باسمه تعالی

پایان‌نامه کارشناسی ارشد

پیش بینی سرطان در شهر عماره عراق با استفاده از مدل‌های یادگیری ماشین

تهیه و تنظیم:

نبیل الراشدی

استاد راهنما:

دکتر عادل قاضی خانی

پاییز 97

تقدیر و تشکر

سپاس مخصوص خداوند مهربان که به انسان توانایی و دانایی بخشید تا به بندگانش شفقت ورزد، مهربانی کند و در حل مشکلاتشان یاری‌شان نماید. از راحت خویش بگذرد و آسایش هم نوعان را مقدم دارد، با او معامله کند و در این خلوص انباز نگیرد و خوش باشد که پروردگار سمیع و بصیر است.

لذا اکنون‌که در سایه‌سار بنده‌نوازی‌هایش پایان‌نامه حاضر به انجام رسیده است بر خود لازم می‌دانم تا مراتب سپاس را از بزرگوارانی به‌جا آورم که اگر دست یاری گرشان نبود، هرگز این پایان‌نامه به انجام نمی‌رسید.

ابتدا از استاد گران‌قدرم جناب آقای دکتر عادل قاضی خانی که زحمت راهنمایی این پایان‌نامه را بر عهده داشتند، کمال سپاس را دارم.

سپاس آخر را به خانواده‌ی عزیزم تقدیم می‌کنم که حضورشان در فضای زندگی‌ام مصداق بی‌ریای سخاوت بوده است.

چکیده

سرطان مرکزی ترین عنصر برای مرگ در سراسر جهان است. پیش بینی اولیه و محل تومور در درمان بیماری مفید می باشد. بنابراین بررسی روش های شناسایی رویداد بیماری در مراحل اولیه گسترش می یابد. پیش بینی دقیق سرطان برای استفاده مفید از داروهای خاص مهم است. گرچه پیش بینی سرطان در سال های اخیر توسعه یافته است، هنوز یک روش کاملا کامپیوتری برای تشخیص سرطان ضروری است. در پایان نامه، ما قصد داریم ویژگی های مؤثر را با استفاده از روش های انتخاب ویژگی شناسایی کنیم و مدلی برای پیش بینی سرطان ارائه کنیم. روش پیش بینی پیشنهادی به تشخیص سرطان کمک می کند. در این پایان نامه، مجموعه داده ای که استفاده خواهد شد دارای 2181 نمونه و 13 ویژگی برای بیماران مبتلا به سرطان در استان میسان در سال های 2013-2018 می باشد. داده ها با استفاده از الگوریتم KMeans بر اساس قابلیت های خوشه بندی تحلیل می شوند. مدل پیش بینی شده با استفاده از سه روش پیش بینی برای داده های گسسته داده شده شامل بیز ساده، J48 و SVM ساخته شده است.

کلمات کلیدی: داده کاوی، خوشه بندی، K-means، پیش‌بینی سرطان، الگوریتم‌های یادگیری ماشین.

فهرست مطالب

فهرست اشکال

فهرست جداول

فصل اول

مقدمه

1-1- مقدمه

به دلیل دسترسی گسترده مقادیر زیاد داده‌ها و نیاز به تبدیل این مقدار عظیم داده‌ها به اطلاعات ارزشمند، موجب شده که استفاده از تکنیک‌های داده‌کاوی ضروری شود. داده‌کاوی و KDD در سال‌های اخیر محبوبیت پیدا کرده است. محبوبیت روش های داده‌کاوی و KDD نباید موجب تعجب شما شود زیرا اندازه مجموعه‌ی داده هایی که اکنون در دسترس هستند بسیار زیاد هستند و همین موضوع موجب شده است که نتوان به صورت دستی با این داده‌ها کار کرد و نیاز به روندهای خودکار برای کاوش و تحلیل این داده‌ها وجود داشته باشد که این روش‌ها بر اساس آمار کلاسیک و یادگیری ماشین برای پردازش داده های عظیم، پویا و مجموعه های شامل اهداف پیچیده می‌باشد.

داده‌کاوی یکی از زمینه های موثر در تحقیقات است که به بررسی داده های مفید در پایگاه داده ها می‌باشد. ما می‌توانیم داده‌کاوی را به عنوان یکی از جدی ترین و جالب ترین زمینه های تحقیقاتی به هدف پوشش داده های مهم از میان مجموعه داده های عظیم، در نظر بگیریم.

در سال های اخیر، ما با تعداد افزاینده ی داده هایی رو به رو هستیم که در سازمان های مختلف مانند بانک ها، بیمارستان ها، دانشگاه ها و غیره، ذخیره شده اند که همین موضوع موجب میشود که راهی را پیدا کنیم که بتوانیم دانش را از این مقدار زیاد اطلاعات به دست بیاوریم و به صورت موثر از آن ها استفاده کنیم [1].

سرطان معمولا با بررسی سلول ها توسط میکروسکوپ تشخیص داده می شود. آزمایش های تصویربرداری مانند توموگرافی کامپیوتری(CT) از طریق نشان دادن رشد غیرطبیعی بافت به نمایش حضور احتمالی سرطان کمک می کنند. تصمیم گیری نهایی معمولا با انجام آزمایش های مختلف آزمایشگاهی از بیمار و مشاهده دقیق سلول های سرطانی انجام می شود. متد دیگری که توسط پزشکان مورد استفاده قرار می گیرد بیوپسی است . بیوپسی با عمل جراحی انجام می شود. پزشکان

نمونه ای از بافت مورد آزمایش می گیرند. سپس با استفاده از میکروسکوپ، این نمونه مورد بررسی قرار می گیرد. ظاهر سلول های طبیعی یکنواخت است؛ آنها مرتب سازماندهی شده و دارای اندازه هایی یکسان هستند. سلول های سرطانی متفاوت از سلول های طبیعی هستند. آنها پراکنده اند، اندازه های آنها متفاوت است و ساختار یافته نیستند. مشکل این است که یک تصویر پزشکی مانند CTاسکن یا MRI نمیتواند تمام الگوها و اطلاعات نوع خاصی از سرطان یا زیرمجموعه از سرطان ها را نشان دهد. مسئله دیگری این است که یک پزشک با چشم غیر مسلح خود و یک میکروسکوپ نمی تواند تعداد زیادی از الگوهای بیماری را به خاطر بیاورد. این برای یک بیمار نگران کننده است که بداند او سرطان دارد [□].

پس از تشخیص سرطان بیمارممکن است همه امید خود را از دست بدهد. بنابراین تشخیص سرطان فرآیندی است که نیازمند به صبر وتامل بسیاری ازهر دو طرف بیمار و پزشک / بیمارستان است. تشخیص زودهنگام سرطان می تواند عمر زندگی یک بیماررا افزایش دهد زیرا سلول های سرطانی باعث تخریب سلول های دیگر و انتشار به سایر قسمت های بدن می شوند. اگر در مرحله اول تشخیص داده شود، درمان زودترآغاز می شود و این می تواند از گسترش بیشتر بیماری جلوگیری کند. سیستم تشخیص موجود در بیمارستان ها در حال حاضر سیستم تشخیص دستی است. به عنوان مثال زمانی که یک بیمارپذیرش شد، او باید از طریق روش تست رادیولوژی یعنی اشعه ایکس، CTیا MRIمعاینه شود. رادیولوژیست اظهارات خود در مورد گزارش آزمایش را ثبت می کند. پس از این فرآیند یک دکتر متخصص گزارش آزمایش اشعه ایکس CT / MRI / را بررسی می کند و اظهار نظر می کند. در برخی از انواع سرطان مثل سرطان سینه و ریه تشخیص مبتنی بر تصمیم نهایی پزشکان است ، اما در سایر انواع سرطان مانند کارسینوم، آزمایش های دیگر مانند بیوپسی نیز لازم است. در سیستم دستی ، رادیولوژیست و پزشک تشخیص سرطان می دهند. این روند به کندی طی می شود و

پس از بازبینی رادیولوژیست پزشک متخصص نیز باید بررسی کند و اظهارات خود را بیان کند و در نهایت بگوید که آیا سرطان وجود دارد یا خیر. خودکار سازی این فرایند ضروری است تا تشخیص سرطان با استفاده از تکنولوژی پیشرفته، کارآمد و سریع باشد [□].

2-1- تعریف مساله

پیش بینی دقیق سرطان از منظر استفاده مفید از داروهای خاص حائز اهمیت است. گرچه پیش بینی سرطان در سال های اخیر پیشرفت کرده است، اما هنوز وجود روشی کاملا کامپیوتری و عینی برای تشخیص سرطان ضرورت دارد [□]. داده کاوی کشف دانش در پایگاه های داده ای است. تکنیک های داده کاوی به پردازش داده ها و تبدیل آنها به اطلاعات مفید کمک می کند. پیش بینی های بدست آمده از نتایج حاصل از داده کاوی نشانگر آنست که این روش در زمینه های مختلف مانند هوش مالی، بیوانفورماتیک، مدیریت بهداشت، امور مالی و غیره مفید است. رشته پزشکی دارای طیف گسترده ای از انواع داده های قابل پردازش ومسائل چالش برانگیز است. این رشته نیازمند تشخیص دقیق و به موقع بیماری است که می تواند زندگی بسیاری از بیماران را نجات دهد. تکنیک های داده کاوی نقش حیاتی در تجزیه و تحلیل مراقبت های بهداشتی ایفا می کنند. تشخیص زودهنگام و نتایج دقیق پزشکان با استفاده از الگوریتم های داده کاوی قابل دستیابی است. الگوریتم های مختلف برای تشخیص بیماری های مختلف قابل استفاده اند . بر اساس داده های مورد استفاده، دقت و عملکرد نیز متفاوت خواهد بود[□]. تکنیک های داده کاوی برای ایجاد یک روش جدید در تشخیص وجود سرطان دریک بیمار خاص اجرا می شوند. هنگام شروع به کار بر روی یک مسئله داده کاوی، ابتدا لازم است تمام داده ها به یک مجموعه منتقل شوند. ادغام داده ها از منابع مختلف معمولا چالش های زیادی را به وجود می آورد. داده ها باید تجمیع ، یکپارچه و تمیز شوند. تنها پس ازآن امکان پردازش از طریق تکنیک های یادگیری ماشینی ممکن است. این سیستم توسعه یافته می

تواند توسط پزشکان و بیماران به راحتی استفاده شود و وضعیت و وخامت سرطان فرد را بدون غربالگری با آزمایش های سرطان نشان دهد. این سیستم همچنین در ضبط و ذخیره حجم زیادی از اطلاعات حساس که می تواند برای بدست آوردن اطلاعات در مورد بیماری و درمان آن مفید باشد قابلیت دارد [□].

