴۱٪ برای شبکه‌های عصبی نشان داد و در نهایت با توجه به نتایج SVM، مهم‌ترین عوامل خطر لنف ادم شامل سنگینی، نوع جراحی، شاخص توده بدنی، نسبت تعداد غدد لنفاوی درگیر به تعداد غدد لنفاوی برداشته شده، رادیوتراپی، متاستاز، سن بیمار در هنگام تشخیص و تعداد غدد لنفاوی درگیر و شغل بودند که در مدل نهایی استخراج شده است.[۱۶]

هدف اصلی مطالعه Mosayebi و همکاران مقایسه الگوریتم‌های مختلف داده‌کاوی برای انتخاب دقیق‌ترین مدل برای پیش‌بینی عود سرطان پستان است.در این مطالعه مقطعی، جمع‌آوری داده‌ها از خرداد ۱۳۹۷ تا خرداد ۱۳۹۸ از آمار رسمی وزارت بهداشت، درمان و آموزش پزشکی و مرکز تحقیقات سرطان ایران برای بیماران مبتلا به سرطان پستان که به‌مدت پنج سال مورد پیگیری بوده‌اند، انجام شده است و شامل ۵۴۷۱ رکورد مستقل می‌باشد.[۱۷]

در این راستا پژوهش حاضر با دو هدف اصلی تعیین مدل مناسب در جهت پیش‌بینی میزان احتمال ابتلا به لنف ادم و تعیین عوامل موثر و پیش‌بینی‌کننده لنف ادم براساس مدل مناسب‌تر انجام شد. در این پژوهش ورودی نمونه‌ها از بانک اطلاعاتی شامل اطلاعات بیماران مراجعه‌کننده به مرکز توانبخشی سید خندان، طی سال‌های ۱۳۸۸ تا ۱۳۹۶ می‌باشد که به شرح زیر است:

نام خانوادگی، تاریخ تولد، قد، وزن، شغل بیمار، دست غالب، تحصیلات، وضعیت تاهل، محل زندگی، بخش درگیر (به گفته بیمار)، نوع نمونه‌برداری، اندازه تومور، نمونه‌برداری از غدد لنفاوی نگهبان، تعداد غدد لنفاوی خارج شده، تعداد غدد لنفاوی درگیر، مرحله بیماری، گیرنده‌های استروژن، گیرنده‌های پروژسترون، وجود و یا عدم وجود متاستاز، تاریخ تشخیص بیماری، نوع عمل جراحی، تاریخ جراحی، شیمی درمانی، رژیم شیمی درمانی، تعداد جلسات شیمی درمانی، پرتو درمانی، هورمون درمانی، داروی هرسپتین (Herceptin)، بیماری‌های همراه، فعالیت فیزیکی، شکایت اصلی بیمار هنگام مراجعه، درد، احساس سنگینی، احساس سوزن سوزن شدن (پارستزی)، سابقه آسیب عضو در اثر انجام کار سنگین، سابقه عفونت، تعداد دفعات عفونت، اندام مبتلا با تشخیص پزشک، تورم، بافت فیبروتیک، دامنه حرکت اندام درگیر، قسمت مبتلا به لنف ادم، درجه لنف ادم، تشخیص بیماری، تعداد جلسات درمانی (روز)، حجم تورم.

در ابتدا داده‌های ۱۱۱۷ بیمار به‌عنوان ورودی خام گردآوری شده‌اند و پس از پیش‌پردازش اولیه (که براساس اطلاعات اولیه پزشکی و از پردازش آماری با استفاده از نرم‌افزار SPSS Clementine Modeler 10.1 انجام شده است)، ۹۷۰ داده نهایی به‌عنوان ورودی اولیه تعریف شده‌اند. تعدادی از متغیرهای کیفی بیماران در جدول ۱ و متغیرهای کمی در جدول ۲ آمده است. این ویژگی‌ها براساس خروجی نهایی که در مدل ما سهم بیشتری در نتیجه را ایفا می‌کردند، انتخاب شده‌اند.

در هر تکرار از الگوریتم ژنتیک دو مرحله کلی وجود دارد: مرحله اول ارزیابی شایستگی راه حل‌های تولید شده، و دیگری بروزرسانی جمعیت (تولید جمعیت جدید). این دو مرحله پی‌درپی به‌صورت تکراری اجرا می‌شوند، تا زمانی‌که شرط خاتمه اجرا شود.

شرط خاتمه در این تحقیق به اتمام رسیدن تعداد تکرارهای الگوریتم تعیین شده است. پس از تعیین متغیرهای مربوطه و پیش‌پردازش داده‌ای که شامل حذف داده‌های تکراری، حذف متغیرهای زاید، شناسایی داده‌های مفقود، کاهش مقادیر متغیرها و غیره است، با استفاده از الگوریتم‌های داده‌کاوی و زبان برنامه‌نویسی مطلب می‌توان راه‌حل‌های زیر را پیشنهاد داد.

