مجله پیاورد سلامت
دوره ۱۵، شماره ٤، مهر و آبان ۱٤۰۰، ۳۵۲-۳٤۰
مقاله پژوهشی

تشخیص ابتلا به سرطان پستان با بهره‌گیری از یادگیری ماشین

کسری دولت‌خواهی[۱]، عادل آذر[۲]، تورج کریمی[۳*]، محمد هادی‌زاده[٤]

دریافت مقاله: تیر ۱٤۰۰
پذیرش مقاله: مهر ۱٤۰۰

* نویسنده مسئول:
تورج کریمی؛
دانشکده مدیریت و حسابداری پردیس فارابی دانشگاه تهران

Email :
tkarimi@ut.ac.ir

چکیده

زمینه و هدف: سرطان و به‌طور جزیی سرطان پستان در زمره بیماری‌هایی به‌شمار می‌روند که در ایران پس از بیماری‌های قلبی بیش‌ترین آمار مرگ ومیر را به خود اختصاص داده است. پیش‌بینی صحیح سرطان پستان دارای اهمیت است و وجود علایم و ویژگی‌های مختلف این بیماری، تشخیص را برای پزشکان دشوار می‌کند. هدف این پژوهش، شناسایی عوامل موثر بر سرطان پستان و تشخیص احتمال ابتلا به سرطان پستان است.

روش بررسی: در مطالعه‌ی حاضر، ابتدا به روش تحلیل محتوا و مطالعات کتابخانه‌ای، عوامل تاثیرگذار در ابتلا به سرطان پستان شناسایی شده سپس با همراهی تیم خبرگان مشتمل بر پزشکان متخصص و یا دارای فوق‌تخصص سرطان‌شناسی و جراحی پستان با کمک روش دلفی، تعدیل گردیده و ۲٦ عامل نهایی که به‌صورت عددی صحیح و رشته‌ای بودند براساس شرایط بومی و اقلیمی تایید شدند. در ادامه و با توجه به عوامل نهایی و براساس پرونده پزشکی ۵۲۰۸ بیمار در مرکز تحقیقات سرطان دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهیدبهشتی به منظور تشخیص ابتلا به سرطان از روش‌های درخت تصمیم (Decision Tree)، جنگل تصادفی (Random Forest) و ماشین‌بردار پشتیبان (Support Vector Machine) به‌عنوان روش‌های یادگیری ماشین بهره گرفته شد.

یافته‌ها: در گام نخست و با روش تحلیل محتوا، ۲۹ عامل تاثیرگذار در ابتلا به سرطان پستان شناسایی شد. در ادامه و با در نظر گرفتن شرایط بومی و اقلیمی و با استفاده از روش دلفی و با بهره‌گیری از نظرات ۱۸ خبره در طی سه دوره، ۲٦ عامل تعدیل و نهایی شد. در گام نهایی و با استفاده از پرونده پزشکی مراجعه‌کنندگان که در طی ۳ سال گردآوری شده و معیارهای استخراج‌شده از سه روش ذکر شده، جنگل تصادفی، بیشترین دقت به میزان ۹٤/۷۵٪ و صحت ۹۷/۲٦٪ را در تشخیص ابتلا به سرطان پستان به خود اختصاص داد، که این میزان در قیاس با سایر پژوهش‌های مشابه که از پایگاه‌های داده بومی بهره گرفته‌اند، دقت‌های به‌دست آمده بسیار نزدیک به کارهای پیشین بوده و در بعضی موارد نیز دقت بهتری داشته است.

نتیجه‌گیری: با استفاده از روش جنگل تصادفی و با بهره‌گیری از عوامل تاثیرگذار بر سرطان پستان، قابلیت تشخیص ابتلا به سرطان با بیشترین دقت فراهم شده است.

واژه‌های کلیدی: سرطان پستان، تحلیل محتوا، روش دلفی، جنگل تصادفی، درخت تصمیم، ماشین‌بردار پشتیبان

۱ دانشجوی دکتری مدیریت صنعتی گرایش تحقیق در عملیات، دانشکده مدیریت و حسابداری، پردیس فارابی دانشگاه تهران، قم، ایران

۲ استاد گروه مدیریت صنعتی، دانشکده مدیریت و اقتصاد، دانشگاه تربیت مدرس، تهران، ایران

۳ استادیار گروه مدیریت صنعتی و تکنولوژی، دانشکده مدیریت و حسابداری، پردیس فارابی دانشگاه تهران، قم، ایران

٤ استادیار، مرکز تحقیقات سرطان، دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهیدبهشتی، تهران، ایران

مقدمه

سرطان یکی از مسایل مهم و اصلی بهداشت و درمان در ایران و تمام دنیاست. در کشورهای در حال پیشرفت نیز بیماری سرطان در ردیف مسایل مهم بهداشتی‌درمانی بوده و روند آن رو به افزایش است. 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(۱). سرطان پستان تهدید بزرگی بر سلامت زنان بوده و از عوامل شایع در کاهش عمر زنان به شمار می‌رود. سرطان پستان ناشی از رشد خارج از قاعده‌ی سلول‌های غیرطبیعی در پستان است(۲).