پس از آنکه فرد بالغ می شود، بیشتر سلول ها فقط برای جایگزینی سلول های فرسوده یا مرگ و یا برای جبران صدمه ایجاد می شوند. سرطان زمانی شروع می شود که سلول های بخشی از بدن خارج از کنترل شروع به رشد می کنند. انواع مختلفی از سرطان وجود دارد، اما همه آنها به دلیل رشد خارج از کنترل سلول های غیر طبیعی آغاز می شود. رشد سلول های سرطانی از رشد طبیعی سلول ها متفاوت است. به جای مرگ، سلول های سرطانی همچنان رشد می کنند و سلول های جدید و غیر طبیعی ایجاد می کنند. سلول های سرطانی همچنین می توانند به دیگر بافت ها حمله کنند (در آنها رشد کنند)، کاری که سلول های طبیعی نمی توانند انجام دهند. رشد خارج از کنترل و حمله به بافت های دیگر اموری است که یک سلول را سرطانی می کند [□]. با وجود اینکه طبقه بندی سرطان طی □□ سال گذشته توسعه یافته است، هیچ رویکرد عمومی برای شناسایی گونه های جدید سرطان (کشف گونه ها) و یا اختصاص دادن تومور به گونه های شناخته شده (پیش بینی گونه ها) صورت نگرفته است. تعیین ویژگی های موثر چالش برانگیز است. در پروژه ه های انجام شده از آن رو که بسیاری از ویژگی های انتخاب ویژگی در انتخاب ویژگی های مؤثر مورد استفاده قرار نگرفته ، دقت پیش بینی این روش ها مناسب نیست؛ استفاده از تکنیک های داده کاوی موجود می تواند در پیش بینی سرطان با دقت مناسب مفید باشد.

1-3- اهداف تحقیق

در این پایان نامه، برآن شدیم تا ویژگی های مؤثر را شناسایی کنیم و بر اساس این ویژگی، یک مدل پیش بینی سرطان ارائه کنیم. همانطور که در آثار قبلی ذکر شده است، اکثر روش ها به طور تجربی در تلاشند تا ویژگی های انتخاب شده را شناسایی کرده و یک مدل با هدف پیش بینی تعداد سرطان بر اساس ویژگی های انتخاب شده ارائه دهند. اکثر این روش ها با دقت غیر قابل قبول مواجهند. ما برآنیم تا از الگوریتم های داده کاوی استفاده کنیم، استفاده از این الگوریتم ها مستلزم پیش پردازش داده ها با هدف آماده سازی آنهاست.

به طور کلی، دو هدف عمده در این پایان نامه در نظر گرفته شده است:

- ارائه مدلی برای پیش بینی سرطان با استفاده از داده های دنیای واقعی برای تشخیص زود هنگام سرطان.
- شناسایی و رتبه بندی عوامل موثر برای تشخیص زودرس سرطان و تعیین تأثیر هر یک از این ویژگی ها از دیگردستاوردهای این پژوهش است.

1-4- ساختار پایان‌نامه

سازمان‌دهی مطالب پایان‌نامه بدین شرح است که در فصل دوم ابتدا بیان مفاهیم اولیه را شرح می‌دهیم. در فصل سوم یک مرور کلی به پیشینه تحقیق ارائه می‌شود. در فصل چهارم ابتدا روش پیشنهادی را شرح می‌دهیم. و سپس ارزیابی نتایج را مورد هدف قرار می‌دهیم در فصل پنچم نتیجه‌گیری و پیشنهاد برای کارهای آتی ارائه می‌شود.

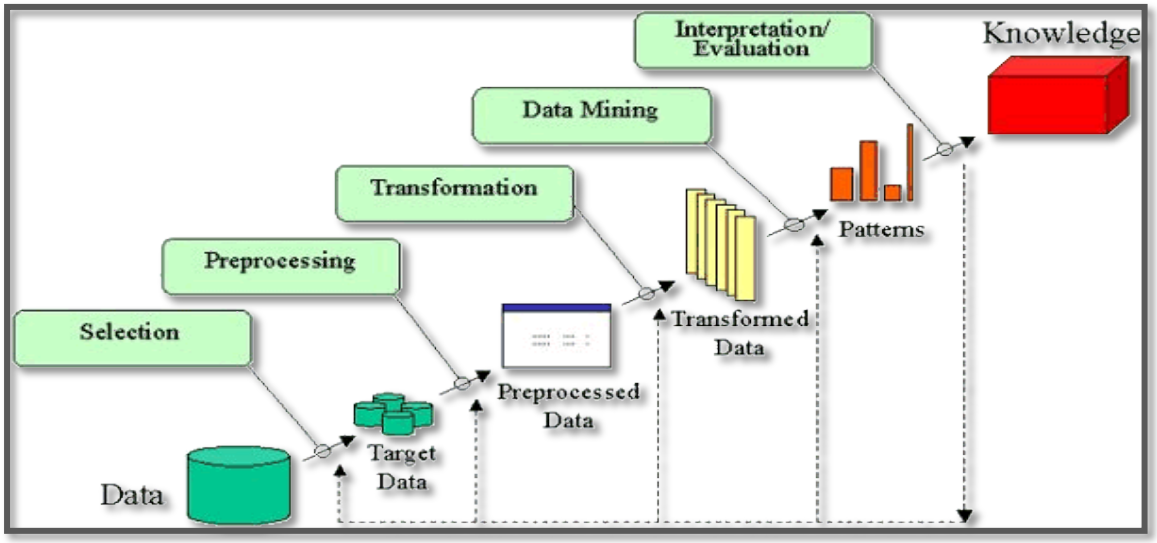
فصل دوم

بیان مفاهیم اولیه

این فصل شامل تعاریف مفاهیم اساسی است. در برگیرنده مفاهیم داده کاوی ، اهمیت داده کاوی و کاربرد داده کاوی . پس از آن داده کاوی در امور پزشکی را مورد بحث قرار می دهیم. سپس به معرفی سرطان و اطلاعات کلی درباره سرطان می پردازیم، چرا سرطان رخ می دهد. در پایان این فصل ما نتیجه گیری خواهیم کرد. که شامل نتیجه گیری محتوای فصل است.

2-1 مفاهیم داده کاوی

داده کاوی به طور کلی فرآیند استخراج اطلاعات جالب پنهان از داده های موجود است که درغیر این صورت دسترسی به آنهاغیرممکن است. با این وجود اینکه این تعریف تصویری نسبتا خام از داده کاوی ارائه می دهد، این مفهوم در قالب های مختلف در گذشته تعریف شده است. این تعاریف متفاوت در ارتباط با معرفی عبارت "کشف دانش در پایگاه های داده ای " در اولین کارگاه کشف دانش در پایگاه داده ای (KDD) (۱۹۸۹) است. از آن به بعد، محققان و نویسندگان KD را با داده کاوی مرتبط دانسته اند و بعضی آنهارا با معنایی یکسان یافته اند ، عمدتا به این دلیل که دانش محصول داده های کشف شده است. در ابتدا چندین روش داده کاوی از دیدگاه کشف دانش شناخته شد، اما محققان از آن زمان به بعد، به تجزیه و تحلیل داده و بر داده کاوی و تکنیک های آن متمرکز شدند. تعریف مشخصی از داده کاوی [۷] بیان می کند که این روش جستجویی برای روابط و الگوهای جهانی در پایگاه های داده های بزرگ است که به علت فراوانی داده ها، مانند رابطه بین بیماران و تشخیص پزشکی آنها پنهان است. علیرغم،تجمیع آرا در استفاده مترادف از اصطلاحات "داده کاوی" و "کشف دانش"، چندین تحلیلگر داده و محقق بر صحت این امر بحث کرده اند [۸].



شکل 2-1- فرآیند کشف دانش در پایگاه داده (KDD) [9]

الف) اهمیت داده کاوی

روش داده کاوی شامل استفاده از ابزارهای تجزیه و تحلیل اطلاعات پیچیده برای کشف الگوهای و روابط ناشناخته پیشین ، الگوهای معتبر و روابط موجود در مجموعه های داده ای بزرگ است. این ابزارها می توانند شامل مدلهای آماری، الگوریتم های ریاضی و روش های یادگیری ماشینی در تشخیص زود هنگام سرطان باشند. در طبقه بندی یادگیری ، طرح یادگیری با مجموعه ای از نمونه های طبقه بندی شده ارائه می شود که انتظار می رود روشی برای طبقه بندی نمونه های نادیده آموزش دهد. در یادگیری ارتباطی، هر گونه ارتباط بین ویژگی های مورد نظرمنظور است، نه فقط آنهایی که یک گروه ارزشی را پیش بینی می کنند. در خوشه بندی، گروهی از نمونه ها که متعلق به هم هستند، دنبال می شوند. در پیش بینی عددی، نتیجه پیش بینی شده یک گروه گسسته نیست، بلکه یک مقدار عددی است. در این پژوهش، برای طبقه بندی داده ها وکشف الگوهای مکرر درمجموعه داده ها از الگوریتم تصمیم گیری درختی استفاده می شود [۱۰].