نکته حایز اهمیت تعیین بهترین الگوریتم‌های داده‌کاوی برای ترکیب در الگوریتم یادگیری جمعی می‌باشد. برای رسیدن به این منظور، با استفاده از الگوریتم‌های پایه C5، KNN و SVM با کرنل‌های متفاوت (Linear، RBF و Polynomial)، LDA، BAYES، MLP، و تایید دقت بالا برای هر کدام از الگوریتم‌ها به این نتیجه رسیدیم که در روش‌های اول و دوم از الگوریتم‌های پایه تست شده در الگوریتم یادگیری جمعی استفاده شود. الگوریتم MLP به‌دلیل ماهیت اولیه رندوم به‌طور کل از راه حل‌های بعدی حذف شده است.

۱- بهینه‌سازی روش یادگیری جمعی با تمام ویژگی‌ها با استفاده از الگوریتم ژنتیک: در این روش از پنج الگوریتم C5، KNN، BAYES، LDA و SVM استفاده شده است با این تفاوت که به خروجی این الگوریتم‌ها وزن اختصاص داده می‌شود. شکل (۱) یک کروموزوم برای الگوریتم ژنتیک است. بنابراین در این حالت فقط وزن‌های خروجی یادگیری جمعی، با الگوریتم ژنتیک بهینه می‌شوند. بنابراین یک راه حل برای این قسمت تعیین ضرایب وزنی M طبقه‌بند می‌باشد ($۰\leq W\leq ۱$). هر اندازه که این وزن‌دهی به سمت یک نزدیک‌تر باشد، الگوریتم طبقه‌بندی مربوطه تاثیر بیشتری در تعیین خروجی نهایی خواهد گذاشت.

۲- روش ترکیبی یادگیری جمعی و استخراج ویژگی: یک بردار (M+N) که M تعداد طبقه‌بندهای یادگیری جمعی و N تعداد ویژگی‌ها را مشخص می‌کند و ۱≥W≥۰ و ویژگی‌ها صفر یا ۱ می‌باشند. (شکل ۲).

