دوفصلنامه محاسبات و سامانه‌های توزیع شده
سال هفتم، شماره ۱، شماره پیاپی ۱۳، صفحه ۵۱ تا ۶۲، سال ۱۴۰۳

مروری بر مفاهیم تکنیک‌های پیش‌بینی زودهنگام سرطان سینه و ارزیابی این تکنیک‌ها بر اساس معیارهای مناسب

فرشید وظیفه دوست[۱]، سمیه کدخدا ده‌خانی[۲] ، مائده رحمانی[۳] ، مهدی قاسمی[۴] و حمید زنگی آبادی زاده[۵]

[۱] فارغ التحصیل مقطع کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز از دانشگاه پیام نور مرکز بین الملل قشم
Vazifehdoostfarshid@gmail.com

[۲]فارغ التحصیل مقطع کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز از دانشگاه پیام نور مرکز بین الملل قشم
Emailsk65@gmail.com

[۳]دانشجوی مقطع کارشناسی ارشد مهندسی کامپیوتر گرایش نرم افزار ،دانشگاه پیام نور مرکز بین الملل کیش
Maede9708@gmail.com

[۴] دانشجوی مقطع کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز ،دانشگاه پیام نور مرکز بین الملل کیش
Mahdikmg1@gmail.com

[۵]دانشجوی مقطع کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز ،دانشگاه پیام نور مرکز بین الملل کیش
Hamid.zangiabadi@gmail.com

چکیده

بروز سرطان سینه در طول سال‌ها به دلیل تغییر در شیوه زندگی و محیط به طور پیوسته افزایش یافته است. در حال حاضر، سرطان سینه یکی از علل اصلی مرگ و میر ناشی از سرطان در میان زنان است که آن را به یک نگرانی حیاتی برای سلامت عمومی جهانی تبدیل کرده است. بنابراین، ایجاد یک سیستم تشخیص خودکار سرطان سینه در جامعه پزشکی اهمیت بالایی دارد. در این مقاله به مفاهیم عمومی سرطان سینه، داده کاوی و یادگیری ماشین و همچنین به معرفی تکنیک‌های مفید یادگیری ماشین که برای طبقه‌بندی و پیش بینی سرطان سینه می‌تواند مفید باشد پرداخته شد. در بخش اصلی مقاله ارزیابی و مقایسه این تکنیک‌ها بر اساس معیارهای مناسب دقت، صحت، بازخوانی، خاصیت و امتیاز F_1 مرور و انجام شد. نتایج ارزیابی این طبقه‌بندها از تحقیقات مناسب که آزمایش‌های آنها روی مجموعه داده‌های استاندارد انجام شده، نشان داد که از بین تکنیک‌های معرفی شده ماشین بردار پشتیبان، جنگل تصادفی و درخت تصمیم به ترتیب به نسبت بقیه دارای پیش‌بینی و طبقه-بندی دارد.

کلمات کلیدی: یادگیری ماشین، طبقه‌بندی، داده‌ها ، سرطان سینه و هوش مصنوعی.

تاریخچه‌ی مقاله:

تاریخ ارسال: ۱۴۰۳/۰۱/۱۸

تاریخ اصلاحات: ۱۴۰۳/۰۶/۰۱

تاریخ پذیرش: ۱۴۰۳/۰۶/۲۱

تاریخ انتشار: ۱۴۰۳/۰۶/۳۰

ایمیل نویسنده مسئول: Vazifehdoostfarshid@gmail.com

۱ – مقدمه

سرطان سینه شایع ترین بدخیمی تشخیص داده شده و کشنده در زنان در سراسر جهان است. مشابه با داده‌های جهانی، سرطان سینه شایع ترین سرطان زنان در ترکیه است[۱]. سرطان سینه از بافت اپیتلیال[1] پستان شروع می‌شود و تقریباً ۹۹ درصد موارد در زنان رخ می‌دهد. آژانس بین المللی تحقیقات سرطان گزارش داد که سرطان سینه زنان دومین سرطان شایع در سراسر جهان و چهارمین علت مرگ و میر ناشی از سرطان است که ۲۳.۸ درصد موارد جدید و ۱۵.۴ درصد از مرگ و میر زنان را در سال ۲۰۲۲ به خود اختصاص داده است. بروز سرطان سینه تنوع جغرافیایی قابل توجهی را نشان داد که از ۲۶.۷ در هر ۱۰۰۰۰۰ زن در جنوب آسیای مرکزی و آفریقای میانه تا ۱۰۰.۳ در هر ۱۰۰۰۰۰ زن در استرالیا و نیوزیلند متغیر بود. علیرغم افزایش بروز در طول زمان، کاهش قابل توجهی در مرگ و میر سرطان پستان در بسیاری از کشورها نشان داده شده است که به رشد در شرایط اجتماعی-اقتصادی، اجرای غربالگری ماموگرافی و استفاده از درمان نئوادجوانت[2] نسبت داده می-شود[۲].

[1] Epithelium

[2] neoadjuvant

عوامل پیش آگهی و پیش‌بینی کننده برای تعیین تاریخچه طبیعی تومور، پیش‌بینی بقا و هدایت درمان استفاده می شود[۱].

علیرغم پیشرفت قابل توجهی که در روش های تشخیص اولیه و درمان به دست آمده است، سرطان سینه یک نگرانی مهم بهداشتی جهانی است که هم زنان و هم مردان را تحت تاثیر قرار می‌دهد. این بیماری شایع‌ترین بیماری بدخیم در بین زنان در سراسر جهان است و دومین عامل علت مرگ و میر ناشی از سرطان است. ناهمگونی و پیچیدگی در مدیریت سرطان سینه توسط زیرگروه های متعدد آن، که رفتارهای بیولوژیکی متفاوتی را نشان می دهند و پاسخ بالینی به درمان را نشان می‌دهند، تاکید می شود . علیرغم پیشرفت‌های اساسی در تشخیص اولیه و استراتژی‌های درمانی سال ، تفاوت‌های قابل توجهی در بقا بین کشورهای پردرآمد و کشورهای با درآمد کم تا متوسط وجود دارد[۳]. سرطان سینه زمانی شناسایی می شود که یک توده بدون درد یا ضخیم شدن پستان در بدن انسان قابل تشخیص نباشد. بهترین راه برای تشخیص سرطان سینه غربالگری زودهنگام یا ماموگرافی پستان است[۴].

دو الگوریتم‌های یادگیری ماشین محدودیت‌های انسانی را حذف می‌کنند و دقت بیشتری در تشخیص بیماری‌هایی مانند سرطان ارائه می‌دهند. سرطان سینه، دومین سرطان تشخیص داده شده در زنان، اغلب به ماموگرافی متکی است که تنها ۷۰ درصد دقیق است و منجر به تشخیص اشتباه احتمالی می‌شود. نمونه‌برداری‌ها، اگرچه قابل اعتمادتر هستند، اما در معرض خطای انسانی و نظرات متخصصان متضاد هستند، که اغلب به بیوپسی‌های متعدد نیاز دارند. کمبود پاتولوژیست تشخیص دقیق و به موقع را پیچیده‌تر می کند. یادگیری ماشینی می‌تواند این خطاها را کاهش دهد و نتایج سریعتر و دقیق‌تری ارائه دهد.

