مروری بر روش های پیش بینی و تشخیص سرطان ریه با استفاده از تکنیک های داده کاوی

علی حسینی *
باشگاه پژوهشگران جوان و نخبگان، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران
ec.ali.hosseini@gmail.com
سعید صدیقی
باشگاه پژوهشگران جوان و نخبگان، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران
sedighi.saeed@gmail.com
محمدامین شایگان
گروه مهندسی کامپیوتر، واحد شیراز، دانشگاه آزاد اسلامی، شیراز، ایران
shayegan@iaushiraz.ac.ir
محمدایمان جم نژاد
گروه مهندسی کامپیوتر، موسسه آموزش عالی زند، شیراز، ایران
e.jamnezhad@gmail.com

چکیده

کشف دانش و داده کاوی، یک حوزه جدید میان رشته ای و در حال رشد است که حوزه های مختلفی چون پایگاه داده، آمار، یادگیری ماشین و سایر زمینه های مرتبط را با هم تلفیق کرده تا اطلاعات و دانش ارزشمند نهفته درحجم بزرگی از داده ها را استخراج نماید. هدف داده کاوی، یافتن الگوها و یا مدل های موجود در پایگاه داده ها است که در میان حجم عظیمی از داده ها مخفی هستند. داده کاوی در زمینه پزشکی دارای کاربردهای بسیار وسیع و در عین حال حساس و حیاتی است. با توجه به این که داده های پزشکی با ارزش ترین و حساس ترین داده ها برای کاوش و تحلیل هستند، تحلیل و کسب دانش از آنها می باید با درجه بالایی از دقت و حساسیت صورت گیرد.

این مقاله مروری بر روش های تشخیص سرطان ریه دارد چرا که در بسیاری از نقاط دنیا ریه ها شایع ترین محل ایجـاد سرطان در بدن بوده است و در پایان به بحث و نتیجه گیری می پردازد.

کلیدواژه‌ها: داده کاوی، سرطان ریه، تشخیص سرطان ریه، الگوریتم های خوشه بندی، الگوریتم های دسته بندی

1- مقدمه

طبق آمار وزارت بهداشت درمان و آموزش پزشکی همه ساله عده زیادی از مردم به انواع سرطان مبتلا می شوند. در بسیاری از نقاط دنیا ریه ها شایع ترین محل ایجاد سرطان بوده وسرطان های گوارش، سینه نیز از انواع دیگر سرطان های شایع می باشند.

در ایران به دلیل مشکلات ثبت اطلاعات بیماران سرطان و محدود بودن مطالعات فراگیر مبتنی بر جامعه، تخمین دقیقی از میزان بروز سرطان های مختلف وجود ندارد. اگر چه با توجه به آمار و مطالعات موجود، به نظر می رسد سرطان های‌گوارش(به ویژه معده، روده بزرگ و مری)، ریه، سینه، مثانه، پروستات، خون، لمفوم و مغز از بروز نسبتا بالایی برخوردار می باشند(موسسه سرطان[1]،2012). بعلاوه سرطان ها امروزه پس از تصادفات و بیماری های قلبی عروقی دلیل اصلی مرگ زود رس در مردم ایران می‌باشند.

تعداد زیادی از مردم ایران طبق آمار وزارت بهداشت درمان و آموزش پزشکی به سرطان ریه مبتلا می‌شوند و این بیماری جز شایعترین سرطان ها در ایران می باشد. میزان بروز سرطان ریه در نقاط مختلف کشور بسیار متغییر بوده و بین1.5%در (زنان اردبیل)تا 10.5%(در مردان تهران)گزارش شده است(موسسه سرطان،2010). سایرانواع سرطان نیز در قسمت های مختلف ایران از بروز و شیوع بسیار متغییری برخوردار هستند.