ب) کاربرد داده کاوی

کاربرد داده کاوی در حال حاضر به طور فزاینده ای در زندگی روزمره قابل مشاهده است. روش های مختلفی که این تکنیک های داده کاوی بر روی آن اعمال شده ازاستخراج الگوهای جالب در منافع تجارت، سلامت و پزشکی وگرفته تا زمینه های آموزشی درذیل مورد بررسی قرار گرفته اند که عبارتند از:

• خرده فروشی و خدمات

داد وستد، تجارت و کارآفرینی بخش مهمی از توسعه را شکل می دهند. بیشتر اطلاعات مربوط به معاملات مرتبط با کسب و کار در بایگانی های داده ذخیره می شوند و هرگز برای اهداف پاکسازی یا تجزیه و تحلیل قابل دسترسی نیستند [8].

• پزشکی و مراقبت های بهداشتی

صرف نظر از موفقیت در عرصه کسب و کار و خرده فروشی، داده کاوی منعکس کننده مزایای خود در زمینه پزشکی و سلامت است. فرمولاسیون الگوریتمها به دلیل پیچیدگی مراقبتهای بهداشتی و کم شدن سرعت تطبیق تکنولوژی هنوز در مرحله بسیار ابتدایی قرار دارد [۱۰].

• تحصیلات تکمیلی

کاربردهای اخیر داده کاوی توسط موسسات آموزش عالی به دلیل افزایش تدریجی میزان داده ها در طول سال های اخیر انجام شده است. داده کاوی در این زمینه ها درراستای درک رفتار دانشجویان مورد استفاده قرار می گیرد ، مثال در روند هایی که نشان دهنده انتقالی دانشجویان، ساعات کاری مفید و همچنین مجموعه مهارت های مختلف دانشجویان و ویژگی های اضافی شخصیتی آنها است[8].

2-2- داده کاوی در فعالیت های پزشکی

داده کاوی در زمینه مراقبت های بهداشتی بسیار محبوب است زیرا در این زمینه نیازی شدید به روش تحلیلی کارآمد برای شناسایی اطلاعات ناشناخته و ارزشمند در داده های بهداشتی وجود دارد. در صنعت بهداشت، داده کاوی مزایای متعددی از قبیل تشخیص تقلب در بیمه درمانی، معرفی روش درمانی پزشکی برای بیماران با هزینه پایین، تشخیص علل بیماری ها و شناسایی روش های درمان پزشکی فراهم می کند. داده کاوی همچنین به محققان بهداشت و درمان برای ایجاد سیاست های کارآمد در این زمینه، ساخت سیستم های تجویزپزشکی و ایجاد پروفایل های بهداشتی افراد کمک می کند. پایگاه های داده ای پزشکی آنچنان بزرگند که به برنامه های کامپیوتری نیازمندند تا روند غایی را درتسهیل تشخیص و درمان پزشکی بیابند. با توجه به تکنیک های داده کاوی، به خصوص تکنیک های داده کاوی پزشکی، بخش مراقبت های بهداشتی پیشرفت های قابل توجهی در استفاده از فن آوری های پیشگیری و تشخیص بیماری ها به دست آورده است. انجمن پزشکی اینفورماتیک آمریکا سلامت اینفورماتیک را به عنوان "کلیه جنبه های درک و ترویج سازماندهی، تجزیه و تحلیل، مدیریت و استفاده از اطلاعات در مراقبت های بهداشتی" تعریف کرده است. فرایندهای داده کاوی شامل فرمول بندی فرضیه، جمع آوری داده ها، اعمال پیش پردازش، برآورد مدل و تفسیر مدل و نتیجه گیری است [11]. از آنجا که کار ما بررسی تشخیص سرطان است، بیماری سرطان را در بخش بعدی توضیح می دهیم.

2-3- بیماریهای سرطان

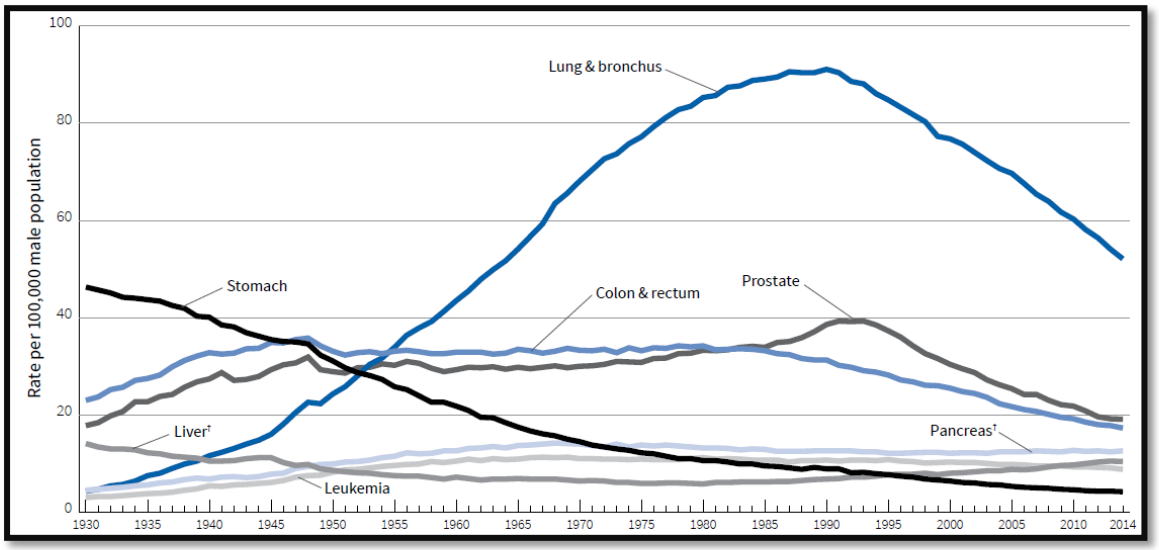
سرطان یک اصطلاح عمومی برای گروهی بزرگی از بیماری هاست که می توانند بر هر قسمت از بدن تاثیر بگذارند. سایر موارد تومورهای بدخیم و نئوپلاسم ها هستند. یکی از ویژگی های تعیین کننده سرطان، ایجاد سریع سلول های غیر طبیعی است که خارج از مرزهای معمول خود رشد می کنند و

پس از آن می توانند به قسمت های مجاور بدن نفوذ کرده و به اندام های دیگر راه یابند، از فرآیند دوم با عنوان متاستاز نام برده می شود. متاستاز علت اصلی مرگ بر اثر سرطان است. سرطان در بین علل عمده مرگ و میر در سراسر جهان است ، با حدود ۱۴ میلیون مورد جدید و ۸.۲ میلیون مورد مرگ و میر ناشی از سرطان در سال ۲۰۱۲. شایع ترین علل مرگ و میر ناشی از سرطان عبارتند از: ریه (۱.۵۹ میلیون مرگ و میر)، کبد (۷۴۵۰۰۰ مرگ و میر) ، معده (۷۲۳۰۰۰ مرگ و میر) و کولورکتال (۶۹۴۰۰۰ مرگ و میر) و سینه (۵۲۱۰۰۰ مرگ و میر) و سرطان مری (۴۰۰۰۰۰ مرگ و میر) است. استفاده از تنباکو، اضافه وزن و چاقی، رژیم غذایی ناسالم با مصرف محدود میوه و سبزی، عدم فعالیت بدنی، مصرف الکل، آلودگی هوای شهری، دود و دم ناشی از استفاده خانگی سوخت جامد، عامل اصلی خطر سرطان است[12].

سرطان بیماری است با ویژگی رشد و گسترش کنترل نشده سلول های غیر طبیعی و قابلیت حمله این سلول ها به بافت های دیگر که ممکن است به واسطه عوامل خارجی چون اشعه، مواد شیمیایی، تنباکو و ... و عوامل داخلی نظیر جهش های ارثی، هورمون ها، شرایط ایمنی، و ... باشد. بیش از ۱۰۰ نوع مختلف از سرطان وجود دارد. اکثر سرطان ها با نام ارگان ها یا نوع سلولهایی که در آنها ظاهر می شوند،نامگذاری می شوند مانند ملانوم، کولون، سرطان سینه و ... [۱۳]. تمام سرطان ها درون سلول هایی که واحد های ساختاری و عملکردی بدن هستند شروع می شوند. این سلول ها رشد می کنند و به صورت کنترل شده ای تقسیم می شوند و سلول های بیشتری تولید می شود، زیرا بدن آنها سالم است. وقتی سلول ها قدیمی یا آسیب دیده اند، می میرند و با سلول های جدید جایگزین می شوند. با این حال، گاهی اوقات چرخه زندگی سلول ها به دلایل بسیاری از بین می رود و یا دچار اختلال می شود . هنگامی که این اتفاق می افتد، سلول ها برخلاف انتظار نمیمیرند و سلول های جدید حتی زمانی که بدن به آنها نیازی ندارد تشکیل می شوند. این سلولهای اضافی ممکن است یک توده بافتی

به نام تومور ایجاد کنند. تومورها می توانند خوش خیم یا بدخیم باشند. برخی از سرطان ها تومورتشکیل نمی دهند. به عنوان مثال، لوسمی یک سرطان خون است که تومور تشکیل نمی دهد [۱۳].

روند کاهش نرخ مرگ و میر سرطان بهترین معیار پیشرفت در برابر سرطان است.نرخ مرگ و میر کلی سرطان در طول قرن بیستم به علت اپیدمی تنفسی افزایش یافت و در سال ۱۹۹۱، ۲۱۵ مرگ و میر ناشی از سرطان به ازای هر ۱۰۰۰۰۰ نفر افزایش یافت. با این حال، از سال ۲۰۱۴ این میزان به ۱۶۱ در هر ۹/ ۱۰۰,۰۰۰ نفر کاهش یافت،(کاهشی ۲۵ درصدی) این کاهش به دلیل کاهش مصرف سیگار، و همچنین بهبود در تشخیص زود هنگام و روند درمان بیماری بود. این کاهش منجربه کمتر شدن بیش از ۲.۱ میلیون مرگ و میر ناشی از سرطان در طی دو دهه گذشته شده است، پیشرفتی که به دلیل کاهش سریع میزان مرگ و میر در چهار گونه از رایج ترین انواع سرطان (ریه، کولورکتال، سینه و پروستات) بوده است (شکل ۲-۲) [14].