۲ – سرطان سینه

سرطان سینه علت اصلی مرگ و میر ناشی از سرطان در میان زنان است و از نظر مرگ و میر کلی سرطان در رتبه پنجم قرار دارد. طبق آمار جهانی سرطان ۲۰۲۰ که توسط آژانس بین‌المللی تحقیقات سرطان سازمان جهانی بهداشت منتشر شده است، سرطان سینه زنان اکنون به عنوان شایع‌ترین سرطان تشخیص داده شده در سراسر جهان از سرطان ریه پیشی گرفته است. برآورد ۲.۳ میلیون مورد جدید نشان می‌دهد که از هر ۸ سرطان تشخیص داده شده در سال ۲۰۲۰، ۱ مورد سرطان سینه است. این بیماری با ۶۸۵۰۰۰ مرگ در سال ۲۰۲۰، پنجمین علت مرگ و میر سرطان در سراسر جهان است. در زنان، سرطان سینه از هر ۴ مورد سرطان، ۱ مورد و از هر ۶ مرگ سرطان ۱ مورد را به خود اختصاص می‌دهد و این بیماری از نظر بروز و مرگ و میر در اکثر کشورهای جهان (به ترتیب در ۱۵۹ و ۱۱۰ کشور) رتبه اول را دارد. علیرغم افزایش میزان بروز و مرگ و میر سرطان پستان، علل و پاتوژنز[3] آن به خوبی شناخته نشده است. شواهد کنونی نشان می‌دهد که سرطان سینه ممکن است تحت تأثیر محیط داخل رحمی قرار گیرد، قرار گرفتن در معرض در دوران نوجوانی از اهمیت ویژه ای برخوردار است، و اینکه بارداری تأثیری دوگانه افزایش زودهنگام و به دنبال آن محافظت طولانی مدت بر خطر سرطان سینه دارد. تنوع زیادی در رشد ساختاری سیستم مجرای پستان در نوزاد تازه متولد شده وجود دارد - و با استنباط داخل رحمی - و بارداری باعث تغییرات ساختاری دائمی در غده پستانی می شود. با تداوم این عوامل خطر تجمعی، احتمال ابتلا به سرطان سینه افزایش می‌یابد. در نتیجه غربالگری منظم، تشخیص دقیق و درمان به موقع و موثر نقش مهمی در پیشگیری از سرطان سینه ایفا می کند و به عنوان ابزار مهمی برای حفظ جان و سلامت زنان عمل می‌کند[۵].

۳- داده‌کاوی

داده‌کاوی روشی است که از مجموعه‌ای از ابزارهای تجزیه و تحلیل اطلاعات برای یافتن روابط و الگوهای موجود در اطلاعات استفاده می‌کند که ممکن است برای پیش‌بینی‌های قانونی مورد استفاده قرار گیرد. اولین و همچنین ساده‌ترین مرحله تحلیلی داده‌کاوی، توصیف اطلاعات - خلاصه کردن ویژگی‌های آماری آن (مانند انحراف استاندارد) و ابزار، بررسی بصری آن با استفاده از نمودارها و همچنین جستجوی یک لینک‌های قابل توجه احتمالی در بین متغیرها است. در داده کاوی، انتخاب، کاوش، جمع‌آوری و پردازش اطلاعات مناسب بسیار مهم است و کشف تخصصی محاسبات هوشمند را در محصول نهایی مورد نظر خود و همچنین جذاب فناوری اطلاعات نشان می‌دهد. برای داشتن توانایی کشف و همچنین استخراج دانش از اطلاعات، فرآیندی است که بسیاری از دست اندرکاران و محققین برای دستیابی به آن تلاش می‌کنند. مقدار

[3] Pathogenesis

زیادی از دانش پنهان در انتظار یافتن است، این چالشی است که با فراوانی اطلاعات امروزی ایجاد می‌شود. کشف داده در دتاست روشی برای تعیین الگوهای قانونی، مفید، جدید و همچنین قابل درک از مجموعه داده‌های بزرگ است[۶].

جامعه عصبی مصنوعیANN، ساده‌سازی شبکه‌های واقعی نورون‌ها است. پارادایم شبکه عصبی که در دهه ۱۹۴۰ آغاز شد، ابزار بسیار مهمی برای یادگیری ارتباط عملکرد ساختار ذهن انسان بود. به دلیل درک ناقص و همچنین پیچیدگی نورون های بیولوژیکی، ساختار متفاوت شبکه عصبی مصنوعی در ادبیات پیدا شده است. جامعه عصبی هدف معمولاً تقلید از نیروی انسانی برای سازگاری با تغییر در شرایط و همچنین شرایط فعلی است. نمایی از زمینه‌های داده‌کاوی در شکل ۱ نمایش داده شده است[۶].

شکل۱: نمایی از زمینه های داده کاوی[۶].

کشف دانش یک فرآیند تعاملی بوده که شامل ایجاد درک دامنه‌ای از کاربرد، انتخاب و ایجاد یک مجموعه داده، پیش پردازش شده برای تبدیل داده‌ها می‌شود. ابزارهای داده‌کاوی به این سؤال پاسخ می‌دهند که به طور سنتی چه میزان زمان طول می‌کشد و پیچیدگی‌های لازم جهت حل داده‌ها چه میزان می‌باشد. پایگاه داده‌ها برای یافتن اطلاعات پیش‌بینی شده توسط سازمان آماده می‌شود. وظایف داده‌کاوی عبارت است از قانون انجمن، الگوها، طبقه‌بندی، پیش‌بینی و خوشه‌بندی . مهمترین اهداف مدل‌سازی طبقه بندی و پیش بینی می‌باشد. شاخه علوم رایانه‌ای که بیشتر به طور فعال و کارآمد در علوم پزشکی مشارکت دارد، هوش مصنوعی است[۷].

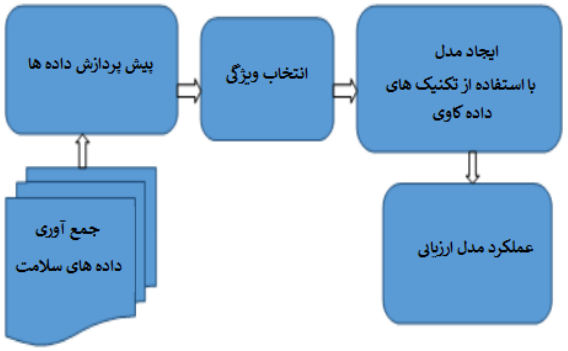
شکل۲: نمایی از معماری داده‌کاوی[۷].