سرطان پستان شایع‌ترین سرطان در ایران است و اولین سرطان زنان ایرانی به شمار می‌رود. میزان بروز آن حدود ۳۰ در ۱۰۰ هزار نفر از زنان ایرانی است. به‌دلیل آگاهی مردم و توسعه‌ی نظام سلامت در کشور درحال حاضر ۵۵٪ مبتلایان به سرطان پستان در مرحله اولیه و ۴۵٪ در مرحله پیشرفت کشف می‌شوند که نسبت به مقیاس جهانی ۳۰٪ دیرتر تشخیص داده می‌شود(۳). تاکنون تحقیقات گسترده‌ای درخصوص پیش‌بینی احتمال ابتلا به سرطان پستان(۴)، تشخیص مبتلابودن به سرطان پستان(۵) پیش‌بینی مدت زمان بقای بیمار مبتلا به سرطان پستان(۶) و در نهایت احتمال عود مجدد سرطان پستان(۷) صورت گرفته است. Delen و همکاران(۸) از شبکه‌های عصبی مصنوعی و درخت تصمیم برای توسعه‌ی مدل‌های پیش‌بینی سرطان پستان با تجزیه و تحلیل پایگاه داده Wisconsin Breast Cancer Database بهره گرفتند که دقت ۹۳/۶٪ حاصل شد. Chou و همکاران(۹) با بهره‌گیری از رویکرد رگرسیون چندمتغیره و شبکه‌های عصبی با دقت ۹۸/۲۵٪ خوشه‌بندی سرطان پستان از نوع خوش‌خیم و بدخیم داده‌های مجموعه WBCD را انجام دادند. روند بهره‌گیری از رویکرد شبکه‌های عصبی در سال‌های بعد ادامه‌دار بوده و تحقیقات Ubeyli (۱۰) با دقت ۹۹/۵٪، Karabatak و Ince (۱۱) با دقت ۹۷/۴٪ و Bhardwaj و Tiwari (۱۲) با دقت ۹۹/۲۶٪ انجام شده است. ارایه یک رویکرد ماشین بردار پشتیبان همراه درخت تصمیم توسط Bennett و Blue با دقت خوشه‌بندی ۹۷/۲٪ آغاز استفاده از این رویکرد است(۱۳). مطالعه‌ای که توسط Chao و همکاران انجام شد از ماشین‌بردار پشتیبان، رگرسیون لجستیک و درخت تصمیم به‌منظور طبقه‌بندی نرخ بقای بیماران مبتلا به سرطان پستان استفاده شد و براساس نتایج این مطالعه، ماشین‌بردار پشتیبان با دقت ۹۵/۱۵٪ بهترین روش بود(۱۴). دهقان و همکاران در مقاله‌ای با عنوان«مدل‌سازی بیماری سرطان پستان با استفاده از روش‌های مبتنی بر داده‌کاوی» و بهره‌گیری از پایگاه داده استاندارد University of California Irvine، بر روی ۶۸۳ مورد و با روش‌های شبکه عصبی و بیزین به‌ترتیب به دقت‌های ۹۷/۵ و ۹۸/۳ دست یافتند(۱۵). همچنین به‌دلیل این‌که ضایعه‌های سرطان پستان در مراحل اولیه قابل دیدن نیستند و به‌صورت ناگهانی ظاهر می‌شود، سیستم‌های تشخیص به کمک کامپیوتر، می‌توانند پزشکان را در شناسایی و کشف این ضایعه‌ها یاری نمایند. بنابراین با استفاده از سیستم‌های تشخیص پزشکی، سرعت تحلیل پزشکان در هر زمان و مکان افزایش می‌یابد و همچنین با توجه به وفور و تداخل متغیرها در تصمیمات پزشکی، پزشکان می‌توانند با به‌کارگیری سیستم‌های هوش مصنوعی، سریع‌تر و یک‌دست‌تر تصمیم‌گیری نمایند و وقت بیشتری را صرف ارزیابی تصمیم نمایند(۱۶). با توجه به جایگاه سرطان پستان در بین زنان ایرانی و با درنظر گرفتن اهمیت زودهنگام سرطان پیش از ابتلا که منجر به کاهش مرگ‌ومیر خواهد شد، هدف این پژوهش، توسعه‌ی مدل بومی تشخیص سرطان پستان با تمرکز بر سه الگوریتم جنگل تصادفی، درخت تصمیم و ماشین بردار پشتیبانی و با تاکید بر تشخیص و انتخاب عوامل موثر بومی و در نهایت انتخاب مدل با بالاترین دقت تشخیص به‌منظور بهبود تصمیم‌گیری پزشکان در امر تشخیص زودهنگام این سرطان می‌باشد. در ادامه این مقاله، با شناسایی عوامل موثر بر ابتلا به سرطان پستان و مجموعه داده‌ی جمع‌آوری‌شده از مرکز تشخیص سرطان پستان، روش‌های ایجاد مدل بررسی می‌گردند. در نهایت نتایج به‌دست آمده از روش‌های مختلف مقایسه و ارزیابی خواهند شد.

روش بررسی

پژوهش حاضر از نظر هدف پژوهش، کاربردی و از لحاظ ماهیت اجرا با بهره‌گیری از روش‌های کیفی و کمی انجام شده است. همان‌طورکه در ادامه مشخص شده است، در گام نخست با استفاده از روش تحلیل محتوا و مطالعات کتابخانه‌ای، عوامل تاثیرگذار در ابتلا به سرطان پستان شناسایی شده سپس با همراهی تیم خبرگان مشتمل بر پزشکان متخصص و فوق‌تخصص سرطان‌شناسی و جراحی پستان و با کمک روش دلفی، عوامل تعدیل گردید. عوامل نهایی بر اساس شرایط بومی و اقلیمی تایید شده و در ادامه با توجه به عوامل نهایی و براساس پرونده پزشکی ۵۲۰۸ بیمار در مرکز تحقیقات سرطان دانشگاه علوم پزشکی و خدمات بهداشتی درمانی شهیدبهشتی، با بهره‌گیری از روش‌های جنگل تصادفی، درخت تصمیم و ماشین‌بردار پشتیبان، احتمال ابتلا به سرطان پستان

محاسبه می‌شود. در ادامه و در بخش یافته‌ها، هرکدام از مراحل بیان شده است.

- **شناسایی عوامل موثر بر ابتلا به سرطان پستان**

جدول ۱: عوامل موثر بر ابتلا به سرطان پستان

ردیف	عوامل موثر بر ابتلا به سرطان پستان	منابع
۱	سن اولین قاعدگی	(۱۵-۱۸)
۲	سن اولین بارداری	(۱۶و۱۸-۲۰)
۳	تعدد بارداری	(۱۶و۱۸و۱۹و۲۱و۲۲)
٤	سن یائسگی	(۱۶و۱۸و۲۳و۲٤)
٥	علت یائسگی(جراحی تخمدان، هورمون‌تراپی)	(۱۶و۱۸و۲۵-۲۷)
٦	سابقه فردی ابتلا به بیماری خوش‌خیم پستان	(۲۰و۲۸-۳۰)
۷	سابقه فردی ابتلا به سرطان تخمدان	(۱۶و۲۵و۳۱)
۸	سابقه فردی ابتلا به سرطان روده	(۳۲و۳۳)
۹	سابقه فردی ابتلا به سایر سرطان‌ها	(۱۹و۳٤)
۱۰	سابقه سرطان پستان در وابستگان درجه اول	(۱۶و۱۷و۱۹و۲۵و۳۵)
۱۱	سابقه سرطان پستان در وابستگان درجه دوم	(۱۶و۲٤و۳۶و۳۷)
۱۲	کشف ژن‌های جهش یافته	(۱۹و۲٤و۲۵و۳۵و۳۸و۳۹)
۱۳	سابقه‌ی تابش یونیزه‌کننده	(۱۶و٤۰)
۱٤	مصرف داروهای ضدباروری	(۱۸و۲۱و۳۰)
۱۵	مصرف داروهای باروری	(۱۱و۱۶و٤۱).
۱۶	سابقه‌ی شیمی‌درمانی	(٤۲و٤۳)
۱۷	شاخص توده بدنی بالا	(۱۵و۲۶و۳۹)
۱۸	بافت متراکم پستان	(۲۵و۳۹و٤٤)
۱۹	عقیم‌سازی(توبکتومی)	(٤۵و٤۶)
۲۰	تعداد زیاد بیوپسی‌های قبلی پستان	(۲۳و٤۷و٤۸)
۲۱	رژیم غذایی نامناسب	(۱۶و۲۱و٤۹و۵۰)
۲۲	مصرف بیش از حد گوشت قرمز	(۲۱و۵۰)
۲۳	نوشیدن الکل	(۱۶و٤۰)
۲٤	مصرف دخانیات	(۱۶و۳۰و۵۱و۵۲)
۲۵	سابقه‌ی تماس با فرد سیگاری در منزل	(۳۰و۵۲)
۲۶	فعالیت جسمانی نامنظم یا بی‌تحرکی	(۱۶و۳۰و۵۳)
۲۷	شغل نامناسب	(٤۰و۵٤)
۲۸	خواب نامناسب	(۵۵-۵۷)
۲۹	مصرف داروی دی‌اتیل استیل بترول	(۵۰و۵۸و۵۹)

تحقیقات بسیاری عوامل موثر بر سرطان پستان را بیان داشته است. بسیاری از عوامل نظیر سن اولین قاعدگی مورد تایید مقالات متعدد قرار گرفته است و این درحالی است که برخی از عوامل تنها در منابع محدودی ذکر شده است. در این بخش از تحقیق سعی شده است، با مطالعه و تحلیل محتوای مقالات متعدد، منابع تاثیرگذار در سرطان پستان استخراج شود(جدول ۱).