جدول۱: توزیع فراوانی متغیرهای کیفی بیماران در گروه‌های مورد مطالعه

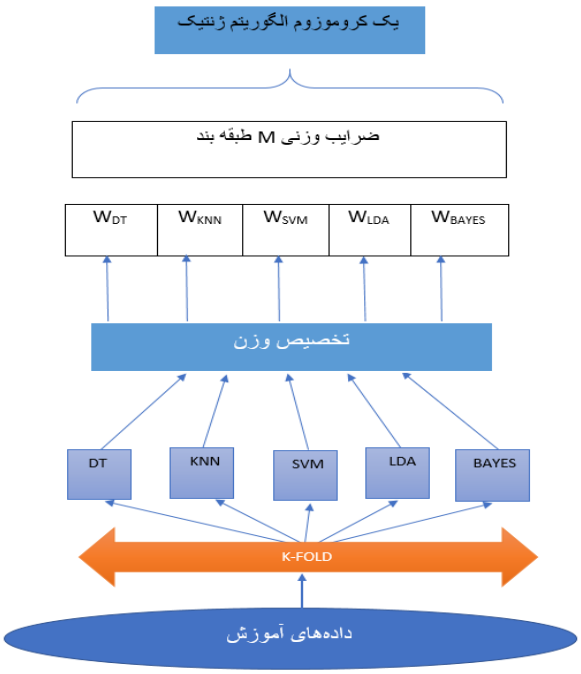
متغیر	سطوح متغیر	ابتلا به لنف ادم (n=۷٤۰) فراوانی (درصد)	عدم ابتلا به لنف ادم (n=۲۳۰) فراوانی (درصد)	P
سن (سال)	کمتر از ٤۰ سال	۱۳۰(۱۷/٥٦)	٥٥(۲۳/۹۱)	۰/۰۲۱
	٦۰-٤۰ سال	٤۸۱(٦٥)	۱٥۹(۸۲/۱۷)	
	بیشتر از ٦۰ سال	۱۲٦(۱۷/۰۲)	۳۱(۱۳/٤۸)	
شغل	خانه‌دار	٥۷۳(۷۷/٤۳)	۱٦۸(۷۳/۰٤)	۰/۷۰۱
	معلم+کار اداری	۱۰۷(۱٤/٤٥)	٤٦(۲۰)	
	کار فیزیکی	۱۷(۲/۲۹)	٥(۲/۱۷)	
تحصیلات	دیپلم و پایین‌تر	٤۹۹(٦۷/۰٦)	۱۳۱(٥٦/۹٥)	۰/۰۰۱
	دانشگاهی	۲۰٥(۲۷/۷)	۹٤(٤۰/۸٦)	
وضعیت تاهل	مجرد	۱۱۱ (۱٥)	۳۸(۱٦/٥۲)	۰/٦٥٤
	متاهل	٦۱٦(۸۳/۲٤)	۱۹۱(۸۳/۰٤)	
نسبت دست درگیر و غالب	همسو	۳۱۱(٤۲/۰۲)	۸٤(۳٦/٥۲)	۰/۰۳۱
	غیر همسو	۳۹۷(۲۳/٦٤)	۱٤۲(٦۱/۷۳)	
شیمی درمانی	دریافت نکرده	۲۸(۳/۷۸)	۱۹(۸/۲٦)	۰/۰۰٤
	دریافت کرده	٦۹۳(۹۳/٦٤)	۲۰۷(۹۰)	
پرتو درمانی	دریافت نکرده	٤٤(٥/۹٤)	۳٥(۱٥/۲۱)	۰/۰۰۰
	دریافت کرده	٦٥۱(۸۷/۹۷)	۱۷۰(۷۳/۹۱)	
هورمون درمانی	دریافت نکرده	۱٤٥(۱۹/٥۹)	٤٥(۱۹/٥٦)	۰/۰۰۲
	دریافت کرده	٥۳۹(۷۲/۸۳)	۱٥۱(٦٥/٦٥)	
احساس سنگینی	ندارد	۲٤٥(۳۳/۱)	۱۱۱(٤۸/۲٦)	۰/۰۰۰
	کم	۲۲۹(۳۰/۹٤)	۷٥(۳۲/٦۰)	
	متوسط	۱۲۹(۱۷/٤۳)	۲۷(۱۱/۷۳)	
	شدید	۱۱۹(۱٦/۰۸)	۱۱(٤/۷۸)	
شاخص توده بدنی	کمبود وزن و معمولی	۱۳٦(۱۸/۳۷)	٦٤(۲۷/۸۲)	۰/۰۰۱
	اضافه وزن	۲٥٤(۳٤/۳۲)	۹۸(٤۲/٦۰)	
	چاق	۲۲۸(۳۰/۸۱)	٤۸(۲۰/۸٦)	

آزمون مقایسه متغیرها با Chi-square test و P=۰/۰٥

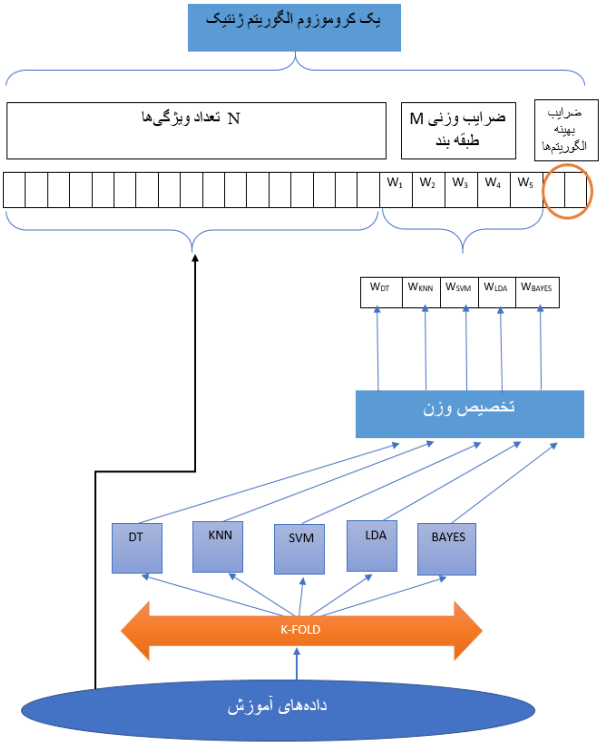
جدول۲: توزیع فراوانی متغیرهای کمی بیماران در گروه‌های مورد مطالعه

متغیرهای کمی	میانگین (±انحراف معیار)	میانگین (±انحراف معیار)	P
اندازه تومور	۱/٤۲(۲/۸۷)	۱/۳٥(۲/٥٥)	۰/۰۰۳
نسبت غدد لنفاوی خارج شده به غدد لنفاوی درگیر	۰/۳۰۳(۰/۳۳۳۷)	۰/۲٤۳(۰/۱۸)	۰/۰۰۰
تعداد دوره شیمی درمانی	۷/۹٥(۲/٦۹)	۷/٦۸(۲/۸۲)	۰/۱۸۷
دامنه حرکت عضو درگیر	۱٥۹/۸۰(۳۸/۰۸)	۱٦٥/۳٥(۲٥/۹۰)	۰/۰٤۱