۴- فرایند داده‌کاوی در بهداشت و سلامت

داده‌کاوی می‌تواند از طریق به حداقل رساندن زمان بیمار، ارائه درمان و تشخیص موثر بیماران بر اساس موارد قدیمی‌تر با علائم مشابه، مزایای بیمارستانی و غیره، به بهبود مراقبت‌های بهداشتی به روشی مقرون به صرفه(با روابط بهتر با مصرف کننده تجزیه و تحلیل توالی یا مسیر، طبقه‌بندی، خوشه بندی و پیش‌بینی پارامترهای اصلی داده‌کاوی مورد استفاده در بخش مراقبت‌های بهداشتی) کمک کند. متن‌کاوی و پردازش زبان طبیعی رشته‌های مطالعاتی با طیف گسترده‌ای از کاربردها در زمینه‌های پزشکی، دارویی و علمی به سرعت در حال گسترش هستند. حفظ، جمع‌آوری و به اشتراک‌گذاری دانش رمزگذاری شده در گزارش‌های بیمارستانی، داده‌های بیماری‌های مزمن، توسط مدیریت دانش متنی نظارت می‌شود. در مراقبت‌های بهداشتی امروزی، داده‌کاوی بیشتر برای پیش‌بینی بیماری‌های مختلف، کمک به تشخیص و مشاوره به پزشکان استفاده می‌شود. در تصمیمات بالینی از سوی دیگر، داده‌کاوی دارای پتانسیل بسیار بیشتری است. ممکن است پاسخ‌های مبتنی بر سوال، اکتشافات مبتنی بر ناهنجاری، تصمیم‌گیری‌های آگاهانه‌تر، اندازه‌گیری احتمال، مدل‌سازی پیش‌بینی کننده و کمک تصمیم را ارائه دهد. سوابق الکترونیکی پزشکی، گزارش‌های اداری و سایر اطلاعات محک‌گذاری، همگی در صنعت مراقبت‌های بهداشتی در دسترس هستند که می‌توان با تکنیک‌های پیش‌بینی‌کننده و توصیفی داده‌کاوی برای به دست آوردن نتایج بهتر در معرض دید قرار گرفت[۸].

داده‌های تولید شده در بخش مراقبت‌های بهداشت و سلامت باید به دانش مفید برای تصمیم‌گیری تبدیل شوند. داده‌کاوی امید بزرگی در مراقبت‌های بهداشتی است که پیچیدگی داده‌ها را برای تولید اطلاعات تجزیه و تحلیل کرده و به کشف دانش از

مرحله انتخاب تا مرحله کشف دانش کمک می‌کند. شکل ۳ فرآیند داده‌کاوی در زمینه پزشکی را نمایش می‌دهد[۹].



شکل۳: نمایی از فرآیند داده‌کاوی در زمینه سلامت[۹].

پیش‌پردازش داده‌ها: تکنیکی است که در داده‌کاوی برای تبدیل داده‌های خام به یک قالب مفید و کارآمد استفاده می‌شود. مراحل پیش‌پردازش داده‌ها شامل: پاک‌سازی داده‌ها، ایجاد داده‌ها و کاهش داده‌ها است. داده‌ها می‌توانند چندین بخش داشته باشند که تمیز کردن یا پاکسازی داده برای رسیدگی به داده‌های پر نویز، از دست رفته و غیره انجام می‌شود. تبدیل داده‌ها برای تبدیل داده‌ها به فرم مناسب جهت فرآیند استخراج انجام شده که شامل نرمال‌سازی، جمع‌آوری ویژگی‌ها، گسسته‌سازی و ایجاد سلسله مراتب است. داده‌کاوی روشی است که برای مدیریت مقادیر عظیم داده استفاده می‌شود. در این موارد، هنگام کار با حجم عظیمی از داده‌ها ، تجزیه و تحلیل دشوارتر می‌شود از این رو برای خلاص شدن از این چالش، از روش کاهش داده استفاده می‌شود که هدف افزایش کارایی ذخیره‌سازی و کاهش هزینه ذخیره‌سازی و تجزیه و تحلیل داده‌ها می باشد. این شامل تجمع مکعب داده، جمع‌آوری ویژگی‌های زیر مجموعه، کاهش عدد و کاهش ابعاد است[۹].

۵- یادگیری ماشین

یادگیری ماشینی به عنوان یک فناوری دگرگون کننده ظهور کرده است که در قلب هوش مصنوعی قرار دارد و رایانه ها را قادر می‌سازد از داده‌ها یاد بگیرند و عملکرد خود را در طول زمان بهبود بخشند. این صنعت و زمینه‌های تحقیقاتی مختلف را متحول کرده است، تصمیم‌گیری مبتنی بر داده‌ها را قدرتمند می‌کند و بینش‌های بی‌سابقه‌ای را از حجم وسیعی از اطلاعات باز می‌کند. در این دوره از رشد سریع داده‌ها، برنامه‌نویسی مبتنی بر قوانین سنتی اغلب برای رسیدگی به پیچیدگی و تنوع چالش های دنیای واقعی کافی نیست. از سوی دیگر، یادگیری ماشینی با اجازه دادن به سیستم‌ها برای یادگیری الگوها و روابط مستقیماً از داده‌ها، انطباق با محیط‌های متغیر و پیش بینی‌ها و تصمیم گیری‌های دقیق، یک تغییر پارادایم ارائه می‌دهد[۱۰].

یادگیری ماشینی محبوب‌ترین تکنیک پیش‌بینی آینده یا طبقه‌بندی اطلاعات برای کمک به افراد در تصمیم‌گیری‌های ضروری است. یک سیستم یادگیری ماشین مستقیماً برای حل مشکل برنامه‌ریزی نشده است، بلکه برنامه یا مدل خود را بر اساس تعداد زیادی مثال و بر اساس تجربه آزمون و خطا در مورد نحوه حل مشکل توسعه می‌دهد. یادگیری ماشین سعی می‌کند یک تابع ناشناخته را بر اساس جفت ورودی-خروجی به نام داده‌های آموزشی یاد بگیرد. الگوریتم‌های یادگیری ماشین بر روی نمونه‌هایی آموزش داده می‌شوند که از طریق آنها از تجربیات گذشته یاد می‌گیرند و همچنین داده‌های تاریخی را تجزیه و تحلیل می‌کنند. به عبارت دیگر، همانطور که الگوریتم‌های یادگیری ماشینی بارها و بارها به مثال‌ها آموزش می‌دهند، قادر به شناسایی الگوها به منظور پیش‌بینی در مورد آینده هستند. تعاریف زیادی از یادگیری ماشینی وجود دارد که در طول سال‌ها تکامل یافته است. اولین تعریف یادگیری ماشینی توسط آرتور ساموئل در سال ۱۹۵۹ ارائه شد: "یادگیری ماشین رشته تحصیلی است که به رایانه ها توانایی یادگیری بدون برنامه ریزی صریح را می‌دهد. در یادگیری ماشینی برنامه ارائه نمی‌شود، اما قرار است تعریف شود که این داده است، این خروجی مورد انتظار است، و کامپیوتر باید این را مدل کند یا خودش برنامه‌ریزی کند تا خروجی لازم هنگام ارائه داده‌ها به عنوان خروجی یادگیری ماشین ارائه شود. این یکی از اولین تعاریف یادگیری ماشینی است که حتی امروزه نیز خوب است زیرا امروزه یادگیری ماشینی همه چیز در مورد یادگیری یک مدل است. دیدگاه دیگر در تعریف یادگیری ماشینی مربوط به یادگیری از داده ها و بهبود عملکرد است. یادگیری ماشینی را به عنوان "هر فرآیندی که توسط آن یک سیستم عملکرد خود را بهبود می بخشد" تعریف کرد. در سال ۱۹۸۳، سایمون مجدداً یادگیری ماشینی را در همین راستا به این صورت تعریف کرد: یادگیری در سیستمهایی که تطبیق‌پذیر هستند به این معنا است که سیستم را قادر می‌سازد همان عمل (یا وظایفی را که از جمعیتی از وظایف مشابه برمی‌آیند) به طور مؤثرتر در مرحله بعد انجام شود[۱۱].