شکل 2- 2- روند تغییرات مرگ و میر ناشی از سابقه سرطان [14]

2-4- تکنیک های یادگیری ماشینی (ML)

ML[1]، شاخه ای از هوش مصنوعی به شمار می آید که به مفهوم کلی مشکل یادگیری نمونه های داده، اشاره دارد. هر فرایند یادگیری شامل دو مرحله است [15]: (1) تخمین وابستگی های ناشناس در سیستم مجموعه داده های ارائه شده (2) استفاده از وابستگی های برآورد شده به منظور پیش بینی خروجی های جدید سیستم.

ML همچنین منطقه ای جالبی در تحقیقات زیست پزشکی همراه با بسیاری از برنامه های کاربردی می باشد که در آن، با جستجوی n بعدی برای یک مجموعه از داده های بیولوژیکی با کمک تکنیک ها و الگوریتم های متفاوت، به تعمیم پذیری قابل قبولی دست می یابد. متداول ترین شیوه ML شناخته شده تحت عنوان (1) یادگیری تحت نظارت و (2) یادگیری بدون نظارت است. در یادگیری تحت نظارت، از مجموعه داده های آموزشی برچسب گذاری شده برای تخمین یا ارائه داده های ورودی به خروجی مورد نظر، استفاده می شود. در عوض، در شیوه یادگیری بدون نظارت، هیچ نمونه برچسب زده ای ارائه نمی شود و در طول فرآیند یادگیری هیچ تصوری از خروجی وجود ندارد. در نتیجه، این برای شیوه یا مدل یادگیری جهت پیدا کردن الگوها یا کشف گروه های داده های ورودی است. در یادگیری تحت نظارت، این رویه را می توان به عنوان یک مشکل طبقه بندی کرد. طبقه بندی به فرایند یادگیری اشاره می کند که داده ها را به مجموعه طبقه های محدود تقسیم می کند. دو تا از دیگر وظایف ML معمولی، رگرسیون و خوشه بندی است. در مورد مشکلات رگرسیون، یک تابع یادگیری داده ها را برای یک متغیر ارزش واقعی ، ترسیم می کند. سپس، برای هر نمونه جدید، ارزش یک متغیر پیش بینی را می توان براساس این فرآیند، برآورد کرد. خوشه بندی یک وظیفه بدون نظارت معمول است که در آن برای یافتن دسته ها یا خوشه ها جهت توصیف اقلام، تلاش می

[1] Machine learning

شود. بر اساس این فرآیند هر نمونه جدید می تواند به یکی از خوشه های شناسایی مربوط به ویژگی های مشترک که آنها به اشتراک می گذارند، اختصاص داده شود. فرض کنید به عنوان مثال پرونده پزشکی مربوط به سرطان سینه را جمع آوری کرده و می خواهیم پیش بینی کنیم که آیا تومور بر اساس اندازه آن بدخیم یا خوش خیم است یا نه. سوال ML به ارزیابی احتمال تومور بدخیم یا نه (1 = بله، 0 = نه) اشاره می شود. شکل 2-3، فرایند طبقه بندی بدخیم بودن یا نبودن تومور نشان می دهد .

شکل 2-3- وظیفه طبقه بندی در یادگیری نظارت [15].

تومورها به صورت X نشان داده شده و تحت عنوان خوش خیم یا بدخیم طبقه بندی می شوند. نمونه های دایره ای، تومورهایی را که به اشتباه طبقه بندی شده اند، نشان می دهد .

سوابق گردی شکل نشان دهنده طبقه بندی نامناسب نوع تومور معرفی شده توسط این روش است. نوع دیگری از روش های ML که به طور وسیعی مورد استفاده قرار گرفته است، یادگیری شبه نظارتی است که ترکیبی از یادگیری نظارت شده و بدون نظارت می باشد. این شیوه داده های برچسب دار و

بدون برچسب را برای ساخت یک مدل یادگیری دقیق، با هم آمیخته می کند. معمولا این نوع یادگیری زمانی مورد استفاده قرار می گیرد که مجموعه داده های بدون برچسب از برچسب زده ها، بیشتر باشند. هنگام استفاده از شیوه ML، نمونه های داده ها جزء اساسی را تشکیل می دهند. هر نمونه همراه با چند ویژگی شرح داده شده و هر ویژگی شامل انواع مختلفی از ارزش ها می باشد. علاوه بر این، دانش مورد استفاده در نوع خاصی از داده ها ، امکان انتخاب مناسب ابزار و تکنیک هایی را فراهم می کند که می تواند برای تجزیه و تحلیل آن ها استفاده شوند . برخی از مسائل مربوط به داده ها به کیفیت داده ها و مراحل پیش پردازش مربوط هستند تا آنها را برای ML مناسب تر سازند. مسائل مربوط به کیفیت داده شامل سر و صدا، ناپایداری ها، داده های گم شده و یا تکراری و داده هایی می باشد که نمایش گر نیستند. با بهبود کیفیت داده ها، معمولا کیفیت تجزیه و تحلیل حاصل نیز افزایش یافته است. علاوه بر این، برای ایجاد داده های خام مناسب تر به منظور تجزیه و تحلیل بیشتر، مراحل پیش پردازشی باید اعمال شوند که بر اصلاح داده ها متمرکز می باشند. برخی از تکنیک ها و استراتژی های مختلف در این زمینه وجود دارند که مربوط به پیش پردازش اطلاعات است که تمرکز آن بر تغییر داده ها جهت سازگاری بهتر در یک روش ML خاص می باشد. برخی از مهم ترین این شیوه ها عبارتند از: کاهش ابعاد ، انتخاب ویژگی و استخراج ویژگی. کاهش ابعاد مزایای زیادی به همراه دارد که مجموعه داده ها از ویژگی بسیاری برخوردار هستند. الگوریتم ML زمانی بهتر است که ابعاد پایین تر هستند. علاوه بر این، کاهش ابعاد می تواند سبب از بین رفتن ویژگی های نامناسب شده، سر و صدا را کاهش داده و هم چنین می تواند الگوهای یادگیری قوی تری را به علت دخالت ویژگی های کمتر، تولید کند. به طور کلی، کاهش ابعاد با انتخاب ویژگی های جدید که یک زیر مجموعه آن ها قدیمی است، به عنوان انتخاب ویژگی، شناخته می شود. انتخاب ویژگی دارای سه شیوه اصلی است، که عبارت است از رویکرد های جاسازی شده، فیلتر شده و بسته بندی.

در مورد استخراج ویژگی، یک مجموعه جدید از ویژگی ها را می توان از مجموعه اولیه به وجود آورد که کلیه اطلاعات قابل قابل ملاحظه ای را در مجموعه داده، ایجاد می کند. ایجاد مجموعه های جدید از ویژگی ها این امکان را فراهم می سازد تا مزایای شرح داده شده درباره کاهش ابعاد جمع اوری کند. با این حال، استفاده از تکنیک های انتخاب ویژگی ممکن است موجب نوسان خاص در ایجاد لیست های پیش بینی شده ، شود. مطالعات متعددی در ادبیات ، پدیده عدم توافق بین لیست ژن های پیش بینی شده توسط گروه های مختلف الزام هزاران نمونه برای دستیابی به نتایج مطلوب، عدم تفسیر بیولوژیکی از امضاهای پیش بینی شده و خطرات فاش سازی اطلاعات، را توضیح می دهند. هدف اصلی تکنیک های ML تولید مدلی است که قادر است جهت انجام طبقه بندی، پیش بینی، برآورد یا هر کار مشابه دیگر مورد استفاده قرار گیرد. رایج ترین وظیفه در فرایند یادگیری طبقه بندی، است. همانطور که قبلا گفته شد، این تابع یادگیری آیتم داده را به یکی از چندین طبقه از پیش تعریف شده، دسته بندی می کند.

هنگامی که یک مدل طبقه بندی توسعه داده می شود، با استفاده از تکنیک های ML، آموزش ها و اشتباهات عمومی می توانند ،تولید شوند. مدل قبلی خطاهای اشتباه در داده های آموزشی و مدل بعدی به اشتباه های مورد انتظار در داده های آزمون، اشاره می کند. یک مدل طبقه بندی خوب باید به خوبی با مجموعه آموزشی تنظیم شده و تمام موارد را به درستی طبقه بندی کرده باشد. اگر میزان خطای آزمون یک مدل، رو به افزایش باشد، حتی اگر میزان اشتباهات آموزش کاهش یابد، پدیده تغییرات در مدل، رخ می دهد. این وضعیت به پیچیدگی مدل مربوط می شود یعنی این که اگر پیچیدگی مدل افزایش یابد، خطاهای آموزش یک مدل ممکن است کاهش یابند. آشکار است که پیچیدگی ایده آل از یک مدل غیر حساس به بیش از حد باعث تولدی کمترین خطای عمومی میشود. تجزیه واریانس تعادلی ،روشی رسمی برای تجزیه و تحلیل خطای بهینه سازی مورد انتظار