- **تعدیل عوامل موثر بر ابتلا به سرطان پستان**

جدول ۲: مشخصات اعضای تیم خبره

ردیف	تخصص	سابقه کار
۱	متخصص زنان و فلوشیپ انکولوژی زنان	۲۶
۲	فلوشیپ فوق‌تخصصی سرطان‌های زنان و متخصص جراحی زنان و زایمان	۲۵
۳	فلوشیپ فوق‌تخصصی سرطان‌شناسی زنان و انکولوژی زنان	۲۵
٤	فوق‌تخصص سرطان‌شناسی زنان، انکولوژی زنان	۲۳
۵	فلوشیپ فوق‌تخصصی سرطان‌شناسی زنان، انکولوژی زنان، متخصص زنان و زایمان و پرتودرمانی	۲۲
٦	متخصص زنان و زایمان، فلوشیپ فوق‌تخصصی سرطان‌شناسی زنان، انکولوژی زنان	۲۲
۷	فلوشیپ فوق‌تخصصی سرطان‌شناسی زنان، انکولوژی زنان، جراح و متخصص زنان و زایمان	۱۷
۸	فوق‌تخصص سرطان‌شناسی زنان، انکولوژی زنان	۱۷
۹	فوق‌تخصص جراحیهای بیماری‌های پستان	۱۷
۱۰	فوق‌تخصص سرطان‌شناسی زنان، انکولوژی زنان	۱۵
۱۱	متخصص زنان و زایمان، فلوشیپ فوق‌تخصصی انکولوژی زنان	۱٤
۱۲	جراح و متخصص زنان و زایمان، فلوشیپ فوق‌تخصصی سرطان‌شناسی زنان، انکولوژی زنان	۱۲
۱۳	فلوشیپ فوق‌تخصصی جراحی سرطان‌های زنان، جراح و متخصص زنان و زایمان و نازایی	۱۰
۱٤	فلوشیپ فوق‌تخصصی سرطان ۲شناسی زنان و انکولوژی زنان و جراح و متخصص زنان و زایمان	۱۰
۱۵	فوق‌تخصص سرطان‌شناسی زنان، انکولوژی زنان	۷
۱٦	فلوشیپ فوق‌تخصصی انکولوژی و سرطان‌شناسی	۷
۱۷	متخصص انکولوژی و رادیوتراپی و پرتودرمانی و شیمی درمانی	۷
۱۸	فلوشیپ سرطان‌شناسی و انکولوژی زنان، جراح و متخصص زنان و زایمان	۵

باتوجه به اینکه جامعه‌ی مورد پژوهش حاضر کشور ایران می‌باشد، ممکن است برخی از عوامل ذکر شده نسبت به شرایط اقلیمی و اجتماعی این کشور وجود نداشته و یا میزان تاثیرگذاری آنها بسیار ناچیز و قابل حذف باشد و همچنین عواملی وجود داشته باشند که در مطالعات پیشین عنوان نشده و یا پژوهش‌گر در مطالعات خود با آنها روبه‌رو نشده است که تمامی این موارد نیاز به تجربه و نظر متخصصان بیماری سرطان پستان در کشور ایران دارد. بنابراین تیم خبرهای شامل ۱۸ پزشک متخصص و فوق‌تخصص جراحی پستان تشکیل شده(جدول ۲) و عوامل مطالعه‌شده از روش تحلیل محتوا، در قالب پرسش‌نامه‌ای باز جهت تایید و یا رد آنها و یا افزودن عاملی جدید، به روش دلفی در اختیار آنها قرار گرفته است.

روش دلفی طی سه مرحله به پایان رسید؛ که در مرحله اول پرسش‌نامه به‌صورت ساختاریافته براساس مطالعات قبلی و بر پایه ۲۹ عامل، طراحی شده و از متخصصان خواسته شد تاثیرگذاری هریک از عوامل مذکور را تایید و یا رد کنند. در دور دوم روش دلفی پرسش‌نامه‌ای طراحی گردید که شامل میزان فراوانی تایید هرعامل حاصل از نظرخواهی اعضای خبره در دور اول و همچنین عوامل استخراج‌شده از پاسخ سوال باز در پرسش‌نامه‌ی اول بوده است. در این مرحله، از خبرگان خواسته شد تا پاسخ‌های خود را با توجه به‌نظر سایر افراد مجدداً تایید و یا درصورت نیاز تغییر دهند و همچنین عوامل جدیدی را که از پاسخ سوال باز

پرسش‌نامه‌ی اولیه به‌دست آمده‌اند، تایید کنند و یا تاثیرگذاری آن را در ابتلا به سرطان پستان رد نمایند. در دور سوم روش دلفی فراوانی‌های تایید در مرحله دوم برای تمام عوامل، دوباره محاسبه گردیده و از آنها درخواست شده است تا درصورت لزوم با تکیه بر تجربه‌ی خود و تجربه‌ی دیگر همکاران که در میزان فراوانی‌ها نهفته است، پاسخ‌های خود را مجدداً تایید نموده و یا تغییر دهند. در روش دلفی، روندهای توزیع پرسش‌نامه تا جایی ادامه خواهند داشت که توافق قابل‌قبولی بین اعضای پنل حاصل شود. در پژوهش حاضر، این روندها شامل سه مرحله بوده و در نتیجه پرسش‌نامه‌ی دور سوم، پرسش‌نامه‌ی نهایی بوده است. میانگین فراوانی تمامی عوامل ۱۳/۷۸ بوده و عواملی که مقدار فراوانی تایید آنها از مقدار میانگین کمتر بوده، حذف شده‌اند که این عوامل عبارتند از: سابقه تماس با فرد سیگاری در منزل، مصرف داروی دی اتیل استیل بترول و مصرف بیش از حد گوشت قرمز.

- **داده‌های سرطان پستان مورد مطالعه**

با توجه به آنکه ۲۲ ویژگی از ۲۶ ویژگی به‌صورت متغیرهای رشته‌ای هستند و همچنین برچسب خروجی مجموعه داده به وضعیت ابتلای بیمار به سرطان اشاره دارد(به سرطان پستان مبتلا شده است/ نشده است) و براساس اینکه ویژگی‌ها قابلیت تفکیک در هر گره را دارند، انتظار می‌رود که روش‌های جنگل تصادفی و درخت تصمیم با توجه به نوع قابلیت دسته‌بندی، نتایج بهتری را نسبت به سایر روش‌ها در بر داشته باشد. اکثر الگوریتم‌های درخت‌های تصمیم، با ساختن یک درخت از بالا به پایین به کمک انتخاب صفات در هر لحظه و جداسازی داده‌ها با توجه به مقادیر صفات‌شان، ایجاد می‌گردند. همچنین ایجاد زیرگره‌ها در هر مرحله، یکنواختی داده‌ها را در زیرگره‌های حاصل، افزایش می‌دهد، به عبارت دیگر خلوص گره در هر مرحله، با توجه به شباهت آن با متغیر هدف، افزایش می‌یابد(۶۰). از ویژگی‌های مهم جنگل تصادفی، عملکرد بالای آن در اندازه‌گیری اهمیت متغیرها برای مشخص کردن این است که هر متغیر چه نقشی در پیش‌بینی پاسخ دارد. قابل ذکر است که عملکرد جنگل تصادفی معمولا بهتر از درخت تصمیم است، اما این بهبود عملکرد تا حدی به نوع داده هم بستگی دارد(۶۱). همچنین از روش ماشین‌بردار پشتیبان که یک جداکننده‌ی دودویی است و در ماکزیمم‌های محلی گیر نمی‌افتد(۶۲)، در این مقاله بهره گرفته‌ایم. برای این منظور از چهار کرنل متفاوت خطی، چندجمله‌ای، تابع پایه شعاعی گاوسی و چندجمله‌ای استفاده کرده‌ایم. نهایتاً برای ارزیابی و کارایی هریک از سه روش، در ادامه بررسی‌های لازم صورت گرفته است.