آزمون متغیرهای کمی t و P=۰/۰٥



شکل ۱: بهینه‌سازی روش یادگیری جمعی با تمام ویژگی‌ها با استفاده از الگوریتم ژنتیک



شکل ۲: روش ترکیبی یادگیری جمعی و استخراج ویژگی

در این روش در ابتدا به‌جای استفاده از گروه داده‌ای آموزش و آزمایش از روش K-fold استفاده شد. یعنی در ابتدا داده‌ها به گروه‌های داده‌ای تقسیم شده و پس از هر بار تکرار گروه داده‌ای آموزش و آزمایش متغیر و متفاوت از دفعه تکرار پیشین الگوریتم هستند.

در مرحله بعد باید داده‌هایی که در نودهای ورودی آخرین مرحله الگوریتم ژنتیک قرار دارند را تعریف کنیم:

۱- ضرایب وزنی برای هر الگوریتم پایه، در نظرگرفته می‌شود. بنابراین هرکدام از خروجی‌ها، براساس ضریب وزنی‌شان به‌عنوان یک داده ورودی برای الگوریتم ژنتیک در نظر گرفته می‌شود.

الف) ویژگی‌های خروجی برای هر طبقه‌بند یادگیری انتخابی، با استفاده از الگوریتم ژنتیک بهینه می‌شود. در واقع در این روش یک کروموزوم به‌صورت زیر در نظر گرفته شده است. (شکل ۳).

بنابراین یک کروموزوم شامل سه قسمت می‌شود که به‌صورت S.FS (یک راه حل انتخاب ویژگی)+S.EL (یک راه حل یادگیری انتخابی)+S.P (یک راه حل برنامه یعنی تعداد بهینه شده K در الگوریتم KNN و نوع بهینه کرنل الگوریتم SVM) می‌باشد.(شکل ۳ و ٤).

S.FS	S.EL	S.P

شکل ۳: بهینه‌سازی روش ترکیبی یادگیری جمعی و استخراج ویژگی

S.FS:

۰	۱	۱	۱	۰	۰	۱	۱	۰

ویژگی‌های انتخاب‌شده

S.EL:

W_{C5}	W_{LDA}	W_{KNN}	W_{Bayes}	W_{SVM}

S.P:

نوع بهینه کرنل الگوریتم SVM	تعداد بهینه شده K در الگوریتم KNN

شکل ٤: اجرای تشکیل‌دهنده یک کروموزوم الگوریتم ژنتیک در مدل ترکیبی بهینه شده

نکته بسیار مهم: در تمامی مدل‌ها، یک کروموزوم الگوریتم ژنتیک علاوه بر متغیرهای ذکر شده، یک متغیر جهت انتخاب بهینه K

در الگوریتم KNN و یک متغیر جهت انتخاب بهینه مدل SVM در نظر گرفته شده است تا بتوان امیدوار بود که در کل الگوریتم هیچ متغیری به‌صورت دستی وارد سیستم نشده است و تمامی متغیرها در بهینه‌ترین حالت آن داده وارد سیستم می‌شود. درایه‌های S.FS در این مدل به‌صورت رابطه (۱) تعریف می‌شود:

رابطه ۱:

$$S.FS_i = \begin{cases} \text{اگر ویژگی i در راه حل حضور داشته باشد: ۱} \\ \text{اگر ویژگی i در راه حل حضور نداشته باشد: ۰} \end{cases}$$

درایه‌های S.EL به‌صورت رابطه (۲) تعریف می‌شود: ضریب تاثیر طبقه‌بند j ام در یادگیری انتخابی:

رابطه ۲: $S.EL_j=1\langle w_j\langle 0$

تابع هدف پروژه: ما تابع هدف پروژه (داده‌های بالینی و پزشکی بیماران) را یک تابع سه هدفه در نظر گرفتیم و این سه هدف را با ضریب وزنی به تابع تک هدفه تبدیل کردیم. دلیل وجود ضریب وزنی این است که متغیرهای مختلف از یک جنس خطا و درصدی هستند، بنابراین می‌توان ضریب وزنی تعریف کرد.

می‌دانیم اگر ضریب وزنی بیشتر باشد یعنی آن متغیر مهم‌تر است و اگر ضریب وزنی کوچکتر باشد آن متغیر بی‌اهمیت‌تر است. (رابطه ۳)

رابطه ۳: $Cost_i=\min\ (W_{c1}\times Err_i+W_{c2}\times FPR_i+W_{c3}\times FRR_i)$

در هر تکرار پس از بروزرسانی جمعیت، راه‌حل‌های تولید شده مورد ارزیابی قرار می‌گیرند. در این مقاله از عملگر ترکیب استفاده شده است که در قسمت بروزرسانی جمعیت به‌طور خاص به ارزیابی راه حل خواهیم پرداخت.