یک الگوریتم یادگیری می باشد. جزء تعصب الگوریتم یادگیری خاص، میزان اشتباه این الگوریتم را اندازه گیری می کند. علاوه بر این، یک منبع ثانویه خطا بر روی تمامی مجموعه های آموزشی ممکن است از اندازه داده شده و تمام مجموعه های ممکن آزمون، واریانس روش یادگیری نامیده می شود. خطای کلی مورد انتظار یک مدل طبقه بندی از مجموع تعصب و واریانس، تجزیه واریانس تعادلی می باشد. هنگامی که یک مدل طبقه بندی با استفاده از یک یا چند تکنیک ML به دست می آید، باید عملکرد طبقه بندی را ارزیابی کنید. تجزیه و تحلیل عملکرد هر مدل پیشنهادی بر اساس حساسیت، ویژگی، دقت و سطح زیر منحنی (AUC) ،اندازه گیری می شود. حساسیت به عنوان نسبت مثبت واقعی تعریف شده که توسط طبقه بندی به طور صحیح مشاهده می شود، در حالی که خصوصیات مربوط به نسبت منفی های واقعی که به طور صحیح شناسایی می شوند، تعریف می شوند. AUC برای ارزیابی عملکرد کلی یک طبقه بندی، معیارهای کمی دقت مورد استفاده قرار می گیرد. به طور ویزه ای، دقت، اندازه گیری مربوط به تعداد کل پیش بینی های درست است. برعکس، AUC یک اندازه از عملکرد مدل محسوب می شود که بر اساس منحنی ROC می باشد که توافق بین حساسیت و 1-خاصیت (شکل 2) را ترسیم می کند. دقت پیش بینی مدل از مجموعه تست محاسبه شده است که نشان دنده برآوردی از خطاهای تعمیم است. برای دستیابی به نتایج قابل اعتماد در مورد پیش بینی عملکرد یک مدل، نمونه های آموزش و آزمایش باید به اندازه کافی بزرگ و مستقل باشند در حالی که برچسب مجموعه های آزمایش، باید شناخته شود. در میان رایج ترین روش های ارزیابی عملکرد یک طبقه بندی با تقسیم داده های اولیه برچسب شده به زیر مجموعه ها، روش های زیر آورده شده اند:1) روش Holdout، 2) نمونه گیری تصادفی، 3) اعتبارسنجی متقابل و 4) بوت استرپ(Bootstrap)

در روش Holdout، نمونه های داده ها به دو مجموعه جداگانه، یعنی آموزش و مجموعه های تست ، تقسیم می شوند. سپس یک مدل طبقه بندی از مجموعه آموزش آماده شده، در حالی که عملکرد آن بر روی مجموعه آزمون تعیین می شود. نمونه گیری تصادفی یک رویکرد مشابه به شیوه Holdout است. در این حالت، به منظور تخمین دقت تر، روش Holdout چندین بار تکرار می شود و نمونه های آموزش و آزمایش بصورت تصادفی، انتخاب می شوند. در رویکرد سوم، یعنی اعتبارسنجی متقابل، هر نمونه از همان تعداد دفعات برای آموزش و فقط یک بار برای آزمایش استفاده می شود. در نتیجه، مجموعه داده های اصلی با هم دیگر در آموزش و در مجموعه تست موفقیت آمیز هستند. نتایج دقت به عنوان میانگین هر چرخه معتبر مختلف، محاسبه می شود. در اخر، در روش بوت استرپ، نمونه ها با جایگزینی به آموزش و مجموعه های آزمایشی، جدا می شوند، به عنوان مثال آنها دوباره پس از این که برای آموزش انتخاب شدند، در کل مجموعه داده ها قرار می گیرند. هنگامی که داده ها پردازش می شوند و نوع کار یادگیری را تعریف می کنیم، لیستی از روش های ML شاملANNs ، DTs، SVMs و BNs را در اختیار داریم.

بر اساس این فرضیه، ما تنها به این تکنیک های ML که در مطالعه فعلی به پیش بینی مرضی سرطان پرداخته شده است، اشاره می کنیم. ما روندهای مربوط به انواع روش های ML مورد استفاده، انواع داده های یکپارچه و همچنین روش های ارزیابی مورد استفاده به منظور ارزیابی عملکرد کلی روش های پیش بینی سرطان یا نتایج بیماری ها را شناسایی می کنیم.

ANNs -2-4-1

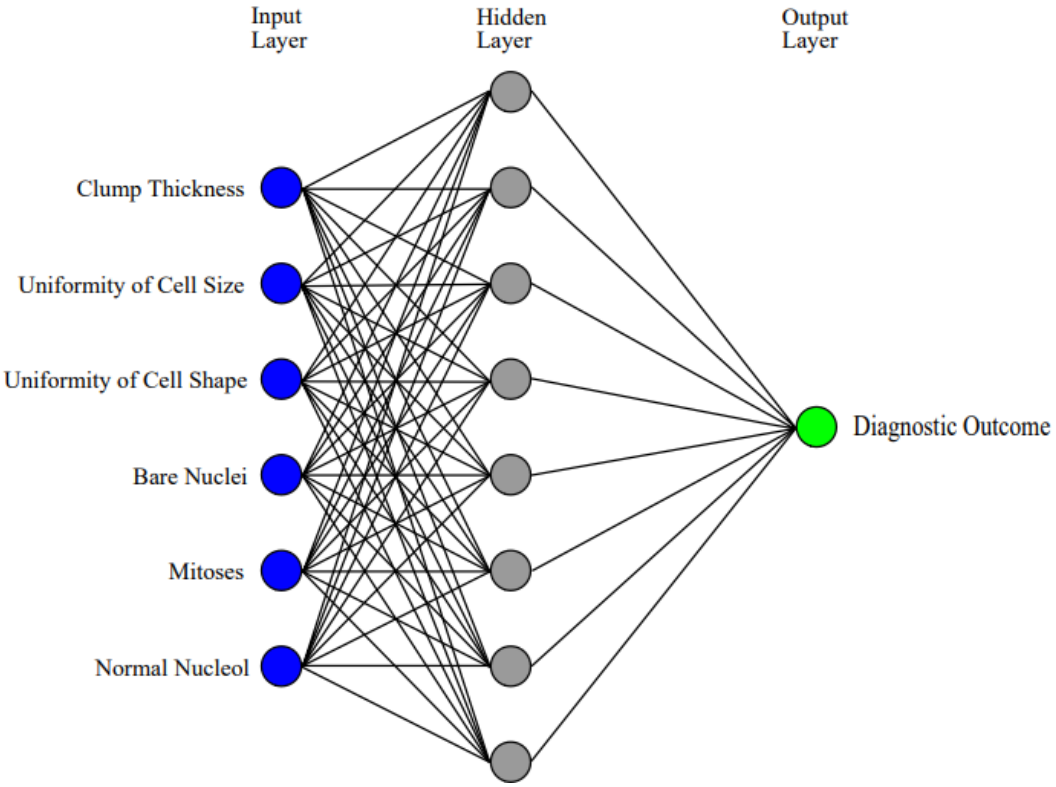
شبکه های ANNs با تنوعی از مشکلات طبقه بندی یا تشخیص الگو سر و کار دارند. آنها برای تولید خروجی به عنوان ترکیبی از متغیرهای ورودی، آموزش دیده اند. لایه های پنهان چندگانه که اتصالات عصبی را به صورت ریاضی را نشان می دهند، معمولا برای این فرآیند استفاده می شوند. هرچند

ANN ها به عنوان روش استاندارد طلایی در چندین وظیفه طبقه بندی، از نقایص خاصی رنج می برند. ساختار لایه بندی عمومی آنها زمان مصرف را اثبات می کند در حالی که می تواند عملکرد بسیار ضعیفی داشته باشد. علاوه بر این، این تکنیک خاص به عنوان یک فناوری " جعبه سیاه" شناخته می شود. شناسایی نحوه انجام فرایند طبقه بندی یا اینکه چرا ANN جواب نداده، تقریبا غیرممکن است. شکل 2-4 ساختار ANN را با گروه متصل به گره نشان می دهد. لایه ها از نورون های متصل شده تشکیل شده اند که شامل تابع فعال سازی جهت تبدیل غیرخطی برای تقویت توانایی بیان غیرخطی است.

لایه ورودی داده ها را دریافت می کند و سپس داده ها را به یک لایه پنهان منتقل می کند که جهت پردازش داده ها و ارائه نتایج آموزشی به لایه خروجی، مورد استفاده قرار می گیرد. لایه خروجی نشان دهنده نتایج طبقه بندی است. با این حال، طبق به مشکلات، روند آموزش ANN ممکن است شامل زنجیره های طولانی از مراحل محاسباتی باشد. از سال 1986، یک الگوریتم شبیه سازی شده به نام Repropagation (BP) برنامه های کاربردی گسترده ای دارد، مخصوصا برای داده های پزشکی.

این امر با تعمیم قانون یادگیری Widrow-Hoff به شبکه های چند لایه و توابع انتقال غیرخطی سبب ایجاد تمایز شد. اگرچه از الگوریتم BP[2] استفاده می شود، اما هنوز این الگوریتم در ارائه اطلاعات گسترده و پیچیده، با ضعف هایی همراه است. محاسبات BP گسترده و در نتیجه، آموزش آهسته هستند، بنابراین الگوریتم BP خالص به ندرت در کاربردهای عملی مورد استفاده قرار می گیرد. محققان در حال تلاش جهت بهبود الگوریتم BP برای افزایش کارایی محاسباتی هستند. در مورد الگوریتم های مختلف ریاضی مبتنی بر ANN که برای تجزیه و تحلیل WBCD مورد استفاده قرار گرفته اند، در بخش های بعدی بحث می شود [16].

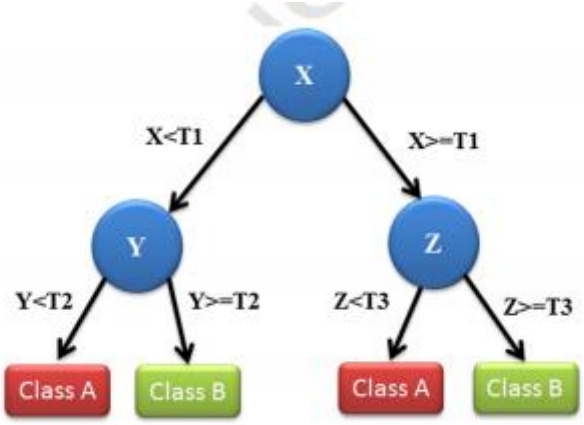
[2] Backpropagation



شکل 2-4- یک مثال ساده از نحوه آموزش ANN برای پیش بینی نتایج تشخیصی از شش ورودی و یک لایه پنهان با 8 نورون آموزش داده شده [16]

2-4-2- درخت تصمیم

درخت تصتمیم یک طرح طبقه بندی درختی را دنبال می کند که گره ها نشان دهنده متغیرهای ورودی هستند و برگ ها مطابق نتایج تصمیم هستند. درخت های تصمیم از اولین و برجسته ترین روش های ML هستند که به طور گسترده ای برای مقاصد طبقه بندی ، استفاده می شوند. بر اساس معماری درخت تصمیم، آنها برای تفسیر ساده هستند و جهت یادگیری، سریع می باشند. وقتی که برای طبقه بندی یک نمونه جدید از نمودار درختی استفاده می کنیم ، قادریم طبقه اش را حدس بزنیم. تصمیمات حاصل از معماری خاص آنها امکان استدلال کافی را برای جذاب کردن آن تکنیک، فراهم می سازد. شکل 2-5 تصویری از درخت تصمیم با عناصر و قوانین آن را نشان می دهد.