جدول ۳: ویژگی‌های مجموعه داده سرطان پستان

ویژگی	نوع داده	بازده مقادیر داده‌ها
سن اولین قاعدگی	عددی صحیح	۹-۱۸
سن اولین بارداری	عددی صحیح	۱۲-۴۷
تعدد بارداری	عددی صحیح	۰-۱۳
سن یائسگی	عددی صحیح	۱۳-۵۹
علت یائسگی(جراحی تخمدان، هورمون‌تراپی)	رشته‌ای	Hormonal therapy/ Bilateral oophorectomy
سابقه فردی ابتلا به بیماری خوش‌خیم پستان	رشته‌ای	Yes/ No
سابقه فردی ابتلا به سرطان تخمدان	رشته‌ای	Yes/ No
سابقه فردی ابتلا به سرطان روده	رشته‌ای	Yes/ No
سابقه فردی ابتلا به سایر سرطان‌ها	رشته‌ای	Yes/ No
سابقه سرطان پستان در وابستگان درجه اول	رشته‌ای	Yes/ No
سابقه سرطان پستان در وابستگان درجه دوم	رشته‌ای	Yes/ No
کشف ژن‌های جهش‌یافته	رشته‌ای	Yes/ No
سابقه تابش یونیزه‌کننده	رشته‌ای	Yes/ No
مصرف داروهای ضدباروری	رشته‌ای	Yes/ No
مصرف داروهای باروری	رشته‌ای	Yes/ No
سابقه شیمی‌درمانی	رشته‌ای	Yes/ No

شاخص توده بدنی بالا	رشته‌ای	Yes/ No
بافت متراکم پستان	رشته‌ای	Yes/ No
عقیم‌سازی(توبکتومی)	رشته‌ای	Yes/ No
تعداد زیاد بیوپسی‌های قبلی پستان	رشته‌ای	High/ Low
رژیم غذایی نامناسب	رشته‌ای	Yes/ No
نوشیدن الکل	رشته‌ای	Yes/ No
مصرف دخانیات	رشته‌ای	Yes/ No
فعالیت جسمانی نامنظم یا بی‌تحرکی	رشته‌ای	Yes/ No
شغل نامناسب	رشته‌ای	Yes/ No
خواب نامناسب	رشته‌ای	Yes/ No

داده‌های مورد مطالعه، داده‌های بانوانی است که با مراجعه به مرکز تحقیقات سرطان دانشگاه شهیدبهشتی در طی سال‌های ۱۳۹٤ تا ۱۳۹۷ از میان ۵۲۰۸ نفر استخراج شده است؛ که از بین مراجعه‌کنندگان، ۱٤۲۳ نفر(۲۷/۳۲٪) مبتلا به سرطان پستان بوده و ۳۷۸۵ نفر(۷۲/٦۸٪) مبتلا نبوده‌اند. همچنین با توجه به معیارهای استخراج‌شده، ۲٦ ویژگی هر فرد درنظر گرفته شده است(جدول ۳).

- **داده‌های نرمال‌شده و استانداردشده**

اهمیت آماده‌سازی داده‌ها به‌دلیل این واقعیت است که فقدان داده با کیفیت برابر با فقدان کیفیت در نتایج کاوش است و ورودی بد، خروجی بد به‌دنبال دارد. وظیفه‌ی اصلی پیش‌پردازش داده‌ها، سازمان‌دهی داده‌ها در شکل‌های استاندارد برای داده‌کاوی است(٦۳).

جدول ۴: میانگین و انحراف‌معیار ویژگی‌های صحیح

ویژگی	سن اولین قاعدگی	سن اولین بارداری	تعداد بارداری	سن یائسگی
میانگین	۱۳/۲۰۸۱٤۱	۲۱/۸٤۸۹۱۷	۲/۲٦۸۲٤۱	٤۷/۵۲۹۸۷۲
انحراف‌معیار	۱/٤٦۱٦۵۲	۵/۰۰۳۸۹۳	۱/۹۵۲۷۸۸	۵/۱۱٤٦۷۱

در این پژوهش ٤ ویژگی«سن اولین قاعدگی»، «سن اولین بارداری»، «تعداد بارداری» و«سن یائسگی» که به‌صورت عدد صحیح هستند، کاملا مشخص است که میزان پراکندگی و مقیاس متغیرها در این ویژگی‌ها، متفاوت هستند(جدول ٤). بنابراین احتیاج به استانداردسازی و نرمال‌سازی است.

یافته‌ها

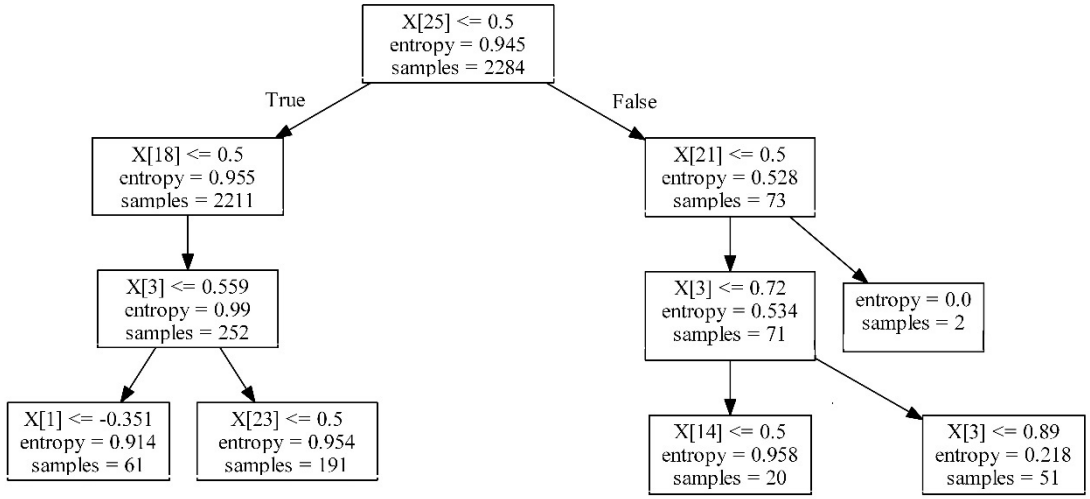
در این مطالعه، مجموعه داده‌ی مراجعه‌کنندگان به مرکز تشخیص سرطان جمع‌آوری گردید و عملیات پیش‌پردازش بر روی نمونه‌ها انجام گرفت. در ادامه سه الگوریتم طبقه‌بندی برای ایجاد و ارزیابی مدل بر روی مجموعه داده‌ها انتخاب گردیدند. تجزیه‌وتحلیل سه الگوریتم با مجموعه داده‌ی آزمون شامل ۳۰٪ از کل داده‌ها و ۷۰٪ از کل داده‌ها برای داده‌ی آموزشی اجرا شده است. نتایج به‌دست آمده از سه تکنیک درخت تصمیم، جنگل تصادفی و ماشین‌بردار پشتیبان با استفاده از چهار معیار دقت (Accuracy)، صحت (Precision)، حساسیت (Sensitivity) و معیار اف (F–measure) ارزیابی شد. در ارزیابی‌های صورت‌گرفته، هر سه مدل، نتایج امیدبخشی را در تشخیص بیماری سرطان پستان نشان دادند.