یادگیری مهمترین جنبه عقل انسان و اساسی‌ترین روش کسب اطلاعات است. یادگیری ماشینی اساسی‌ترین روش برای هوشمندسازی ماشین است. اگر کامپیوتر قادر به یادگیری نباشد، هوشمند در نظر گرفته نخواهد شد. یادگیری یک فرآیند ذهنی یکپارچه است که شامل به خاطر سپردن، تفکر، ادراک، عواطف و سایر عملکردهای ذهنی است که همگی در هم تنیده هستند. در نتیجه، محققان حوزه‌های مختلف، تفاسیر گوناگونی را بر اساس تخصص‌های خود و همچنین دیدگاه‌های گوناگون ارائه می‌کنند. با رشد مداوم داده ها، پلتفرمی لازم است که بتواند این حجم عظیم داده را مدیریت کند. تکنیک‌های یادگیری ماشینی، مانند یادگیری عمیق، تولید صحیح پیش‌بینی‌ها را برای اکثریت قاطع اطلاعات امکان‌پذیر می‌سازد. یادگیری ماشینی طرز تفکر ما در مورد داده‌ها یا انواع بینش‌هایی که انسان‌ها می‌توانند از آن به دست آورند را تغییر داده است[۱۲].

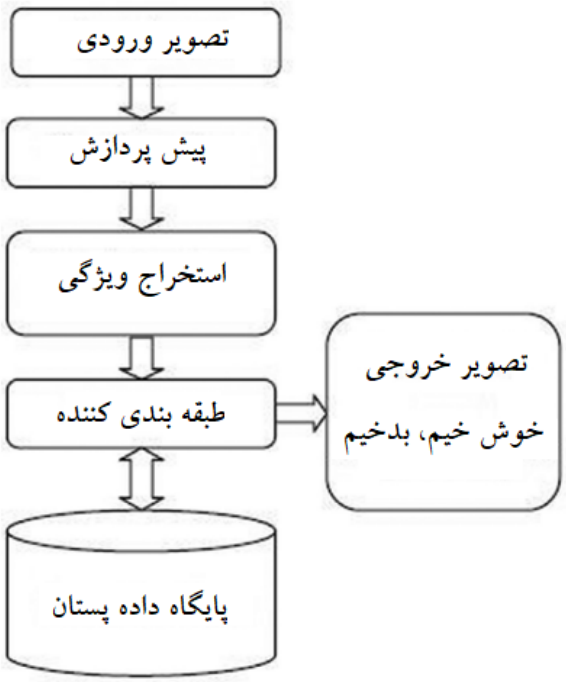
شکل ٤: نمایی از نحوه کار یادگیری ماشین[۱۲].

۶- سیستم تشخیص سرطان پستان مبتنی بر یادگیری ماشین

فن‌آوری‌ها در مراقبت‌های بهداشتی شامل نگهداری و بازیابی سوابق پزشکی الکترونیکی بیماران و دستگاه های درگیر است. تشخیص سرطان همیشه یک چالش در طرح تشخیص و درمان بیماری های هماتولوژیک بوده است. در حال حاضر، درصد قابل توجهی از جمعیت به یک یا چند بیماری مبتلا هستند. سال‌های اخیر شاهد پیشرفت‌های شگرفی در علم پزشکی بوده است. علیرغم این پیشرفت‌ها، هنوز فقدان عظیمی از اطلاعات عمومی در مورد سلامت و بیماری وجود دارد. بخش بزرگی از جمعیت احتمالاً از مشکلات سلامتی رنج می‌برند که برخی از آنها حتی ممکن است کشنده باشند. علاوه بر بهبود دقت در تشخیص سریع شرایط کشنده، اتخاذ تکنیک‌های ایمن، واقع بینانه و استفاده از فناوری مدرن می‌تواند نیاز به مراقبین را کاهش دهد و هزینه‌های کلی مراقبت‌های بهداشتی را کاهش دهد. جان چندین نفر را می‌توان از طریق نوآوری در استراتژی‌های تصمیم‌گیری هوشمند و فناوری‌ها نجات داد. سرطان با رشد بی‌رویه و نابهنجار سلول ها به دلیل ترکیبی از نقص‌های ژنتیکی و اپی ژنتیک مشخص می‌شود. این رشد کنترل نشده سلول‌ها به توسعه تومور کمک می‌کند. اگر تومور با پیشرفت سرطان شروع به متاستاز سریع به سایر اندام‌ها و سیستم‌های بدن کند، ممکن است بیماری در زمان کشف غیرقابل درمان باشد. سرطان سینه عمدتاً زنان را تحت تأثیر قرار می‌دهد (با کمتر از ۱٪ موارد غیر زن). تقریباً از هر هشت زن یک نفر در طول زندگی خود به سرطان سینه مبتلا می‌شود. سالانه تقریباً ۲.۱ میلیون زن مبتلا به سرطان سینه تشخیص داده می‌شوند و شدیدترین آنها زنان بین ۴۰ تا ۷۰ سال هستند. بنابراین، تشخیص زودهنگام سرطان پستان برای پیش آگهی خوب مهم است. علیرغم این واقعیت که علائم ممکن است در مراحل اولیه ضعیف باشند، در صورت تشخیص زودهنگام، شانس بقا به طور چشمگیری افزایش می‌یابد. روش‌های مختلف غربالگری که برای تشخیص سرطان سینه استفاده می‌شود شامل سیتولوژی آسپیراسیون با سوزن ظریف، بیوپسی جراحی با هدایت اولتراسوند و ماموگرافی است. در سینه‌های متراکم، میزان تشخیص سرطان با استفاده از ماموگرافی بسیار ضعیف است و حدود ۱۰ تا ۳۰ درصد موارد تشخیص داده نمی شود. شناسایی دقیق سلول‌های سرطانی برای کاهش میزان مرگ و میر بسیار مهم است و این شامل تشخیص و درمان زودهنگام سرطان موثر برای افزایش نرخ بقای بیماران سرطانی است.

یادگیری ماشینی یکی از محبوب‌ترین مدل‌ها برای آموزش آسان ماشین‌ها و ایجاد مدل‌های پیش‌بینی برای تصمیم‌گیری موفق است. یادگیری ماشینی به تشخیص زودهنگام سرطان سینه کمک می‌کند و با تجزیه و تحلیل اندازه تومور ماهیت سرطان را تعیین می‌کند. روش‌های یادگیری ماشینی، رویکردهای پیشرو برای به دست آوردن نتایج مطلوب در میان مسائل طبقه‌بندی و پیش‌بینی هستند. تحقیقات سرطان سینه می‌تواند از تکنیک‌های یادگیری ماشین که برای شناسایی سرطان و پیش‌بینی وجود یا عدم وجود تومورها استفاده می‌شود، سود ببرد. تکنیک‌های یادگیری ماشینی همچنین می‌تواند برای پیش بینی بدخیمی تومور مورد استفاده قرار گیرد. برای تشخیص و پایش بیماری‌ها، روش‌های مرسوم مورد استفاده به شدت مبتنی بر تشخیص وجود ویژگی‌های سیگنال خاص توسط ناظر انسانی است. در دهه گذشته، توسعه چندین رویکرد تشخیص به کمک کامپیوتر توسط بسیاری از

بیماران در بخش‌های مراقبت‌های ویژه که نیاز به نظارت دائمی دارند، انجام شده است. معیارهای تشخیص عمدتاً کیفی به یک ویژگی کمی مشخص‌تر برای بهبود موضوع طبقه‌بندی در این تکنیک‌ها تبدیل می‌شوند. مدل سرطان سینه با الگوریتم های یادگیری ماشین در شکل ۵ نشان داده شده است. این مدل می‌تواند برای پیش‌بینی سلول‌های سرطانی خوش خیم و بدخیم استفاده شود. ابتدا داده‌های تصویر پستان بارگیری می‌شوند، سپس استخراج ویژگی انجام می‌شود، و سپس مدل طبقه‌بندی نهایی را می‌توان برای انجام وظیفه ذکر شده در بالا آموزش داد. سرطان بدخیم با رشد نامنظم سلولی شروع می‌شود و می‌تواند به سرعت به بافت‌های اطراف گسترش یابد یا به بافت‌های اطراف نفوذ کند، که آن را به یک وضعیت تهدید کننده زندگی تبدیل می‌کند. در مقابل، تومورهای خوش خیم غیرسرطانی و معمولاً غیر کشنده در نظر گرفته می شوند.