شکل 2-5- یک تصویر درخت تصمیم نشان دهنده ساختار درختی است[16].

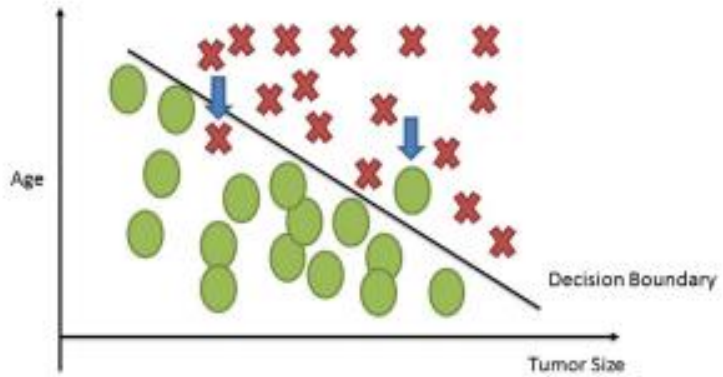
هر متغیر (X، Y، Z) توسط یک دایره و نتایج تصمیم گیری توسط مربع (طبقه A، طبقه B) نشان داده می شود. T(1–3) آستانه ها (قوانین طبقه بندی) را نشان می دهد تا به طور موفقیت آمیزی هر متغیر را به یک برچسب طبقه، دسته بندی کند.

3-4-2- طبقه بند SVM

یک ماشین بردار پشتیبانی (SVM)، یک مفهوم در آمار و علوم رایانه ای برای مجموعه ای از روش های یادگیری تحت نظارت است که داده ها را پردازش و الگوهای مورد استفاده برای طبقه بندی و تجزیه و تحلیل رگرسیون را تشخیص می دهد. طبقه بند SVM استاندارد، مجموعه ای از داده های ورودی را برای هر ورودی داده شده و هر کدام از دو طبقه ممکن که ورودی را تشکیل می دهند، می گیرد و SVM را یک طبقه بندی خطی جفتی غیر احتمالی می کنند [17].

طبقه بندهای SVM یک رویکرد جدیدتر از روش های ML در زمینه پیشگیری و پیش بینی مرضی سرطان هستند. در ابتدا طبقه بند SVM، ورودی را به یک فضای مشخصه از ابعاد بالاتر ترسیم می کند و شبه صفحه ای را شناسایی می کند که نقاط داده را به دو دسته جدا می کند. فاصله مرزی بین شبه صفحه تصمیم گیری و نمونه هایی که نزدیک ترین به مرز هستند ، به حداکثر می رسد. طبقه بندی نهایی به طور قابل ملاحظه ای عمومی می شود و بنابراین می تواند برای طبقه بندی معتبر

نمونه های جدید، مورد استفاده قرار گیرد. باید ذکر شود که خروجی های احتمالی نیز برای طبقه بندهای SVM به دست می آیند. شکل 2-6 نشان می دهد که چگونه یک طبقه بند SVM ممکن است به منظور طبقه بندی تومور ها بین خوشخیم و بدخیم بر اساس اندازه و سن بیمار آنها، کار کند.

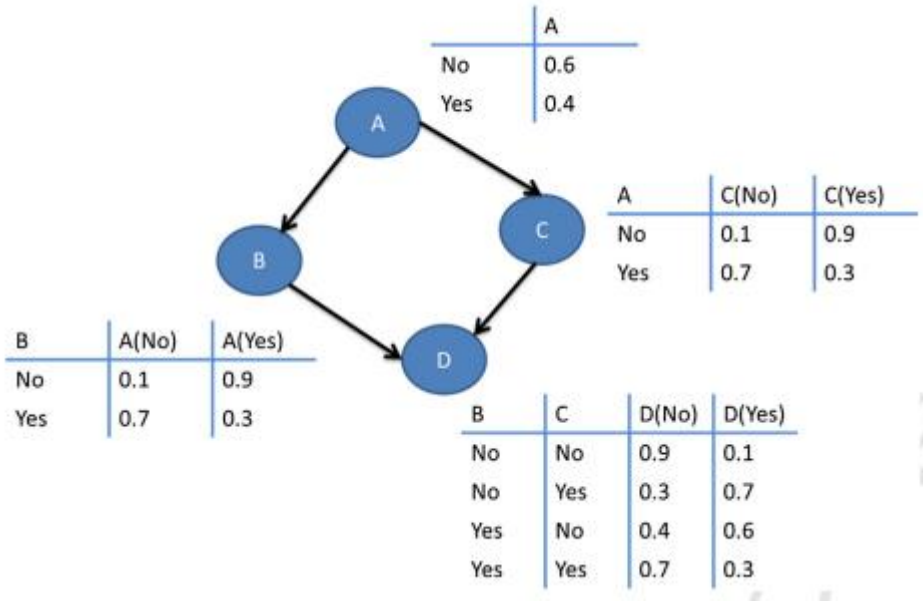


شکل 2- 6- یک تصویر ساده از طبقه بندی SVM خطی داده های ورودی [17]

تومورها بر اساس اندازه و سن بیمار طبقه بندی می شوند. فلش های نشان دهنده تومورهای طبقه بندی نشده هستند.

2-4-2- طبقه بندی شبکه بیزین

شبه صفحه شناسایی شده می تواند به عنوان یک مرز تصمیم گیری میان دو خوشه باشد، به طور آشکاری، وجود یک مرز تصمیم برای تشخیص هر طبقه بندی اشتباه تولید شده توسط این روش، امکان پذیر است. طبقه بندی های شبکه بیزین به جای پیش بینی ها، تخمین های احتمالی را، تولید می کنند. همانطور که که از نام آن ها پیداست، آنها برای نشان دادن دانش مرتبط با وابستگی احتمالی در میان متغیرهای مورد علاقه از طریق یک نمودار دوره یا یا چرخشی، استفاده می شوند. شبکه های بیزین به طور گسترده ای به چند وظیفه طبقه بندی شده اند و همچنین برای ارائه دانش و اهداف استدلال، اعمال می شوند. شکل 2-7 تصویری از شبکه بیزین در سراسر احتمال شرطی محاسبه شده برای هر متغیر را نشان می دهد.



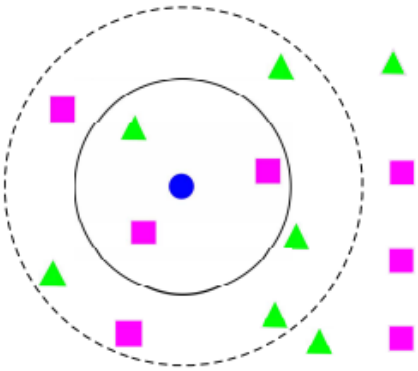
شکل 2-7- یک تصویر از شبکه بیزین

گره ها (A-D) مجموعه ای از متغیرهای تصادفی را با احتمال شرطی آنها که در هر جدول محاسبه می شوند، نشان می دهند.

k-NN -2-4-5

k-NN یکی از تکنیک های مرکزی ML در طبقه بندی است. k-NN یک الگوریتم یادگیری تنبل غیر پارامتری است که برای طبقه بندی استفاده می شود که اشیا را با استفاده از نزدیکترین همسایگان "k" طبقه بندی می کند. k-NN تنها همسایگان اطراف جسم را در نظر می گیرد، نه توزیع داده های پایه.

علاوه بر این، مرحله آموزش با داده های آموزشی وجود ندارد. در شکل 2-8 یک مثال از ساختار k-NN برای تعیین تشخیص BC و پیش گیری در صورت k = 3 ارائه شده است، نمونه آزمون (دایره) برای بدخیمی BC (مربع)اختصاص داده می شود ، به دلیل اینکه 2 مربع و تنها 1 مثلث در داخل دایره وجود دارد. اگر k = 5 است، نمونه آزمون به BC خوش خیم(مثلث) اختصاص داده می شود.



شکل 2- -8نزدیک ترین همسایه برای تشخیص سرطان سینه[17]

دایره آبی به معنی نمونه آزمون، مثلث سبز به معنای BC بدخیم و مربع صورتی به معنی BC خوشخیم است.

2-5-کارهای انجام شده در زمینه پیش بینی سرطان

داده کاوی در اواسط دهه 1990 آغاز شد و به عنوان یک ابزار قدرتمند برای استخراج الگوی ناشناخته و اطلاعات مفید مجموعه داده های بزرگ، شناخته شده است. مطالعات مختلف حاکی از آن است که تکنیک های داده کاوی به مالک داده ها این امکان را می دهد تا رابطه نامشخص در میان داده هایشان را با تجزیه و تحلیل، کشف کنند که در نتیجه می تواند برای تصمیم گیری مفید و موثر باشد [18]. شیوه های داده کاوی جهت تجزیه و تحلیل این مجموعه غنی از داده های دیدگاه های مختلف و یافتن اطلاعات مفید، به کار گرفته می شود. مقایسه‌ای از کارهای قبلی در زمینه پیش بینی سرطان در جدول 3-1 نشان داده شده است.