جدول ۵: ارزیابی مدل‌های تشخیص سرطان پستان

روش	دقت	صحت	حساسیت	معیار اف
درخت تصمیم	٪۹۲/۵۰	٪۹٦/٦٤	٪۹۵/۰۱	٪۹۵/۲۵
جنگل تصادفی	٪۹٤/۷۵	٪۹۷/۲٦	٪۹۵/۱۹	٪۹٦/۲۹
ماشین‌بردار پشتیبان	٪۹۱/۱٦	٪۹۵/۵۹	٪۹۳/٤۳	٪۹٤/۱٤

جدول ۵، نتایج ارزیابی سه مدل مذکور را نشان می‌دهد. طبق نتایج حاصل‌شده، جنگل تصادفی با دقت تشخیص ۹۴/۷۵٪، یک مدل با دقت قابل‌قبول به‌منظور استفاده در فرایند تشخیص ابتلا به سرطان پستان نسبت به دو روش درخت تصمیم و ماشین‌بردار پشتیبان ارایه داده است.

در شکل ۱ ساختار درخت تصمیم‌گیری با عمق ٤ قابل مشاهده است. صحت قوانین با خبرگان بررسی و تحلیل گردید. یکی از قوانین برگرفته از جنگل تصادفی بدین صورت است که زنانی که باردار نشده‌اند، دارای سابقه فردی ابتلا به بیماری خوش‌خیم پستان هستند، سن یائسگی آنها کمتر از ٤٠ سال و سن منارک ایشان بالای ١٤ سال است، بیش از سایر زنان در معرض ابتلا به سرطان پستان قرار دارند.



شکل ۱: ساختار درخت تصمیم‌گیری

بحث

بیماری سرطان پستان واقعیتی غیرقابل انکار است که زمانی‌که اتفاق بیفتد و پیشرفت داشته باشد، درمان آن به مراتب بسیار سخت‌تر و گاه غیرقابل کنترل می‌باشد. باتوجه به پیامدهای غیرقابل جبران بیماری سرطان پستان، مبحث پیشگیری از آن که در رأس آن شناسایی عوامل تاثیرگذار، میزان اهمیت و سطح تاثیرگذاری آنها بریکدیگر و برابتلا به سرطان پستان می‌باشد، توجه محقق را به خود جلب نموده است. ازاین‌رو در این تحقیق ابتدا به شناسایی عوامل تاثیرگذار در ابتلا به سرطان پستان به‌ویژه در ایران پرداختیم. این امر در گام نخست با بررسی مطالعات پیشین و مرتبط به روش تحلیل محتوا صورت پذیرفته و در گام بعدی عوامل به‌دست آمده در قالب پرسش‌نامه‌ای باز جهت تایید و یا رد تاثیرگذاری عوامل بر سرطان پستان و نیز ارایه عوامل جدید باتوجه به تجربه خبرگان، در اختیار تیم خبره قرار گرفته است. در ادامه و در گام نهایی با توجه به معیارهای اثرگذار بر سرطان پستان، و با استفاده از روش‌های درخت تصمیم، جنگل تصادفی و نیز ماشین‌بردار پشتیبان که در زمره روش‌های یادگیری ماشین به شمار می‌آیند و با درنظر گرفتن ٢٦ عامل موثر بر سرطان پستان که از مراحل قبل استخراج شده‌اند، احتمال ابتلا به سرطان پستان تخمین زده شد. در راستای بهبود تشخیص، داده‌ها نرمال و سپس استاندارد شدند، که میزان تشخیص بهبود داده شد و با روش جنگل تصادفی دقت ۹٤/۷۵٪ به‌دست آمد.

تاکنون پژوهش‌های داخلی متفاوتی در زمینه‌ی سرطان پستان صورت گرفته است. لطیف و همکاران با مطالعه بر روی ٥٧٤ بیمار مبتلا به سرطان پستان در بیمارستان فوق‌تخصصی مرتاض یزد با ٣٢ ویژگی و بهره‌گیری از روش‌های درخت تصمیم، بیزین و نزدیک‌ترین همسایه به‌ترتیب به دقت‌های ۹۲/۹٪، ۹۱/۳٪ و ۹٥/۱٪ دست یافتند(٦٤). محمودی و همکاران در تحقیقی مشابه بر روی داده‌های اندازه تومور، درگیری غدد لنفاوی و متاستاز در ٧٣٢ بیمار مبتلا به سرطان پستان در بیمارستان ولی‌عصر بیرجند، از ٤ روش درخت تصمیم، شبکه عصبی، شبکه بیزین و نزدیک‌ترین همسایه به دقت ۹٦٪ با روش نزدیک‌ترین همسایه دست‌یافتند. همچنین دقت روش درخت تصمیم با داده‌ها و ویژگی‌های مذکور ۹۳/٤٪ حاصل شد(٦٥). گنجی و آباده از الگوریتم مورچگان برای تشخیص خوش‌خیم یا بدخیم بودن سرطان پستان استفاده کردند و به دقت ۹٥٪ رسیدند(٦٦). طلوعی اشلقی و همکاران در پژوهشی مشابه و با مطالعه بر

روی ۱۱۸۹ بیمار مبتلا به سرطان پستان با ۳۰ ویژگی با روش درخت تصمیم به دقت ۹۳/٦٪، با روش ماشین‌بردار پشتیبان به دقت ۹٤/۷٪ و با نهایتاً با روش شبکه عصبی به‌دقت ۹۵/۷٪ دست یافتند(٦۷). یزدانی و همکاران با مطالعه بر روی ٦۵٤ پرونده‌ی در دسترس از بیماران کلینیک تخصصی سرطان پستان مطهری شیراز با ۱۰ ویژگی، و با بهره‌گیری از روش‌های درخت تصمیم، بیزین و شبکه عصبی به‌عنوان روش‌های یادگیری ماشین به‌ترتیب به دقت‌های ۹۳/۵۰٪، ۹۰/٦۷٪ و ۹٤/٤۹٪ دست یافتند(٦۸). قیومی‌زاده برای تشخیص خوش‌خیم یا بدخیم بودن سرطان پستان از ترکیب شبکه عصبی خودسازمان‌ده و شبکه‌ی پرسپترون چندلایه استفاده کردند(٦۹). کیانی و آتشی با مطالعه بر روی داده‌های ۸۰۹ بیمار و با استفاده از الگوریتم درخت تصمیم‌گیری و با بهره‌گیری از ۳۲ ویژگی به دقت ۸۵٪ دست یافتند(۷۰).