در رابطه (۲)، کلاس هدف افراد سالم هستند و $Cost_i$ خطای کروموزوم i ام است. Err، FPR و FRR به‌ترتیب خطای طبقه‌بندی، نرخ پذیرش اشتباه و نرخ رد اشتباه به‌ازای زیر مجموعه ویژگی متناظر با هر راه حل هستند.

بروزرسانی جمعیت در الگوریتم ژنتیک از سه بخش تشکیل شده است: انتقال مستقیم درصدی از بهترین کروموزوم‌های نسل جاری، عملگر ترکیب، و عملگر جهش. درصد تولید کروموزوم‌های نسل بعد با استفاده از این سه عملگر به‌ترتیب برابر با $P_{Recombination}$، $P_{Crossover}$ و $P_{Mutation}$ است که مجموع این سه درصد برابر با ۱ است.

جدول ۱: ارزیابی الگوریتم کلاس‌بندی

در این تحقیق مقدار این سه ضریب به‌ترتیب برابر با۰/۱، ۰/۵، ۰/٤ در نظر گرفته شده است. روش‌های مختلفی برای انتخاب والدین در الگوریتم ژنتیک وجود دارد. در این مقاله از روش انتخاب چرخ رولت (Roulette–Wheel) با توان ۲ برای انتخاب والدین در عملگر ترکیب و جهش استفاده شده است.

هر یادگیر پایه به‌طور جداگانه تحت ویژگی‌های انتخاب شده خود آموزش می‌بینند. سپس، مدل‌های آموزش دیده جمع می‌شوند تا مشخص کند بیمار جدید آیا مبتلا به لنف‌ادم هست یا خیر (رابطه ٦).

رابطه ٦:

$$Out_i = \frac{W_1 Out^i_{KNN} + W_1 Out^i_{SVM} + W_1 Out^i_{BAYAS} + W_1 Out^i_{C5} + W_1 Out^i_{LDA}}{\sum_{j=1}^{5} W_j}$$

جایی که w1 تا w5 به‌ترتیب وزن یادگیرندگان پایه هستند به‌دلیل ماهیت فرضیات و صفر و یک بودن، هدف نهایی در این قسمت از پروژه باید یک حد آستانه به‌صورت زیر تعریف کرد: اگر خروجی از ۰/۵ بزرگتر باشد، جواب نهایی را یک در نظر می‌گیریم. به این معنا که بیمار مربوطه مبتلا به لنف ادم است. در غیر این‌صورت جواب را برابر صفر در نظر می‌گیریم یعنی بیمار مبتلا به لنف ادم نمی‌باشد.

یافته‌ها

خروجی قسمت اول که از الگوریتم‌های کلاس‌بندی استفاده شد، و علاوه بر ماتریس در هم‌ریختگی، ارزیابی FPR، FRR، Accuracy و Cost را در جدول (۱) نشان داده شده است.

همان‌طور که از خروجی جدول (۱) مشخص است، الگوریتم SVM با کرنل RBF بهترین نتیجه را در قسمت داده‌های پزشکی و بالینی دارد.

خروجی مربوط به دو روش ترکیبی، به این صورت است که علاوه بر ماتریس در هم ریختگی، ارزیابی FPR، FRR، Accuracy و Cost را در جدول (۲) نشان داده شده است.

نتایج ارزیابی و مقایسه حساسیت و دقت الگوریتم‌ها نشان داد، ترکیب روش یادگیری جمعی با الگوریتم‌های منتخب کلاس‌بندی دارای ضریب صحت ۸۷٪ و در روش یادگیری جمعی با الگوریتم‌های منتخب با روش بهینه شده استخراج ویژگی‌ها دارای ضریب صحت ۹۰٪ است.

نام الگوریتم	ماتریس در هم‌ریختگی		ارزیابی دقت الگوریتم
SVM (Linear)	۵۴۷	۱۴۳	۰/۷۷۰۶
	۱۸۵	۵۵۵	
SVM(RBF)	۶۳۷	۵۳	۰/۸۷۶۰
	۱۱۰	۶۳۰	
SVM(Polynomial)	۶۶۸	۲۲	۰/۸۶۲۲
	۱۷۵	۵۶۵	
LDA	۵۴۸	۱۴۲	۰/۷۵۸۰
	۲۰۴	۵۳۶	
KNN	۶۱۹	۷۱	۰/۷۴۵۵
	۲۹۳	۴۴۷	
Bayes	۵۷۰	۱۲۰	۰/۶۸۲۵
	۳۳۴	۴۰۶	
C5	۵۵۲	۱۳۸	۰/۸۱۹۶
	۱۲۰	۶۲۰	
	۵۰۱	۱۸۹	۰/۷۵۳۱
	۱۶۴	۵۷۶	