در مقاله [19] یک سیستم پیش بینی مبتنی بر داده کاوی انجام گرفته است. هدف اصلی این مدل ارائه هشدار قبلی به کاربران است، از این رو، از جهت هزینه و زمان نیز به سود کاربر است. سه خطر ویژه، سرطان را پیش بینی می کند. به طور ویژه ای، سیستم پیش بینی ،خطر ابتلا به سرطان سینه، پوست و ریه را با بررسی تعدادی از عوامل ژنتیکی و غیر ژنتیکی ارائه شده توسط کاربر، ارزیابی می

کند. این سیستم با مقایسه نتایج پیش بینی شده آن، با پرونده پزشکی قبلی بیمار معتبر است و همچنین با استفاده از سیستم سنجش آن،تجزیه تحلیل می شود. این سیستم پیش بینی در اینترنت در دسترس است، افراد به راحتی می توانند خطر خود را بررسی و اقدامات مناسب را براساس وضعیت خطر آنها، انجام دهند. این سیستم عملکرد خوبی نسبت به سیستم موجود دارد.

در این کار، سیستم پیش بینی مبتنی بر داده های معماری با استفاده از سیستم پیش بینی شده با تکنولوژی داده کاوی، مورد استفاده قرار گرفت. در این مدل نویسندگان از یکی از الگوریتم های طبقه بندی به نام درخت تصمیم گیری، استفاده کرده اند.

هنگامی که کاربر وارد سیستم پیش بینی سرطان می شود، آنها باید به پرسش های مربوط به عوامل ژنتیکی و غیر ژنتیکی پاسخ دهند. سپس سیستم پیش بینی مقدار خطر را برای هر سوال طبق پاسخ کاربر، تعیین می کند. هنگامی که ارزش خطر پیش بینی می شود، محدوده خطر را می توان با سیستم پیش بینی، تعیین کرد. نویسندگان از چهار سطح خطر پایین ، متوسط، سطح بالای و بسیار بالا برخوردار هستند. محدوده خطر، بر اساس ارزش های پیش بینی شده خطر، تعیین می شود. نتیجه می تواند به وسیله پایگاه داده به کاربر نشان داده شود.

در مقاله [20]، از سه الگوریتم یادگیری ماشین استفاده شده است تا به طور خودکار توده های بیماری را طبقه بندی کنند: جنگل تصادفی، ماشین پشتیبانی بردار و بیز ساده و سه تکنیک یادگیری ماشین را به پایگاه داده سرطان سینه ویسکانسین اعمال می کند. جنگل تصادفی به روش های جعبه سازی و تصادفی زیرمجموعه بستگی دارد، SVM یک طبقه بندی بسیار محتاطانه است که اگر تنظیم دقیق پارامترهای آن به خوبی انجام شده باشد، اغلب به عنوان بهترین طبقه بندی کننده، ، امکان پذیر است و بیز ساده یک طبقه بندی ساده است که پیش بینی ها را مستقل فرض می کند. این سه الگوریتم با استفاده از همان مجموعه های آموزشی و تست مقایسه می شوند و عملکرد هر یک از آنها

با اندازه گیری ماتریس پریشانی، اندازه گیری می شود. سه مدل توسعه یافته پیش بینی می کنند که آیا آسیب های بیمار، خوش خیم است یا بدخیم. هدف مقاله، مقایسه عملکرد این سه الگوریتم از طریق صحت، دقت، فراخوانی و اندازه گیری f است. نتایج نشان می دهد که جنگل تصادفی بهترین نتیجه را با 99.42٪ بدست می آورد، که کمی بهتر از SVM و بیز ساده است که دارای اطمینان 98.8٪ و 98.24٪ می باشند. این نتایج بسیار رقابتی هستند و می توانند برای تشخیص، پیش آگهی و درمان استفاده شوند.

در مقاله [21]، فرمول یک سیستم تشخیص کامپیوتری (CAD) را برای تشخیص سرطان ریه، با استفاده از یک رویکرد بین رشته ای بر اساس تکنیک های پردازش تصویر و یادگیری ماشین، مورد بحث قرار می دهد. این مقاله یک فرایند پردازش تصویر با استفاده از تشخیص سرطان ریه است و نتایج حاصل از استخراج ویژگی و انتخاب ویژگی پس از تقسیم بندی را، تولید می کند. در این جا، مدل پیشنهادی با استفاده از الگوریتم SVM به منظور انتخاب و طبقه بندی ویژگی ها، طراحی شده است. این سیستم تصاویر رادیویی (Tomography) CT را به عنوان ورودی می پذیرد. این کار حاضر، روشی را برای شناسایی سلول های سرطانی به طور موثر توسط CT اسکن و تصاویر پشینهاد می کند. برای محاسبه تقسیم بندی، از ابزارهای C-Measurement Fuzzy Modified (MFPCM) استفاده شده است و برای فیلتر کردن تصاویر پزشکی ، فیلتر گابور مورد استفاده قرار گرفته است. نتایج شبیه سازی برای سیستم تشخیص سرطان با استفاده از نرم افزار MATLAB بدست می آید. نتایج دقت سیستم CAD پیشنهاد شده با هسته های مختلف SVM ، در مقاله نشان داده شده است که دقت استفاده از هسته شعاعی (RBF) نسبت به دیگر هسته SVM بهتر می باشد. همانطور که جدول نشان می دهد، دقت متوسط هسته های دیگر برای تصاویر ریه 10-10 برابر با 89/ 5349 درصد است، در حالی که استفاده از RBF با دقت 98.4496 درصد انجام می شود؛ که بهتر از سایر شیوه ها است.

برای تصاویر ریه 11-20، RBF ،دقت و صحت 96.124٪ را تولید می کند، در حالیکه سایر تصاویر فقط دقت 89.1473٪ را تولید می کنند. دقت کلی برای RBF ، بیشتر از سایر تکنیک های SVM است.

در مقاله [22] یک ارزیابی تحلیلی از برخی از الگوریتم های انتخاب شده ماشین یادگیری بر روی مجموعه داده های سرطان سینه با استفاده از ابزار منبع باز WEKA انجام شد. برخی از پیش پردازش ها نیز بر روی داده های ورودی ، با استفاده از WEKA خاص در فیلترهای ساخت، انجام شده است و همچنین تاثیر کلی آن بر دقت پیش بینی نیز مشخص شد. نتایج نشان داد که آنالیز جنگل تصادفی از درختان تصمیم گیری با دقت 69٪ قبل از اعمال فیلتر و 98٪ پس از اعمال فیلتر، بهترین دقت را با استفاده از فیلترها دارد. به طور مشابه، رگرسیون لجستیک رتبه دوم را با 96٪ و بدون فیلتر 68٪ بدست آورد و در نهایت، نایوبیز با 91٪ و بدون فیلتر 71٪ بود. در سایر مجموعه های داده های آینده نیز می توان با استفاده از روش های مختلف الگوریتم کاوش داده، تجزیه و تحلیل کرد. همچنین انتخاب بهترین ویژگی ها، می تواند دقت پیش بینی را افزایش دهد و علاوه بر افزایش سرعت، می توان از انتخاب ویژگی دقت نیز، استفاده کرد.

در مقاله [23] یک سیستم به نام سیستم پیش بینی مبتنی بر داده کاوی انجام گرفته است که پیش بینی سه نوع خطر خاص سرطان (پستان، پوست و سرطان ریه) می باشد. به طور ویژه ای، ابزار پیش بینی سرطان، خطر ابتلا به سرطان سینه را با بررسی تعدادی از عوامل ژنتیکی و غیر ژنتیکی ارائه شده توسط کاربر، تخمین می زند. معماری این سیستم پیش بینی مبتنی بر روش داده کاوی، ترکیبی از سیستم پیش بینی با تکنولوژی استخراج است. در این مدل، آنها از یک الگوریتم طبقه بندی به نام درخت تصمیم گیری استفاده می کنند. این ابزار با مقایسهٔ نتایج پیش بینی شده با پرونده پزشکی قبلی بیمار و همچنین با استفاده از ابزار Weka مورد تجزیه و تحلیل قرار می گیرد.

هنگامی که کاربر وارد سیستم پیش بینی سرطان می شود، آنها باید به پرسش های مربوط به عوامل ژنتیکی و غیر ژنتیکی پاسخ دهند. سپس سیستم پیش بینی مقدار خطر را برای هر سوال بر اساس پاسخ کاربر تعیین می کند. هنگامی که ارزش خطر پیش بینی می شود، محدوده خطر را می توان با سیستم پیش بینی تعیین کرد. آنها دارای چهار سطح خطر پایین، سطح متوسط، سطح بالا و بسیار بالا هستند. بر اساس ارزش های پیش بینی شده خطر، محدوده خطر تعیین می شود. نتیجه می تواند به وسیله پایگاه داده به کاربر نشان داده شود. روش فوق می تواند به طور موفقیت آمیزی برای مجموعه داده های سرطان سینه مورد استفاده قرار گیرد، زیرا با موفقیت بر سرطان سینه تأیید شد. در نهایت این سیستم پیش بینی معتبر است که از طریق یک ابزار کمکی، دقت بیشتری را در مقایسه با سیستم موجود، فراهم می کند. هدف اصلی این مدل برای ارائه هشدار قبلی به کاربران است و همچنین هزینه و زمان به سود کاربر می باشد.

در مقاله [24] به ساخت یک ساختار سازماندهی نرم افزار مبتنی بر نرم افزار (SOM) اشاره شد که برای کشف الگوهای پنهان در تصاویر سیگنال اختلال ریه با استفاده از تکنیک های داده کاوی، مورد استفاده قرار می گیرد. این روش با استخراج مناطق ریه از تصویر CT با استفاده از تکنیک های پردازش تصویر آغاز می شود، از جمله برش تصویر بیتی، فیلتر Erosion و Weiner. تکنیک برش Bit plane در فرآیند استخراج برای تبدیل تصویر CT به یک تصویر باینری استفاده می شود. تکنیک برش Bit plane سریع تر می باشد و داده ها و کاربر مستقل هستند. الگوریتم های بسیاری برای تشخیص سرطان ریه ایجاد شده اند، اما اگر فرضیه های وابسته در نظر گرفته شوند، ثابت نمی شوند. در عصر الگوریتم های مورد استفاده برای تشخیص سرطان ریه، ساختار SOM مبتنی بر نرم افزار توسعه نمی یابد. این مقاله با تجسم ساختار بسته های منطقه ریه شروع می شود و سپس پایگاه داده های اختلال با استفاده از جعبه ابزار SOM برای ایجاد SOM آموخته شده با استفاده از خوشه بندی

K-means ، در نظر گرفته می شود. استفاده از این تکنیک استخراج داده ها با SOM، دارای مزایای قوی برای تحلیل روش موثر، است. روش داده کاوی از روش یادگیری برای درک الگوهای داده استفاده می کند. SOM می تواند در فرایند تجزیه و تحلیل داده ها و تجزیه و تحلیل داده های اکتشافی، استفاده شود.