شکل۵: بلوک دیاگرام سیستم تشخیص سرطان پستان

۷- معرفی طبقه بند های موثر برای سیستم تشخیص سرطان پستان

الگوریتم‌های داده‌کاوی و یادگیری ماشین در تشخیص و پیش‌بینی انواع سرطان پستان کمک می‌کنند. تکنیک‌های داده کاوی مانند طبقه‌بندی، رگرسیون و خوشه‌بندی کمک می‌کنند تا اطلاعات معنی‌داری در مورد بیماران مبتلا به سرطان پستان به‌دست آید. این الگوریتم‌ها از مجموعه داده‌های آموزشی تشکیل شده‌اند که با کمک این مجموعه داده‌ها می‌توان شانس پیش‌بینی انواع مختلف سرطان سینه را پیدا کرد[۱۳]. در زیر الگوریتم های پرتکرار و مناسب که طبقه‌بندی جهت تشخیص و پیش‌بینی انواع سرطان پستان می‌تواند مناسب باشد معرفی شده است[۱۴]:

ماشین بردار پشتیبان: یک الگوریتم یادگیری ماشین قدرتمند است که برای طبقه‌بندی و رگرسیون استفاده می‌شود و با یافتن ابر صفحه بهینه که نقاط داده را به بهترین شکل از هم جدا می‌کند، کار می کند. این الگوریتم می‌تواند داده‌های خطی و غیرخطی را با استفاده از توابع هسته مختلف مدیریت کند تا داده‌ها را به فضاهای ابعادی بالاتر تبدیل کند و آن را برای انواع مختلف مشکلات متنوع کند[۱۴]. این الگوریتم جز مدلهای یادگیری تحت نظارت هستند که داده‌های مورد استفاده برای طبقه‌بندی و تحلیل رگرسیون را شناسایی می‌کنند. این تکنیک می‌تواند از طریق آموزش طبقه‌بندی، کلاس نمونه جدید را پیش‌بینی کند. ماشین بردار پشتیبان اجرای تقریبی اصل حداقل‌سازی ریسک ساختاری است که اولین اجرای عملی آن در اوایل دهه ۹۰ بود. دو نوع تکنیک ماشین بردار پشتیبان (برنامه‌نویسی ریاضی و عملکرد هسته) وجود دارد که برنامه‌نویسی ریاضی یک جداکنندهٔ خطی (آبر صفحه) بهینه بین نقاط داده کلاس‌های مختلف در فضای ابعادی بالا را جستجو می‌کند و عملکرد هسته، طبقه‌بندی غیرخطی، نقشه‌برداری کامل از ورودی‌ها در فضای بعد بالا است[۱۵].

درخت تصمیم: یک الگوریتم اساسی یادگیری ماشینی است که با تقسیم بازگشتی مجموعه داده بر اساس اطلاعاتی‌ترین ویژگی‌ها، تصمیم‌گیری می‌کند و بصری هستند به طوریکه تفسیر آنها آسان است و می‌توانند داده‌های دسته‌بندی و عددی را مدیریت کنند[۱۴].

جنگل تصادفی: یک روش یادگیری مجموعه ای است که به موضوع بیش از حد درختان تصمیم می‌پردازد. چندین درخت تصمیم را ترکیب می‌کند و با میانگین‌گیری یا رای دادن به خروجی‌های این درخت‌ها، پیش‌بینی می‌کند. این روش مجموعه، دقت و تعمیم را بهبود می بخشد در حالی که قابلیت تفسیر را حفظ می‌کند[۱۴]. جنگل تصادفی یکی از معروف‌ترین و تاثیرگذارترین تکنیک‌ها برای داده‌کاوی است. یک نرم افزار داده کاوی به نام بگینگ یا وب اپلیکیشن خوشه‌بندی سلسله مراتبی است. بوت استرپ یک روش ریاضی واقعاً مؤثر برای یک

مقدار تقریبی از یک مجموعه داده مانند یک رسانه است. چندین آزمون از داده‌ها گرفته می‌شود، میانگین اندازه‌گیری شده و سپس همه متغیرهای نمونه جمع می‌شوند تا میانگین واقعی بهترین پیش‌بینی را به دست آورند. از همین رویکرد برای برچسب‌گذاری استفاده می‌شود، اما معمولاً به جای محاسبه میانگین هر نمونه از درخت‌های فرضی استفاده می‌شود. نمونه-های زیادی از داده‌های آزمون گرفته شده، مدل‌هایی برای هر نمونه داده تولید و مدل پیش‌بینی ارائه می‌شود، اما این پیش‌بینی‌ها برای تخمین بهتر ارزش واقعی خروجی میانگین می‌شوند (اگرچه پیش‌بینی هر داده ضروری است) [۱۶].

تقویت گرادیان شدید[4]: یک الگوریتم تقویت گرادیان است که برای سرعت و عملکرد طراحی شده است. درخت‌های تصمیم را به صورت متوالی می‌سازد، جایی که هر درخت خطاهای قبلی را تصحیح می‌کند. این فرآیند تکراری به تقویت گرادیان شدید اجازه می‌دهد تا مدل‌های بسیار دقیق ایجاد کند و شامل تکنیک‌های منظم‌سازی برای جلوگیری از بیش‌افزاری می‌شود. به دلیل کارایی و قدرت پیش بینی آن به طور گسترده در مسابقات علم داده و برنامه های کاربردی در دنیای واقعی استفاده می شود[۱۴].

بیز ساده[5]: بیز ساده یک راه آسان برای ایجاد مدل‌های پیشگوی آماری است که مبنای آن تئوری بیز است. در این روش کلاس‌های مختلف هر کدام به شکل یک فرضیه دارای احتمال در نظر گرفته می‌شوند. هر رکورد آموزشی جدید احتمال درست بودن فرضیه‌های پیشین را افزایش و یا کاهش می‌دهد و درنهایت، فرضیاتی که دارای بالاترین احتمال شوند، به‌عنوان یک کلاس در نظر گرفته‌شده و برچسبی بر آن‌ها زده می‌شود. این تکنیک با ترکیب تئوری بیز و رابطه سببی بین داده‌ها، به طبقه‌بندی می‌پردازد[۱۷, ۱۸].

k-نزدیکترین همسایه: این طبقه‌بند در واقع یک دسته‌بند مبتنی بر نمونه است. واحد پارامترها شامل نمونه‌هایی هستند که در این روش مورد استفاده قرار می‌گیرند و سپس این الگوریتم فرض می‌کند که تمام این نمونه‌ها در واقع نقاطی در یک فضای n-بُعدی RN هستند. این الگوریتم بسیار مفید است زیرا اطلاعات موجود در داده‌های آموزشی هرگز از بین نمی‌روند. از این رو، این الگوریتم برای زمانی مناسب است که مجموعه‌داده‌ی آموزشی بسیار بزرگ باشد، این الگوریتم وقت‌گیر است، زیرا در هنگام دسته‌بندی یک داده‌ی جدید باید هر نمونه داده‌ی موجود در مجموعه‌داده‌ی آموزشی پردازش شود و این فرآیند باعث افزایش مدت زمان دسته‌بندی می‌شود. با توجه به تحقیقات صورت گرفته در مراجع ذکر شده به این نتیجه می‌توان رسید که دقت دسته‌بندی همان چیزی است که نویسندگان این مراجع به دنبال آن بوده‌اند و به همین دلیل مدت زمان دسته‌بندی مورد توجه واقع نشده است، چرا که دقت دسته‌بندی در تشخیص پزشکی در مقایسه با زمان دسته‌بندی، بسیار مهمتر و حیاتی‌تر است[۱۹].