جدول ۲: ارزیابی الگوریتم‌های ترکیبی

نام الگوریتم	ماتریس در هم‌ریختگی		ارزیابی دقت الگوریتم
			FPR=۰/۱۹۷۳
روش اول (وزن‌دهی یادگیری جمعی)	۶۵۶	۳۴	**FRR=۰/۰۴۹۳**
	۱۴۶	۵۹۴	**Accuracy=۰/۸۷۴۱**
			Cost=۰/۱۲۴۸
			FPR=۰/۱۴۵۹
روش دوم (یادگیری جمعی و انتخاب ویژگی بهینه شده)	۶۵۸	۳۲	**FRR=۰/۰۴۶۴**
	۱۰۸	۶۳۲	**Accuracy=۰/۹۰۲۱**
			Cost=۰/۰۹۷۲

براساس ارزیابی نهایی، تاثیرگذارترین عوامل خطر لنف ادم به‌ترتیب شامل موارد زیر بود: تعداد غدد لنفاوی خارج شده (۷۳٪)، احساس سنگینی (۷۱٪)، تعداد غدد لنفاوی اثرگذار به تعداد غدد لنفاوی خارج شده (۷۱٪)، غیر همسو بودن دست غالب و درگیر (۶۸٪)، دامنه حرکت اندام درگیر (۶۸٪)، پرتو درمانی (۶۶٪)، نوع عمل جراحی (۶۵٪)، سن (۶۵٪)، تعداد غدد لنفاوی اثرگذار (۶۴٪)، درجه سرطان پستان (۶۰٪)، بافت فیبروتیک (۶۰٪)، احساس سوزن سوزن شدن (۶۰٪)، شاخص توده بدنی (۵۷٪)،

اندازه تومور (۵۷٪)، آسیب دست (۵۷٪)، تعداد جلسات شیمی درمانی (۵۷٪)، رژیم شیمی درمانی (۵٤٪)، شغل (٤۸٪)، متاستاز (٤۸٪)، هورمون درمانی (۳٤٪)، درجه لنف ادم (٤۳٪)، شیمی درمانی (۳۸٪).

این موارد پس از ۵۰ بار تکرار الگوریتم نهایی و پس دریافت خروجی و تعیین تعداد دفعات تکرار عامل موثر، به‌دست آمده است. طبق نتیجه نهایی که بر روی داده‌های بیماران در سال ۹۷ انجام شد، مدل ماشین بردار پشتیبان با دقت ۷۷/٤۹ بهترین کارایی را در بین الگوریتم‌های C5، C&RT، CHAID، QUEST، شبکه عصبی تخمین زدند (جدول ۳).[۱٦]

همان‌طور که از جدول (۳) مشخص است، روش ترکیبی الگوریتم‌های یادگیری جمعی و انتخاب ویژگی کلی به‌همراه الگوریتم بهینه‌سازی ژنتیک در میزان دقت الگوریتم موثر است.

همان‌طور که مشخص است در روش حاضر به‌دلیل استفاده از الگوریتم‌های منتخب در روش یادگیری جمعی می‌توان نتیجه گرفت که می‌توان به‌صورت جمعی الگوریتم‌ها در نتیجه نهایی بهبود ایجاد کرد. در هر دو مقاله ذکر شده از الگوریتم‌های تکی و پایه برای رسیدن به هدف استفاده شده است.[۱٦و۱۷]

بحث

در این پژوهش با ترکیب الگوریتم‌های منتخب داده‌کاوی در الگوریتم یادگیری جمعی، به این دلیل که از الگوریتم‌های داده‌کاوی به‌طور جداگانه استفاده نشده است و از تمام عملکردهای الگوریتم‌های داده‌کاوی مربوطه به‌صورت جمعی استفاده کرده‌ایم، دقت مدل را به‌طور چشمگیری نسبت به استفاده تکی از الگوریتم‌ها افزایش دادیم.