نتیجه‌گیری

در این تحقیق با استفاده از روش‌های یادگیری ماشینی در زمینه تشخیص احتمال ابتلا به سرطان پستان و با بهره‌مندی از نظر و تجربه‌ی زنده‌ی متخصصان حاذق در زمینه‌ی سرطان‌شناسی برای شناسایی عوامل تاثیرگذار در ابتلا به سرطان پستان، نتایج مطلوبی به‌دست آمد. ۲٦ عامل تاثیرگذار بر سرطان پستان استخراج شده در این پژوهش علاوه بر ویژگی‌های آناتومی، سایر بخش‌های بیماری‌های فردی، شیوه‌ی تغذیه، داروهای خوراکی، وضعیت بارداری، عوامل شخصی و اجتماعی، سوابق درمانی و سوابق خانوادگی را نیز شامل می‌شود که از طریق تحلیل محتوا استخراج شده‌اند و از تحلیل خبرگی جهت بومی‌سازی بهره گرفته شده است؛ همچنین مجموعه داده‌ی مورد مطالعه شامل ۵۲۰۸ مراجعه‌کننده‌ی بومی است که در طی ۳ سال گردآوری شده و در آموزش سیستم خبره موثرتر واقع خواهد شد. در این پژوهش از سه روش درخت تصمیم، جنگل تصادفی و ماشین‌بردار پشتیبان جهت تشخیص احتمال ابتلا به سرطان پستان استفاده شده است. با استفاده از روش جنگل تصادفی با دقت ۹٤/۷۵٪ و صحت ۹۷/۲٦٪ بهترین نتیجه نسبت به دو روش دیگر حاصل شد. همچنین استفاده از شبکه عصبی مصنوعی به‌دلیل آنکه توانایی پیش‌بینی و دسته‌بندی اطلاعات را داشت و قادر به کشف روابط موجود بین داده‌ها با کمترین خطا بود، در تحقیقات آتی پیشنهاد می‌گردد. در مقایسه با سایر پژوهش‌های مشابه که از پایگاه‌های داده بومی بهره گرفته‌اند، دقت‌های به‌دست آمده بسیار نزدیک به کارهای پیشین بوده و در بعضی موارد نیز دقت بهتری داشته است. همچنین این پژوهش در مقایسه با سایر پژوهش‌ها که از پایگاه داده‌ی استاندارد نظیر WBCD استفاده کرده‌اند، در بعضی موارد دقت کمتری داشته است که می‌توان دلیل اصلی آن را پژوهش‌های متعدد بر روی یک پایگاه داده و استانداردسازی متعدد داده‌ها دانست. با توجه به اینکه به‌منظور ساخت مدل در این مطالعه از داده‌های مرکز تشخیص سرطان پستان دانشگاه شهیدبهشتی بهره گرفته شده است، به‌منظور تشخیص ابتلا به سرطان پستان در بیماران این مرکز، استفاده از این مدل نسبت به سایر مدل‌های ساخته شده بر روی سایر مجموعه داده‌ها، می‌تواند نتایج بهتری داشته باشد. با توجه به اینکه این مدل بر روی مجموعه داده‌های بومی ایجاد گردیده است می‌توان نقش مجموعه داده‌های داخلی را در جهت بهبود فرایند تشخیص بیماری‌های بومی ارزیابی نمود؛ اما با توجه به در دسترس نبودن مدل‌های دیگر، از این موضوع به‌عنوان محدودیت مطالعه صورت‌گرفته می‌توان یاد کرد.

براساس ارزیابی صورت‌گرفته در این مقاله، سه مدل ایجاد شده به‌ویژه جنگل تصادفی، بر روی مجموعه داده‌های جمع‌آوری‌شده، قابلیت تشخیص ابتلا به سرطان پستان را به‌خوبی فراهم می‌کند و عواملی نظیر سن یائسگی، سن منارک، تعداد بارداری، سابقه‌ی فردی ابتلا به بیماری خوش‌خیم پستان، کشف ژن‌های جهش‌یافته و سابقه‌ی سرطان پستان در وابستگان درجه اول را می‌توان مهم‌ترین عوامل دانست. همچنین از دیگر عوامل خطرزا، در این پژوهش رژیم غذایی نامناسب، فعالیت جسمانی نامنظم یا بی‌تحرکی و خواب نامناسب است و پیشنهاد می‌گردد که در برنامه‌ریزی‌های پیشگیری از ابتلا به سرطان پستان نسبت به عوامل ذکر شده توجه ویژه شده و در سبک زندگی افراد جهت پیشگیری، اقدامات لازم اجرا شود.

تشکر و قدردانی

این مقاله، حاصل بخشی از پایان‌نامه با عنوان«طراحی سیستم خبره تشخیص ریسک ابتلا به سرطان پستان» در مقطع دکتری‌تخصصی مصوب دانشگاه تهران در سال ۱۳۹٦ است. در این پژوهش تمامی ملاحظات اخلاقی از جمله شرط امانت و صداقت و حفظ محرمانگی اطلاعات هویتی شرکت‌کنندگان رعایت شده است. نویسندگان مراتب سپاس و قدردانی خود را از تمامی مشارکت‌کنندگان در پژوهش اعلام می‌نمایند.

References

1. Sharifian AH, Pourhoseingholi MA, Emadedin M, Rostami Nejad M, Ashtari S, Hajizadeh N, et al. Burden of Breast cancer in Iranian women is increasing. Asian Pacific Journal of Cancer Prevention 2015; 16(12): 5049-52.

2. Sheikhpour R, Agha Saram M, Zare Mirak Abad MR & Sheikhpour R. Breast cancer detection using two-step reduction of features extracted from fine needle aspirate and data mining algorithms. Iranian Quarterly Journal of Breast Disease 2015; 7(4): 43-51[Article in Persian].

3. Hadizadeh M. Pink ribbon, what men and women need to know about Breast diseases. Tehran: Boroj Publications; 2015: 10-3[Book in Persian].

4. Chaurasia V, Pal S & Tiwari BB. Prediction of benign and malignant Breast cancer using data mining techniques. Journal of Algorithms and Computational Technology 2018; 12(2): 119-26.

5. Lu J, Hales A, Rew D, Keech M, Frohlingsdorf C, Mills Mullett A, et al. Data mining techniques in health informatics: A case study from Breast cancer research. Switzerland: Proceedings of the 6th International Conference on Information Technology in Bio- and Medical Informatics, 2015.

6. Ruiz A, Sebagh M, Wicherts DA, Castro Benitez C, Van Hillegersberg R, Paule B, et al. Long-term survival and cure model following liver resection for Breast cancer metastases. Breast Cancer Research and Treatment 2018; 170(1): 89-100.

7. Ojha U & Goel S. A study on prediction of Breast cancer recurrence using data mining techniques, Noida, India: 7th International Conference on Cloud Computing, Data Science and Engineering-Confluence, 2017.

8. Delen D, Walker G & Kadam A. Predicting Breast cancer survivability: A comparison of three data mining methods. Artificial Intelligence in Medicine Journal 2010; 34(2): 113-27.

9. Chou ShM, Lee TSh, E Shao Y & Chen IF. Mining the Breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications 2004; 27(1): 133-42.

10. Ubeyli ED. Implementing automated diagnostic systems for Breast cancer detection. Expert Systems with Applications 2007; 33(4): 1054-62.

11. Karabatak M & Ince MC. An expert system for detection of Breast cancer based on association rules and neural network. Expert Systems with Applications 2009; 36(2): 3465-9.

12. Bhardwaj A & Tiwari A. Breast cancer diagnosis using genetically optimized neural network model. Expert Systems with Applications 2015; 42(10): 4611-20.