شبکه عصبی مصنوعی: استعاره‌ای از مغز انسان است که برای پردازش اطلاعات استفاده می‌شود. مجموعه پیوندی از واحدهای ورودی و خروجی در شبکه عصبی گنجانده شده است. شبکه برای تشخیص الگو با استفاده از داده‌های ورودی آموزش دیده و شبکه می‌داند و می‌تواند برچسب‌های گروه را با اعمال تغییرات وزن به طور دقیق تخمین بزند. روش پس انتشار آموزش انجام می‌شود و اگر حلقه‌ای از پیوندها وجود نداشته باشد، یک شبکه عصبی را شبکه عصبی پیشخور می نامند. سه لایه در آن وجود دارد که اولین و آخرین لایه به ترتیب به عنوان لایه اولیه و لایه خروجی و یک لایه بین آنها به عنوان لایه پنهان شناخته می‌شود. همه این لایه‌ها دارای پیوند هستند. ویژگی‌های نمونه آموزشی به عنوان ورودی به شبکه منتقل می‌شوند. سپس ورودی‌ها با هم در وزن‌ها به لایه‌های پنهان ارسال می‌شوند. لایه خروجی نیز خروجی وزنی از لایه پنهان قبلی است که برچسب های گروه مورد انتظار را توصیف می‌کند[۱۶].

8- معرفی معیارهای مناسب تست الگوریتم های طبقه بند یادگیری ماشین

در این قسمت برای ارزیابی طبقه بندهای معرفی شده قسمت قبل، نویسندگان از یکسری معیارهای مهم استفاده میکنند. در واقع برای تست نیاز به مجموعه داده استاندارد است که معمولا مجموعه داده‌های پزشکی از پایگاه داده Kaggle که در دسترس عموم قرار دارد در https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data به دست می‌آید. ویژگی های مجموعه داده معمولا از یک تصویر دیجیتالی از نمونه سرطان سینه که با

[4] XGBoost

[5] Naive *Bayes*

فرآیند آسپیراسیون با سوزن ظریف گرفته شده است، به دست آمده است. ویژگی‌های هسته‌های سلولی مشاهده‌شده در عکس فوری برای تعیین ویژگی‌های آنها استفاده می‌شود.

معیارهای عملکرد برای طبقه بندی معیارهای ارزیابی مورد استفاده برای سنجش اثربخشی این تحلیل به شرح زیر است:

صحت: اثربخشی یک مدل با نسبت پیش‌بینی‌های دقیق تولید شده در تمام انواع پیش‌بینی‌ها ارزیابی می‌شود. فرآیند ارزیابی شامل ارزیابی صحت طبقه‌بندی است یعنی نمونه‌های طبقه‌بندی شده با تعداد کلی رخدادها. اندازه‌گیری صحت به ویژه در مواردی ارزشمند است که توزیع کلاس‌ها در متغیر هدف به طور یکنواخت در سراسر مجموعه داده پخش شود. رابطه صحت در معادله (۱) بیان شده است[۲۰]:

$$\text{Accuracy} = \frac{\text{TP}+\text{T}N}{\text{TP}+\text{FP}+\text{FN}+\text{T}N} \quad (۱)$$

یادآوری: یادآوری اندازه‌گیری نرخ مثبت واقعی در زمینه یک سیستم نقص نرم‌افزاری است. در این زمینه خاص، تعداد رخدادهایی را نشان می‌دهد که به عنوان نرم‌افزار معیوب طبقه‌بندی شده‌اند که به‌طور دقیق توسط مدل پیش‌بینی شده‌اند. معادله ۲ نشان دهنده نسبت نمونه‌های نرم‌افزاری مشکل‌ساز است که به طور دقیق توسط مدل شناسایی شده‌اند[۲۰]:

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \quad (۲)$$

دقت: دقت پیش‌بینی‌های مثبت انجام‌شده توسط مدل را اندازه‌گیری می‌کند و نسبت نمونه‌های پیش‌بینی‌شده به‌عنوان مثبت و مثبت واقعی را نشان می‌دهد و با استفاده از فرمول زیر محاسبه می شود[۱۴]:

$$\text{precision} = \frac{\text{TN}}{\text{TN}+\text{FP}} \quad (۳)$$

خاصیت: ویژگی، که به عنوان نرخ منفی واقعی شناخته می شود، در حوزه نقص نرم افزار دارای ارتباط است و از رابطه ۴ به دست می‌آید، درصد نمونه‌هایی را در سیستم نرم‌افزاری که بدون نقص هستند و به درستی توسط مدل به عنوان نمونه طبقه بندی می‌شوند، ارزیابی می کند[۲۰]:

$$\text{Specifity} = \frac{\text{TP}}{\text{TP}+\text{FP}} \quad (٤)$$

امتیاز F1: معیاری است که دقت و یادآوری را در یک مقدار واحد ترکیب می‌کند. این به ویژه زمانی مفید است که توزیع کلاس ناهموار (کلاس های نامتعادل) وجود داشته باشد و اغلب در مسائل طبقه بندی باینری استفاده می شود.این معیار میانگین هارمونیک دقت و یادآوری است و با استفاده از فرمول زیر محاسبه می شود[۱۴]:

$$F_1\text{score} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision}+\text{recall}} \quad (۵)$$

بر اساس معیارهای بالا مقایسه‌ای برای ارزیابی تکنیک های نامبرده در بخش قبل که توسط نویسندگان تنظیم شده است، در این بخش اشاره شده است.

جدول ۱ مقایسه جامعی از الگوریتم‌های یادگیری ماشین و همچنین طبقه‌بندی‌کننده‌های مجموعه در مجموعه داده سرطان پستان ارائه می‌کند. در میان طبقه‌بندی‌کننده‌های یادگیری ماشین، ماشین بردار پشتیبان نتیجه بالا صحت و دقت متعادل، یادآوری، ویژگی و امتیاز F1 را نشان می‌دهد. بالاترین مقادیر برای صحت، دقت، یادآوری، ویژگی و امتیاز F1 با متن پررنگ برجسته شده است. روندهای مشابهی با تقویت گرادیان شدید، جنگل تصادفی، درخت تصمیم مشاهده می‌شود، به طور کلی، می‌توان ترتیب الگوریتم‌ها را در افزایش پیش‌بینی سرطان سینه مشاهده کرد. در این ارزیابی همه تکنیک‌ها روی مجموعه داده های سرطان پستان مختلف انجام شده است. در واقع در مقاله [۱۴] آزمایش‌ها بر روی سه مجموعه داده، ویسکانسین ، دانشگاه کالیفرنیا، ایروین و مجموعه داده‌های سرطان سینه کویمبرا انجام شد که مجموعه داده‌ها به ترتیب شامل ۵۶۹، ۲۸۶ و ۱۱۶ نمونه بودند[۱۴].