در پروژه [25] یک سیستم پیشگیرانه پیشگیری از سرطان پوست براساس داده کاوی ارائه شده است و روشی کارآمد برای استخراج الگوی قابل توجه از پایگاه داده برای پیش بینی کارایی سرطان پوست، ارائه شده است. روش پیشنهادی با استفاده از یادداشت های لوتوس اجرا می شود. روش پیشنهادی می تواند سرطان پوست را به طور موثر و موفقیت آمیزی پیش بینی کند و نرم افزار اجرا شده از طریق آنلاین ارائه خواهد شد به طوری که هر فرد به راحتی می تواند سطح خطر سرطان پوست خود را بررسی کند.

مقاله [26] حاوی طراحی، توسعه و ارزیابی یک سیستم غربالگری خودکار AML در تصاویر میکروسکوپی خون است. از تصاویر با کیفیت بالا که توسط انجمن هماتولوژی آمریکا دریافت شده، استفاده می شود. سیستم ارائه شده پردازش خودکاری را از همبستگی رنگ، تقسیم سلول های هسته ای، و اعتبار سنجی و طبقه بندی موثر، انجام می دهد. مجموعه ای از ویژگی های استخراج پارامترهای شکل، رنگ و بافت یک سلول برای ایجاد تمام اطلاعات مورد نیاز برای طبقه بندی کارآمد، ساخته شده است. اپراتور LBP و HD نتایج قابل توجهی را برای این تحلیل، ارائه دادند. خروجی سیستم منجر به وضعیت طبیعی یا غیر طبیعی بیمار شد. سیستم پیشنهادی در مقایسه با سیستم موجود بهتر عمل می کند. دقت 85.71٪ حاصل شده است.

در مقاله، توالی مراحل نشان داده می شود که برای طبقه بندی کارآمد لوسمی حاد مگنوم، دنبال شده است. این سیستم دارای چهار مرحله اصلی است. مرحله پیش پردازش، مرحله تقسیم بندی،

مرحله استخراج ویژگی و مرحله طبقه بندی. در مرحله پیش پردازش محتوای نویز های ناخواسته در تصویر حذف می شود. همچنین تصویر RGB به تصویر فضای رنگ L * a * b تبدیل می شود. مرحله پیش پردازش به دنبال مرحله تقسیم بندی است که با استفاده از خوشه k-means انجام می شود. پس از آن ویژگی های استخراج شده در مرحله استخراج ویژگی که عمدتا از LBP و HD استفاده می کنند. این مرحله به مرحله طبقه بندی می پردازد که از SVM استفاده می کند. درنهایت، اعتبار سنجی انجام می شود.

در مقاله [27] توانایی نایو بیز و طبقه بندی های ماشینی بردار پیشتیبانی با توجه به دقت طبقه بندی دو مجموعه داده های مختلف سرطان، مقایسه می شود. بهترین نتایج با استفاده از طبقه بندی نایو بیز و ماشین بردار پشتیبانی می شود. اعتقاد بر این است که نتایج امیدوار کننده است و با پیش پردازش داده ها و تنظیم طبقه بندی ها می توان، آنها را بهبود بخشید. در ارزیابی آزمایشات بر روی طبقه بندی نایو بیز، آنها بهبود 96٪ برای سرطان سینه و 100٪ برای سرطان لوسمی را یافتند. پس از ارزیابی همان آزمایش ها بر روی ماشین بردار پشتیبانی ، بهبود 97٪ برای سرطان سینه و 100٪ برای سرطان لوسمی یافت شد.

در مقاله [28] تمرکز بر یافتن الگوریتم درست برای طبقه بندی داده هایی است که بر روی مجموعه داده های مختلف کار می کند، آن ها اهدافی را برآورده می سازند که هفده الگوریتم طبقه بندی انتخاب شده را بر اساس ابزار Weka برای پیش بینی بهترین مدل بیماری های لوسمی ، تجزیه و تحلیل می کردند. بهترین الگوریتم مبتنی بر داده های لوسمی، طبقه بندی تصادفی درخت با دقت 100٪ است و کل زمان ساخت این مدل در 0.02 ثانیه می باشد. این نتایج نشان می دهد که در این میان، الگوریتم یادگیری ماشینی آزمایش شده است زیرا می تواند به طور قابل توجهی بهبود روش های طبقه بندی متعارف را در زمینه پزشکی و یا به طور کلی، زمینه بیوانفورماتیک را بهبود بخشد.

جدول 2-1- مقایسه کارهای انجام شده در زمینه پیش بینی سرطان

شماره	عنوان	نوع سرطان	ایده اصلی	شیوه	سال
1	اثربخشی سیستم پیش بینی سرطان بر اساس داده کاوی (DMBCPS)	سرطان سینه، پوست و ریه	ارائه هشدار قبلی به کاربران و همچنین هزینه و زمان برای کاربر مفید است.	پیاده سازی سیستم پیش بینی پیشگیری از سرطان داده ها (DMBCPS) . این سیستم خطر ابتلا به سرطان را برآورد می کند.سرطان پستان، پوست و سرطان ریه با استفاده از درخت تصمیم گیری.	2013
2	پیش بینی سرطان سینه با استفاده از جنگل تصادفی، پشتیبانی از ماشین های بردار و نایو بیز	سرطان سینه	مقایسه عملکرد این سه الگوریتم از طریق صحت، دقت، فراخوانی و اندازه گیری f.	با استفاده از سه تکنیک یادگیری ماشین: نایوبیز ، SVM و جنگل تصادفی به پایگاه داده سرطان سینه ویسکانسین برای پیش بینی اینکه آسیب های بیمار خوش خیم یا بدخیم است.	2017
3	سیستم CAD SVM مبتنی بر تشخیص سرطان ریه است	سرطان ریه	پیشنهاد سیستم های CAD برای افزایش دقت و کاهش زمان تشخیص	با استفاده از پردازش تصویر برای تشخیص سرطان ریه و تولید نتایج حاصل از ویژگی استخراج و پس از تقسیم بندی. با استفاده از الگوریتم SVM	2017
4	یک روش آموزش مکانی برای پیش بینی اولیه سرطان پستان	سرطان سینه	انتخاب بهترین طبقه بندی برای تشخیص زودرس و بیماری بیان شده با دقت بهتر	سه نوع مدل مختلف در مجموعه داده های سرطان سینه تحت عنوان نایوبیز، رگرسیون لجستیک و جنگل تصادفی ، اجرا شد	2017
5	نقش سیستم پیش بینی سرطان مبتنی بر داده کاوی (DMBCPS) در آگاهی سرطان	سرطان سینه	سیستم برنامه برای پیش بینی خطر ابتلا به سرطان سینه در مرحله اولیه	پیشنهاد سیستم پیش بینی سرطان بر اساس تکنولوژی داده کاوی (DMBCPC)	2013
6	بررسی سیستم پیش بینی سرطان ریه با استفاده از تکنیک های داده کاوی و نقشه سازماندهی خود (SOM)	سرطان ریه	اجرای مدل موثر برای پیش بینی بیماران مبتلا به سرطان ریه به نظر می رسد الگوریتم SOM است.	تجسم ساختار بسته ناحیه ریه ها و سپس مجموعه داده های اختلال، با استفاده از جعبه ابزار SOM برای ایجاد SOM آموخته شده با استفاده از طبقه بندی K-means است.	2017
7	پیشگیری اولیه از ابتلا به	سرطان پوست	اجرای یک سیستم با	جمع آوری داده ها با استفاده از	2013

	سرطان پوست با استفاده از داده کاوی		استفاده از یادداشت های لوتوس برای پیش بینی سطح خطر سرطان پوست با پیشنهاداتی که آسان تر، با هزینه و زمان تر هستند.	الگوریتم خوشه بندی K-means برای جداسازی داده های مربوطه و غیر مرتبط به سرطان پوست.	
8	سیستم تشخیص کامپیوتری برای لوسمی حاد مگنوم در تصاویر میکروسکوپهای خون	سرطان لوسمی	الگوریتمی را ایجاد می کند که به طور خودکار لوسمی حاد مگنی را طبقه بندی می کند	ارائه یک روش کارآمد که به طور خودکار سلول های AML را در اسمیر خون با استفاده از الگوریتم (svm) کشف می کند	2016
9	طبقه بندی مجموعه داده های سرطان در الگوریتم های داده کاوی با استفاده از ابزار R	سرطان سینه و لوسمی	پیش بینی حضور دو نوع سرطان لوسمی و سینه با تجزیه و تحلیل مجموعه داده های بالینی.	توانایی نایو بیز و طبقه بندی های SVM با توجه به دقت بر طبقه بندی دو مجموعه داده های مختلف سرطان، مقایسه می شود.	2017
10	مدل پیش بینی بیماری های لوسمی بر اساس الگوریتم های طبقه بندی داده ها با بهترین دقت	سرطان لوسمی	استفاده از بهترین مدل برای پیش بینی بیماری های لوسمی با مقایسه دقت طبقه بندی و زمان پاسخ در میان یازده درخت تصمیم گیری روش ها و شش روش طبقه بندی قانون با استفاده از پنج عملکرد شاخص.	کارایی درخت تصادفی نتیجه بهتر را می دهد. همچنین زمان کمتری را برای ساخت مدل در طبقه بندی درخت دارد. طبقه بندی قوانین الگوریتم های به عنوان نزدیک ترین الگوریتم (NNge) بهترین الگوریتم است که دقت بالا و زمان کمتری را برای ساخت مدل در طبقه بندی به خود اختصاص می دهد.	2017