13. Bennett KP & Blue JA. A support vector machine approach to decision trees, Anchorage, AK, USA: IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence, 1998.

14. Chao CM, Yu YW, Cheng BW & Kuo YL. Construction the model on the Breast cancer survival analysis use support vector machine, logistic regression and decision tree. Journal of Medical Systems 2014; 38(10): 106.

15. 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): 266-78[Article in Persian].

16. Sadoogh F & Sheikhtaheri A. Applications of artificial intelligence in clinical decision making: Opportunities and challenges. Health Information Management 2011; 8(3): 440-5[Article in Persian].

17. Afshari E & Farahi A. Application of fuzzy expert system in the diagnosis of Breast cancer, Tehran, Malek Ashtar University of Technology: 3rd International Conference on Applied Research in Computer Engineering and Information Technology, 2015.

18. Key TJ, Verkasalo PK & Banks E. Epidemiology of Breast cancer. The Lancet Oncology 2001; 2(3): 133-40.

19. Rosner B & Colditz GA. Nurses' health study: Log-incidence mathematical model of Breast cancer incidence. Journal of the National Cancer Institute 1996; 88(6): 359-64.

20. Anderson KN, Schwab RB & Martinez ME. Reproductive risk factors and Breast cancer subtypes: A review of the literature. Breast Cancer Research and Treatment 2014; 144(1): 1-10.

21. Ferrer J, Neyro, JL & Estevez A. Identification of risk factors for prevention and early diagnosis of a-symptomatic post-menopausal women. The European Menopause Journal (Maturitas) 2005; 52(1): 7-22.

22. Beckmann MV, Bani MR, Fasching PA, Strick R & Lux MP. Risk and risk assessment for Breast cancer: Molecular and Clinical aspects. The European Menopause Journal (Maturitas) 2007; 57(1): 56-60.

23. Porter P. "Westernizing" women's risks? Breast cancer in lower-income countries. The New England Journal of Medicine 2008; 358(3): 213-6.

24. Breyer JZ, Wendland EM, Kops NL, Caleffi M & Hammes LS. Assessment of potential risk factors for Breast cancer in a population in southern Brazil. Breast Cancer Research and Treatment 2018; 169(1): 125-31.

25. Hamsian Etefagh M & Nadimi Shahraki MH. Provide a model to determine the risk of Breast cancer using the EM algorithm in risk factors. Iranian Journal of Breast Disease 2016; 9(1): 21-30[Article in Persian].

26. Sakorafas GH, Krespis E & Pavlakis G. Risk estimation for Breast cancer development; A clinical perspective. Surgical Oncology 2002; 10(4): 183-92.

27. Hulka BS & Moornan PG. Breast cancer: Hormones and other risk factors. The European Menopause Journal (Maturitas) 2001; 38(1): 103-13.

28. Colditz GA & Rosner B. Cumulative risk of Breast cancer to age 70 years according to risk factor status: Data from the nurses' health study. American Journal of Epidemiology 2000; 152(10): 950-64.

29. Ross RK, Paganini Hill A, Wan PC & Pike MC. Effect of hormone replacement therapy on Breast cancer risk: Estrogen versus estrogen plus progestin. Journal of the National Cancer Institute 2000; 92(4): 328-32.

30. Lesser ML, Rosen PP & Kinne DW. Multicentricity and bilaterality in invasive Breast carcinoma. Surgery 1982; 91(2): 234-40.

31. Lee JSY, Grant CS, Donohue JH, Crotty TB, Harmsen WS & Ilstrup DM. Arguments against routine contralateral mastectomy or undirected biopsy for invasive lobular Breast cancer. Surgery 1995; 118(4): 640-8.

32. Roohparvarzadeh N. Prevalence of risk factors for Breast cancer in women (20 to 69 Years old) in Isfahan 2012-2013. Iranian Quarterly Journal of Breast Disease 2014; 7(1): 52-61[Article in Persian].

33. Seyyed Noori T, Zahmatkesh T, Malahi T, Akbari P, Haghi Z & Mohsseni Azad P. Breast cancer risk assessment using the Gail model. Iranian Journal of Breast Diseases 2008; 1(2): 53-7[Article in Persian].

34. Lai JH, Park G & Gerson LB. Association between Breast cancer and the risk of colorectal cancer. Gastrointestinal Endoscopy 2017; 86(3): 429-41.

35. Muller A, Edmonston TB, Corao DA, Rose DG, Palazzo JP, Becker H, et al. Exclusion of Breast cancer as an integral tumor of hereditary nonpolyposis colorectal cancer. Cancer Research 2002; 62(4): 1014-9.

36. Buehring GC, Shen HM, Jensen HM, Jin DL, Hudes M & Block G. Exposure to bovine Leukemia virus is associated with Breast cancer: A case-control study. PLoS ONE 2015; 10(9): e0134304.

37. Robson ME, Bradbury AR, Arun B, Domchek SM, Ford JM, Hampel HL, et al. American society of clinical oncology policy statement update: Genetic and genomic testing for cancer susceptibility. Journal of Clinical Oncology 2015; 33(31): 3660-7.

38. Marcus JN, Watson P, Page DL & Lynch HT. Pathology and heredity of Breast cancer in younger women. Journal of the National Cancer Institute, Monographs 1994; 16(1): 23-34.

39. Henderson IC. What can a woman do about her risk of dying of Breast cancer? Current Problems in Cancer 1990; 14(4): 161-230.

40. Thorlacius S, Sigurdsson S, Bjarnadottir H, Olafsdottir G, Jonasson JG, Tryggvadottir L, et al. Study of a single BRCA2 mutation with high carrier frequency in a small population. American Journal of Human Genetics 1997; 60(5): 1079-84.

41. Buyukavcu A, Albayrak YE & Goker N. A fuzzy information-based approach for Breast cancer risk factors assessment. Applied Soft Computing 2016; 38(1): 437-52.

42. Hiatt RA & Green Brody J. Environmental determinants of Breast cancer. Annual Review of Public Health 2018; 39(1): 113-33.

43. Reigstad MM, Storeng R, Myklebust TA, Oldereid NB, Omland AK, Robsahm TE, et al. Cancer risk in women treated with fertility drugs according to parity status-a registry-based cohort study. Cancer Epidemiology, Biomarkers and Prevention 2017; 26(6): 953-62.

44. Huszno J, Budryk M, Kołosza Z & Nowara E. The risk factors of toxicity during chemotherapy and radiotherapy in Breast cancer patients according to the presence of BRCA gene mutation. Contemporary Oncology, Wspolczesna Onkologia 2015; 19(1): 72-6.

45. Chaudhuri A, Sinha D, Bhattacharya K & Das A. An integrated strategy for data mining based on identifying important and contradicting variables for Breast cancer recurrence research. International Journal of Recent Technology and Engineering 2020; 8(6): 1096-106.

46. Dembrower K, Liu Y, Azizpour H, Eklund M, Smith K, Lindholm P, et al. Comparison of a deep learning risk score and standard mammographic density score for Breast cancer risk prediction. Radiology 2020; 294(2): 265-72.

47. Gaudet MM, Patel A, Sun J, Teras L & Gapstur S. Tubal sterilization and Breast cancer incidence: Results from the cancer prevention study ii nutrition cohort and meta-analysis. American Journal of Epidemiology 2013; 177(6): 492-9.