طبق نتیجه نهایی که بر روی داده‌های بیماران در سال ۹۷ انجام شد، Fazeli و همکاران، با استفاده از پایگاه داده یکسان با پژوهش ما برای تعیین عوامل تاثیرگذار از الگوریتم‌های پایه استفاده کردند و به این نتیجه رسیدند که مدل ماشین بردار پشتیبان با دقت ۷۷/٤۹، بهترین کارایی را در بین الگوریتم‌های C5، C&RT، CHAID، QUEST، SVM، شبکه عصبی تخمین زدند.[۱٦] همچنین در مطالعه Mosavebi و همکاران پس از پیش‌پردازش اولیه در مجموعه داده‌ها و متغیرها، هفت الگوریتم داده‌کاوی به‌کار گرفته شده است که هر یک نشان‌دهنده یک نوع رویکرد داده‌کاوی است.[۱۷]

جدول ۳: مقایسه روش الگوریتم‌های تکی SVM با روش مورد استفاده، با داده‌های یکسان[۱٦]

ارزیابی[۱۷]	ارزیابی	نام الگوریتم
	FPR=۰/۲٤۵٦	SVM (Linear)
	FRR=۰/۲۰۱۸	
	Accuracy=۰/۷۷۱۸	
FPR=۰/۸۸۱۸	FPR=۰/۱۳۱۲	SVM (RBF)
FRR=۰/۵٤٤۱	FRR=۰/۰۵۲۸	
Accuracy=۰/۷۷٤۹	Accuracy=۰/۸۹٦٤	
	FPR=۰/۲۱٤۸	SVM (Polynomial)
	FRR=۰/۳۰۸۱	
	Accuracy=۰/۸۸۵۱	

نتایج نشان می‌دهد که الگوریتم C5.0 احتمالا می‌تواند ابزار مفیدی برای پیش‌بینی عود سرطان پستان در مرحله عود دور و عدم عود، به‌ویژه در سال‌های اول تا سوم باشد. در این مقاله بیشترین دقت

مربوط به الگوریتم C5، ۸۱٪ و پس از آن KPCA-SVM با دقت ۷۸٪ در جایگاه دوم قرار دارد.[۱۸]

در کارهای اخیر، الگوریتم‌های به‌کار رفته به‌صورت الگوریتم‌های پایه کامپیوتری و یا بهبود یافته با الگوریتم ژنتیک و الگوریتم‌های دیگر است.

در مطالعه حاضر از نظر کامپیوتری، به‌دلیل استفاده همزمان الگوریتم‌های منتخب کلاس‌بندی و الگوریتم‌های K-fold و یادگیری انتخابی و در نهایت برای بهبود کار الگوریتم ژنتیک، دقت مدل تا حد قابل قبولی افزایش پیدا کرده است. بنابراین با استفاده از الگوریتم‌های منتخب داده‌کاوی در روش یادگیری جمعی و ترکیب بهینه این روش با انتخاب ویژگی و نهایتا الگوریتم ژنتیک، به دقت بالایی برای احتمال ابتلا به لنف ادم در بیماران سرطان پستان رسیدیم. می‌توان از این روش در مدل‌های داخلی سیستم‌های تصمیم‌یار استفاده کرد تا پزشک متخصص به‌راحتی و با دقت بالا بتواند احتمال ابتلا را بررسی و از بروز این بیماری تا حد قابل توجهی بکاهد. این موارد باعث رضایت‌مندی و افزایش امید به زندگی در بیماران سرطان پستان می‌شود.