همچنین یکسری از مقالات فقط به برسی معیار صحت یکسری از طبقه بندها پرداخته اند. اکنون جدول ۲ خلاصه مقایسه ای از تکنیک‌های یادگیری ماشینی برای پیش بینی سرطان پستان را بر اساس ابزارها، منابع داده، نوع داده، روش پیش پردازش داده ها، روش ارزیابی داده ها، روش اعتبار سنجی و سطح دقت هر الگوریتم در موقعیت های مختلف ارائه می‌دهد.

جدول ۱: مقایسه طبقه‌بندی‌کننده‌های یادگیری ماشین در مجموعه داده های مختلف سرطان پستان [۱۴].

روش	مجموعه داده	**صحت**	دقت	**بازخوانی**	**خاصیت**	امتیازF1
ماشین بردار پشتیبان		96.50	96.67	97.75	94.44	97.21
تقویت گرادیان شدید		94.40	95.56	95.56	92.45	95.56

جنگل تصادفی	ویسکانسین	96.50	96.67	97.75	94.44	97.21
درخت تصمیم		95.10	93.33	98.82	89.66	95.99
ماشین بردار پشتیبان	دانشگاه کالیفرنیا، ایروین	95.10	94.23	92.45	96.67	93.33
تقویت گرادیان شدید		95.80	96.08	92.45	97.78	94.23
جنگل تصادفی		95.10	91.07	96.22	94.44	93.57
درخت تصمیم		90.20	85.45	88.68	91.11	87.03
ماشین بردار پشتیبان	کویمبرا	62.02	62.49	62.56	67.31	62.06
تقویت گرادیان شدید		70.69	70.40	70.01	63.46	70.12
جنگل تصادفی		70.69	70.50	69.83	61.53	69.97
درخت تصمیم		65.52	65.07	98.82	57.70	64.85

جدول2: بررسی مقایسه ای تکنیک های یادگیری ماشینی برای پیش بینی سرطان سینه

روش	مجموعه داده	تعداد ویژگی ها	نوع داده	روش پردازش داده ها	پیش پردازش	صحت
ماشین بردار پشتیبان	مخزن UCI WDBC WPBC	32WDBC 34WPBC	عددی	انتخاب ویژگی	مقدار گسسته	93%
K نزدیکترین همسایه						95%
بیز ساده						92%
ماشین بردار پشتیبان	مخزن UCI	۳۲ ویژگی	عددی	انتخاب ویژگی کاهش ابعاد	مقدار گسسته	92.78%
تقویت گرادیان شدید						92.78%
ماشین بردار پشتیبان	مرکز ایرانیان ICBC	۲۲ ویژگی	عددی	پاکسازی و آماده سازی داده ها	مقدار ترکیبی	95.7%
شبکه عصبی مصنوعی						94.7%
درخت تصمیم						93.6%

در این بخش الگوریتم‌های مختلف یادگیری ماشین و داده‌کاوی برای پیش بینی سرطان سینه بررسی شد تمرکز اصلی یافتن مناسب‌ترین الگوریتمی است که می‌تواند بروز سرطان سینه را به طور مؤثرتری پیش‌بینی کند. هدف اصلی این بخش، برجسته کردن تمام مطالعات قبلی در مورد الگوریتم‌های یادگیری ماشینی است که برای پیش‌بینی سرطان سینه استفاده می‌شوند، این بخش اطلاعاتی لازم را در اختیار مبتدیانی قرار می‌دهد که می‌خواهند الگوریتم‌های یادگیری ماشین را تجزیه و تحلیل کنند تا پایه یادگیری عمیق را به دست آورند.

9- نتیجه‌گیری و کارهای آتی

بروز سرطان سینه در طول سال‌ها به دلیل تغییر در شیوه زندگی و محیط به طور پیوسته افزایش یافته است. در حال حاضر، سرطان سینه یکی از علل اصلی مرگ و میر ناشی از سرطان در میان زنان است که آن را به یک نگرانی حیاتی برای سلامت عمومی جهانی تبدیل کرده است. بنابراین، ایجاد یک سیستم تشخیص خودکار برای سرطان سینه اهمیت زیادی در جامعه پزشکی دارد .در چند سال گذشته، روش‌های مختلف تشخیص سرطان پستان با استفاده از الگوریتم‌های یادگیری ماشین و حسگرهای زیستی توسعه یافته‌اند. علاوه بر این، تکنیک‌های غربالگری سینه و تکنیک‌های تجزیه و تحلیل داده-های بزرگ نیز در طول زمان در حال افزایش هستند. بیشتر

پایگاه‌های داده به پیش‌پردازش وابسته به مدل یادگیری ماشین مورد استفاده نیاز دارند. با این حال، تفاوت‌هایی در فاکتورهای اندازه‌گیری عملکرد وجود دارد و مدل‌های مختلف بسته به مجموعه داده از عناصر متریک متفاوتی استفاده می‌کنند. در این مقاله بخش های اول به مفاهیم عمومی سرطان سینه، داده کاوی و یادگیری ماشین اشاره شد سپس در بخش اصلی به معرفی تکنیک های مفید یادگیری ماشین شامل ماشین بردار پشتیبان، درخت تصمیم، جنگل تصادفی، تقویت گرادیان شدید، بیز ساده، k-نزدیکترین همسایه و شبکه عصبی مصنوعی که برای طبقه‌بندی و پیش بینی سرطان سینه می‌تواند مفید باشد اشاره شد. در ادامه برای ارزیابی این تکنیک‌ها معیارهای مناسب دقت، صحت, بازخوانی، خاصیت و امتیاز F_1 معرفی و تعریف شد و جهت ارزیابی این تکنیک ها از مقالاتی که در این زمینه تحقیقات مروری مناسبی ارائه داده بودند استفاده شد و از آزمایش‌هایی که روی مجموعه داده‌های استاندارد انجام شد، بخش‌های مناسب تحلیل و بررسی شد. به طور کلی می‌توان نتیجه گرفت از بین تکنیک های معرفی شده ماشین بردار پشتیبان، جنگل تصادفی و درخت تصمیم به ترتیب به نسبت بقیه دارای دقت و صحت بهتری و بهینه تری را نتیجه دادند. این مقاله می‌تواند برای علاقه‌مندان به این موضوع بسیار حائز اهمیت باشد تا از این این سه تکنیک به صورت ترکیبی با الگوریتم‌های تکاملی و یا ترکیبی از این سه تکنیک بتوانند روش‌های بسیار مفیدی جهت پیش بینی و تشخیص سرطان سینه پیشنهاد دهند.