48. Calle EE, Rodriguez C, Walker KA, Wingo PA, Petrelli JM & Thun MJ. Tubal sterilization and risk of Breast cancer mortality in US women. Cancer Causes and Control 2001; 12(2): 127-35.

49. Kerlikowske K, Gard CC, Tice JA, Ziv E, Cummings SR & Miglioretti DL. Risk factors that increase risk of estrogen receptor-positive and negative Breast cancer. Journal of the National Cancer Institute 2016; 109(5): 1-9.

50. Visscher DW, Frank RD, Carter JM, Vierkant RA, Winham SJ, Heinzen EP, et al. Breast cancer risk and progressive histology in serial benign biopsies. Journal of National Cancer Institute 2017; 109(10): 1-7.

51. Hunter DJ & Berman PC. Public health management: Time for a new start? European Journal of Public Health 1997; 7(3): 345-9.

52. Alberg AJ, Lam AP & Helzlsouer KJ. Epidemiology, prevention, and early detection of Breast cancer. Current Opinions in Oncology 1999; 11(6): 435-41.

53. Gaudet MM, Gapstur SM, Sun J, Diver WR, Hannan LM & Thun MJ. Active smoking and Breast cancer risk: Original cohort data and meta-analysis. Journal of the National Cancer Institute 2013; 105(8): 515-25.

54. Lagiou A & Lagiou P. Tobacco smoking and Breast cancer. A life course approach. European Journal of Epidemiology 2017; 32(8): 631-4.

55. Sanchis Gomar F, Lucia A, Yvert T, Ruiz Casado A, Pareja Galeano H, Santos Lozano A, et al. Physical inactivity and low fitness deserve more attention to alter cancer risk and prognosis. Cancer Prevention Research 2015; 8(2): 105-10.

56. Kolstad HA. Nightshift work and risk of Breast cancer and other cancers--a critical review of the epidemiologic evidence. Scandinavian Journal of Work, Environment and Health 2008; 34(1): 5-22.

57. Chen Y, Tan F, Wei L, Li X, Lyu Z, Feng X, et al. Sleep duration and the risk of cancer: A systematic review and meta-analysis including dose-response relationship. BMC Cancer 2018; 18(1149): 1-13.

58. Zhao H, Yin JY, Yang WS, Qin Q, Li TT, Shi Y, et al. Sleep duration and cancer risk: A systematic review and meta-analysis of prospective studies. Asian Pacific Organization for Cancer Prevention 2013; 14(12): 7509-15.

59. Palmer JR, Wise LA, Hatch EE, Troisi R, Titus Ernstoff L, Strohsnitter W, et al. Prenatal diethylstilbestrol exposure and risk of Breast cancer. Cancer Epidemiology, Biomarkers and Prevention 2006; 15(8): 1509-14.

60. Breiman L. Random forests. Machine Learning 2001; 45(1): 5-32.

61. Hastie T, Tibshirani R & Friedman J. The elements of statistical learning: Data mining, inference, and prediction. 2nd ed. USA: Springer; 2009: 144-7.

62. Noble WS. What is a support vector machine? Nature Biotechnology 2006; 24(1): 1565-7.

63. Blake RH & Mangiameli P. The effects and interactions of data quality and problem complexity on data mining, MA, USA: Proceedings of the 13th International Conference on Information Quality, MIT, Cambridge, 2008.

64. Latif AM, Momeny M, Agha Sarram M, Pour Ahmadi A & Haj Ebrahimi Z. Using data mining and genetic algorithm for diagnosis of Breast cancer. Iranian Quarterly Journal of Breast Disease 2016; 9(1): 45-56[Article in Persian].

65. Mahmoodi MS, Mahmoodi SA, Haghighi F & Mahmoodi SM. Determining the stage of Breast cancer by data mining algorithms. Iranian Quarterly Journal of Breast Diseases 2014; 7(2): 36-44[Article in Persian].

66. Ganji MF & Abadeh MS. Parallel fuzzy rule learning using an ACO-based algorithm for medical data mining, Changsha, China: IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA), 2010.

67. Toloiee Ashlaghi A, Pourebrahimi A, Ebrahimi M & Ghasemahmad L. Using data mining techniques for prediction Breast cancer recurrence. Iranian Journal of Breast Diseases 2013; 5(4): 23-34[Article in Persian].

68. Yazdani A, Safaei AA, Safdari R & Zahmatkeshan M. Diagnosis of Breast cancer using decision tree, artificial neural network and naive bayes to provide a native model for Fars province. Journal of Payavard Salamat 2019; 13(3): 241-50[Article in Persian].

69. 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[Article in Persian].

70. Kiani B & Atashi A. A prognostic model based on data mining techniques to predict Breast cancer recurrence. Journal of Health and Biomedical Informatics Medical Informatics 2014; 1(1): 26-31[Article in Persian].

Journal of Payavard Salamat
Vol. 15, No. 4, Oct & Nov 2021

Diagnosing Breast Cancer by Machine Learning

Kasra Dolatkhahi[1] (M.S.), Adel Azar[2] (Ph.D.), Tooraj Karimi[3*] (Ph.D.), Mohammad Hadizadeh[4] (Ph.D.)

1 Ph.D. Candidate in Industrial Management Operations Research Orientation, Faculty of Management and Accounting, College of Farabi University of Tehran, Tehran, Iran
2 Professor, Department of Industrial Management, Faculty of Management and Economics, Tarbiat Modares University, Tehran, Iran
3 Assistant Professor, Department of Industrial and Technology Management, Faculty of Management and Accounting, College of Farabi University of Tehran, Iran
4 Assistant Professor, Cancer Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran

Abstract

Received: Jun 2021
Accepted: Sep 2021

Background and Aim: Cancer and in particular Breast cancer are among the diseases that have the highest mortality rate in Iran after heart disease. The accurate prognosis for Breast cancer is important, and the presence of various symptoms and features of this disease makes it difficult for doctors to diagnose. This study aimed to identify the factors affecting Breast cancer, modeling and ultimately diagnosing the risk of Breast cancer.

Materials and Methods: In the present study, first, by content analysis and library studies, the effective factors in Breast cancer were identified, then with the help of a team of experts consisting of physicians and subspecialists in Breast oncology and Breast surgery; With the help of the Delphi method, the factors were adjusted and 26 final factors that were numerically correct and string based on local and climatic conditions were approved. Then, according to the final factors and based on the medical records of 5208 patients in the Cancer Research Center of Shahid Beheshti University of medical sciences, to diagnose cancer, Decision Tree, Random Forest, and Support Vector Machine methods were used as machine learning methods.

Results: In the first step, by content analysis method, 29 effective factors in Breast cancer were identified. Then, taking into account the indigenous and climatic conditions and using the Delphi method and also using the opinions of 18 Experts during three years, 26 factors were finalized. In the final step, using the medical records of the patients and the results obtained from the three methods mentioned, random forest, had the highest accuracy of 94.75% and precision of 97.26% in diagnosing Breast cancer. It has been noted that, compared to other similar studies, indigenous databases have been exploited, the accuracy obtained has been very close to previous studies, and in many cases much better.

Conclusion: Using the random forest method and taking advantage of the factors affecting Breast cancer, the ability to diagnose cancer has been provided with greatest accuracy.

*** Corresponding Author:**
Karimi T
Email:
tkarimi@ut.ac.ir

Keywords: Breast Cancer, Content Analysis, Delphi Method, Random Forest, Decision Tree, Support Vector Machine