References

1. Keihanian S, Ghaffari F, Fotokian Z, Shoormig R, Saravi M. Risk factors of breast cancer in Ramsar and Tonekabon. *J Qazvin Univ Med Sci* 2010;14(2):12-9
2. Prabha S, Sujatha C. Proposal of index to estimate breast similarities in thermograms using fuzzy C means andanisotropic diffusion filter based fuzzy C means clustering. *Infrared Phys Technol* 2018;93:316-25.
3. Díaz-Cortés M-A, Ortega-Sánchez N, Hinojosa S, Oliva D, Cuevas E, Rojas R, et al. A multi-level thresholding method for breast thermograms analysis using Dragonfly algorithm. *Infrared Phys Technol* 2018;93:346-61.
4. Suganthi S, Ramakrishnan S. Anisotropic diffusion filter based edge enhancement for segmentation of breastthermogram using level sets. *Biomed Signal Process Control* 2014;10:128-36.
5. Ng E-K. A review of thermography as promising non-invasive detection modality for breast tumor. *Int J Therm Sci* 2009;48(5):849-59.
6. Asadi s. Evaluation of risk factors for lymphedema in patients with breast cancer. (PhD Thesis), Shahid Beheshti University of Medical Sciences, School of Medicine. (2008).
7. Zakeri ZS. Determining the epidemiologic charecterestics of patients suffering from breast cancer related Lymphedema reffering to Martyr Motahari Lymphedema Clinic (1388-1391). *Iran Breast Dis* 2014;7(1):24-8.
8. Azizi F, Hatami H, Janghorbani M. Epidemiology and control of common diseases in Iran. *Tehran: Eshtiagh Publications* 2000:602-16.
9. Fu MR, Ridner SH, Armer J. Post-breast cancer. Lymphedema: Part 1. *Am J Nurs* 2009;109(7):48-54.
10. Dahl AA, Nesvold IL, Reinertsen KV, Fosså SD. Arm/Shoulder problems and insomnia symptoms in breast cancer survivors: Cross-sectional, controlled and longitudinal observations. *Sleep Med* 2011;12(6):584-90.
11. Hemmati S, Jabbari H, Akbari M, Tajvidi M, Rankoohi PA. Factors Associated with the Severity of Lymphedema after the Treatment of Invasive Breast Cancers. *J Isfahan Med School* 2013;30(211).
12. Shahpar H, Atieh A, Maryam A, Fatemeh HS, Massoome N, Mandana E, et al. Risk factors of lymph edema in breast cancer patients. *Int J Breast Cancer* 2013;2013.
13. Qasem Ahmad L, Tolouei Ashlaghi A, Poorabrahimi M. Predicting the recurrence of breast cancer using three data mining techniques. *J Breast Patients* 2013;5(4):23-34.
14. Mona S, Somayeh A, Abbasi M, Ameri H. Providing a model for predicting the risk of osteoporosis using decision tree algorithms. *J Mazandaran Univ Med Sci* 2014;24(116):110-8.
15. Safdari, R, Ghazi SM, Gharooni M, Nasiri M, Arji G. Predicting the risk of myocardial infarction by Decision Tree Method. *J Paramedical Sci Rehabilitation* 2013;3:26-37.
16. Fazeli M, Kazemi A, Haghighat S. Predicting the Risk of Lymphedema in Breast Cancer Patients by Using Data Mining Techniques. *Multidisciplinary Cancer Investigation* 2017;1:0-.
17. Mosayebi A, Mojaradi B, Bonyadi Naeini A, Khodadad Hosseini SH. Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. *PloS One* 2020;15(10):e0237658

TUMJ Tehran University Medical Journal, February 2022; Vol. 79, No. 11: 863-872 Original Article

The prediction of lymphedema via the combination of the selected data mining algorithms

Anaram Yaghoobi Notash Ph.D.[1]
Peiman Bayat Ph.D.[1*]
Shahpar Haghighat Ph.D.[2]
Ali Yaghoobi Notash Ph.D.[1,3]

1- Department of Computer Engineering, Faculty of Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran.
2- Breast Cancer Research Center, Motamed Cancer Institute, ACECR, Tehran, Iran.
3- Department of Surgery, Sina Hospital, School of Medicine, Tehran University of Medical Sciences, Tehran, Iran.

[*]Corresponding author: Department of Computer Engineering, Rasht Branch, Islamic Azad University, Rasht, Iran.
Tel: +98-13-33422153
E-mail: bayat@iaurasht.ac.ir

Abstract

Received: 17 Oct. 2021 Revised: 24 Oct. 2021 Accepted: 14 Jan. 2022 Available online: 21 Jan. 2022

Background: Breast cancer is the second leading cause of cancer death in women, after lung cancer. Due to the importance of predicting this disease, the use of data mining methods in medical research is more significant than before. Data mining algorithms can be a great help in preventing the development of lymphedema in patients. The aim of this study was to create a diagnosis system that can predict the probability of lymphedema in breast cancer patients.

Methods: In the present study, the factors of lymphedema in 1117 patients with breast cancer have been collected. The likelihood of developing lymphedema is predicted using ensemble learning via 5 heterogeneous classification algorithms, feature selection and the genetic algorithm (The Two-layer Ensemble Feature Selection method). After collecting the data of patients with breast cancer from 2009 to 2018, and data preprocessing using the optimized ensemble learning algorithm and feature selection, we will examine the likelihood of developing lymphedema for the new patient. Finally, the factors affecting the disease have been extracted. Excluding the time of collecting statistical data, the period of the study was from September 2019 to February 2021. This study is performed at Seyed Khandan Rehabilitation Center, Tehran, Iran.

Results: The results of algorithms showed that the accuracy of the ensemble learning method with selected classification algorithms (SVM with RBF kernel) is 87% and the accuracy of the ensemble learning with feature selection method is 90%. According to the final evaluation of the proposed method, the most effective risk factors for lymphedema have been extracted.

Conclusion: Unfortunately, treatment and diagnosis are not without complications, and one of the most important of these complications in breast cancer is lymphedema in the upper extremities, which can affect the quality of life in patients. It is essential to have a method that can accurately suggest to a specialist whether a new patient will develop lymphedema in the future or how likely it is to develop it, using patient's own clinical and demographic characteristics.

Keywords: breast cancer lymphedema, classification, data mining.