10- منابع

1. *Arici, M.Ö. and M. Kocer, The Impact of Prognostic Factors on Survival in Patients with Non-Metastatic Invasive Breast Cancer: A Single-Center Experience: Prognostic factors and breast cancer survival. Archives of Breast Cancer, 2025. 12(1): p. 38-46.*
2. *Tang, D.-D., et al., Survival feature and trend of female breast cancer: A comprehensive review of survival analysis from cancer registration data. The Breast, 2025: p. 103862.*
3. *Badu-Peprah, A., et al., Five-Year Survival Outcomes for Breast Cancer Patients Across Continental Africa: A Contemporary Review of Literature with Meta Analysis. medRxiv, 2025: p. 2025.01. 03.25319952.*
4. *Yara, D. and T. Oroszi, Understanding Breast Cancer: A Comprehensive Review of Epidemiology, Risk Factors, and Treatment Strategies. Advances in Breast Cancer Research, 2025. 14(1): p. 1-15.*
5. *Zhou, S., et al., Breast cancer prediction based on multiple machine learning algorithms. Technology in Cancer Research & Treatment, 2024. 23: p. 15330338241234791.*
6. *Mohammad Asim khan, S.A., An Emergence of AI in Data Mining and KDD: ANN its Strength & Weakness. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-9 Issue-1, May 2020, 2020.*
7. *Singh, R. and T.H. Sheikh, An Overview of Data Mining Applications in Healthcare. International Journal, 2016. 4(2).*
8. *Kavyasree S Anil, R.J., A REVIEW ON DATAMINING TECHNIQUES IN HEALTHCARE SECTOR. Electronic copy available at: https://ssrn.com/abstract=4019442, 2022.*
9. *Reddy, R.P., Ch Mandakini, and Ch Radhika. , A Review on Data Mining Techniques and Challenges in Medical Field. International Journal of Engineering and Technical Research V9(08), 2020.*
10. *Khan, J., Machine Learning: An Introduction to Algorithms and Applications. 2023.*
11. *Geetha, T. and S. Sendhilkumar, Machine Learning: Concepts, Techniques and Applications. 2023: CRC Press.*
12. *Chopra, R.K., A Review on Machine Learning and Its Applications. International Journal of Innovative Research in Computer Science & Technology, 2022. 10(2): p. 300-304.*
13. *Fatima, N., et al., Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 2020. 8: p. 150360-150376.*
14. *Islam, M.R., M.S. Islam, and S. Majumder, Breast cancer prediction: a fusion of genetic algorithm, chemical reaction optimization, and machine learning techniques. Applied Computational Intelligence and Soft Computing, 2024. 2024(1): p. 7221343.*
15. *Chowdary, B.V., and Y. Radhika. , "A survey on applications of data mining techniques.". International Journal of Applied Engineering Research 13, no. 7 (2018): 5384-5392., 2018.*
16. *Saleh, B.J., et al., A REVIEW P A REVIEW PAPER: AN APER: ANALYSIS OF WEKA D SIS OF WEKA DATA MINING A MINING TECHNIQUES FOR HEART DISEASE PREDICTION SYSTEM. University of*

Nebraska - Lincoln DigitalCommons@University of Nebraska - Lincoln 2020.

17. *Fdez-Glez, J., David Ruano-Ordas, José Ramón Méndez, Florentino Fdez-Riverola, Rosalía Laza, and Reyes Pavón. , "A dynamic model for integrating simple web spam classification techniques.". Expert Systems with Applications 42, no. 21 (2015): 7969-7978, 2015.*
18. *Abd AL-Nabi, D.L., and Shereen Shukri Ahmed. , "Survey on classification algorithms for data mining:(comparison and evaluation)." International Journal of Computer Engineering and Intelligent Systems 4, no. 8 (2013): 18-27., 2013.*
19. *Jothi, N. and W. Husain, Data mining in healthcare–a review. Procedia computer science, 2015. 72: p. 306-313.*
20. *Kene Tochukwu Anyachebelu, S.H.H., Muhammad Umar Abdullahi, Maimuna Abdullahi Ibrahim, Comparative Analysis of Machine Learning Algorithms for Breast Cancer Prediction Dutse Journal of Pure and Applied Sciences (DUJOPAS), Vol. 9 No. 4b December 2023, 2023.*

مائده رحمانی، دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش نرم افزار ، دانشگاه پیام نور مرکز بین الملل کیش می باشد و نشانه رایانامه ایشان عبارتند از :
Maede9708@gmail.com

مهدی قاسمی، دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز ، دانشگاه پیام نور مرکز بین الملل کیش می باشد و نشانه رایانامه ایشان عبارتند از:
Mahdikmg1@gmail.com

حمید زنگی آبادی زاده، دانشجوی کارشناسی ارشد مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز، دانشگاه پیام نور مرکز بین الملل کیش می باشد و نشانه رایانامه ایشان عبارتند از:
Hamid.zangiabadi@gmail.com

فرشید وظیفه دوست، فارغ التحصیل در مقطع کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز از دانشگاه پیام نور مرکز بین الملل قشم می باشد و نشانه رایانامه ایشان عبارتند از:
Vazifehdoostfarshid@gmail.com

سمیه کدخدا ده خانی، فارغ التحصیل مقطع کارشناسی ارشد رشته مهندسی کامپیوتر گرایش هوش مصنوعی و رباتیکز دانشگاه پیام‌نور قشم می باشد، او به عنوان کارشناس فناوری در دانشگاه پیام‌نور استان کرمان مشغول به کار می باشد و نشانه رایانامه ایشان عبارتند از:
Emailsk65@gmail.com

ح. زنگی آبادی زاده، م. قاسمی، ف. وظیفه دوست ، س. کدخدا ده خانی و م. رحمانی. امنیت رایانش ابری: مروری بر مفاهیم پایه‌ای، چالش‌های امنیتی ، مسائل ، الزامات، استانداردهای امنیتی و انواع حملات در رایانش ابری. دوفصلنامه محاسبات و سامانه‌های توزیع شده، سال هفتم، شماره ۱، شماره پیاپی ۱۳، صفحه ۵۱ تا ۶۲، سال ۱۴۰۳

How to cite: F.Vazifehdoost، S.kadkhodadehkhani، M.Rahmani، M.Ghasemi ،H.Zangiabadi Zadeh، **A survey of the concepts of early breast cancer prediction techniques and evaluation of these techniques based on appropriate criteria,** Journal of Distributed Computing and Systems (JDCS), Vol 7, Issue 1, Pages 51 - 62, 2024.

A survey of the concepts of early breast cancer prediction techniques and evaluation of these techniques based on appropriate criteria

F.Vazifehdoost[1]، S.kadkhodadehkhani[2] ، M.Rahmani[3] ،M.Ghasemi،[4] H.Zangiabadi Zadeh[5]

[1]Payame Noor University, Qeshm International Center.
[2]Payame Noor University, Qeshm International Center.
[3] Payame Noor University, Kish.
[4] Payame Noor University, Kish International Center.
[5] Payame Noor University, Kish International Center.

Abstract

The incidence of breast cancer has been increasing steadily over the years due to changes in lifestyle and environment. Currently, breast cancer is one of the leading causes of cancer mortality among women, making it a critical concern for global public health. Therefore, the development of an automated breast cancer detection system is of great importance in the medical community. This article discusses the general concepts of breast cancer, data mining, and machine learning, and also introduces useful machine learning techniques that can be useful for breast cancer classification and prediction. In the main part of the article, the evaluation and comparison of these techniques were reviewed and performed based on appropriate criteria of accuracy, precision, recall, specificity, and F1 score. The results of the evaluation of these classifiers from appropriate researches whose experiments were conducted on standard datasets showed that among the introduced techniques, support vector machine, random forest, and decision tree have better prediction and classification than the others, respectively.