اگر تهران را متوسطی از کشور و نمونه تعمیم پذیری از جامعه ایران فرض کنیم و جمعیت ایران را حدود70میلیون نفر وحدود نیمی از آن را مذکر در نظر بگیریم، با توجه به میزان بروز خام سرطان های گزارش شده درمردان و زنان تهران تخمین زده می شود در ایران سالیانه حداقل40000نفر(22600زن و17400مرد)به سرطان های ریه، سینه، لمفوم، مری، روده بزرگ، مغز، پانکراس، سرو گردن، ملانوم، تخمدان، بیضه، مغز و بافت های عضلانی مبتلا می گردند(موسسه سرطان،2010).

امروزه در ایران و بسیاری از کشور های دنیا کیفیت و امید به زندگی در انواع مختلف سرطان بهبود یافته است، اما همچنان بسیاری از این بیماران در مرحله ای تشخیص داده می شوند که به علت گسترش بیماری نمی توان برای آنها کار زیادی انجام داد. به عنوان مثال در اسکاتلند به علت تشخیص دیر هنگام و گسترش بیماری، تنها 15 %بیماران دارای سرطان ریه از نوع سلول های غیر کوچک یا NSCLCدر زمان تشخیص قابل جراحی شدن بوده اند(فیسی و همکاران[2]،2007) و پیش بینی می شود این میزان در ایران از این حد هم کمتر باشد. همچنین با توجه به آمار وزارت بهداشت در سال 2006 حدود6 میلیون مرگ به دلیل ابتلا به سرطان گزارش شده در حالی که این آمار در سال 2012 به 8 میلیون نفر رسیده و پیش بینی می‌شود در سال 2030 این میزان به

[1] The Cancer Institute
[2] Facey et al.

درمـان زودتـر، دقیـق تـر و پیشـگیری ازعوارض بیماری در مراحل اولیه آن می باشند، که با استفاده از تکنیک های داده کاوی می توان پیش بینی زود هنگام سرطان ریه را انجام داد.
بخش های آتی این مقاله بصورت زیر می باشند. بخش دوم مبانی نظری تعریف خواهد گردید. بخش سوم مروری از کارهای انجام شده در این حوزه را معرفی می کند و در بخش چهارم به جمع بندی و نتیجه گیری پرداخته خواهد شد.

2- مبانی نظری

2-1- داده کاوی

داده کاوی عبارت از اقتباس یا استخراج دانش از مجموعه ی بسیار حجیم داده است. به عبارتی دیگر داده کاوی فرایندی است که با استفاده از تکنیکهای هوشمند، دانش را از مجموعه ای از داده ها استخراج می کند(هان و کمبر[1]،2011).

2-1-1- خوشه بندی

گروه بندی مجموعه ای اشیاء به کلاس هایی از اشیاء مشابه را خوشه بندی گویند. اعضای درون یک خوشه شباهت های زیادی به یکدیگر دارند ولی اعضای هر خوشه با اعضای موجود در خوشه های دیگر شباهت خیلی کمتری دارند(بنسال و همکاران[2]،2017). برخی از روش‌های خوشه بندی عبارتند از: روش سلسله مراتبی، روش میانگین K ، روش های بر مبنای چگالی و روش دو مرحله‌ای

2-1-2- دسته بندی

دسته بندی عملیات اختصاص داده ها به یکی از گروه های از پیش تعیین شده است(جاهان و جیسوال[3]،2016). برخی از روش‌های دسته بندی ماشین های بردار پشتیبان Support Vector Machines(SVM)، نزدیکترین همسایگی(KNN)k-Nearest Neighbor، درخت تصمیم، شبکه های عصبی و شبکه های بیزین می باشند.

3-مروری بر کارهای گذشته

پورکودی[1] پنج الگوریتم SVM، Naive Bayes، CN2، Random Forest، که از جمله الگوریتم های طبقه بندی معروف و قدرتمند در زمینه داده کاوی هستند، از لحاظ کارایی با هم مقایسه شده اند.
در این مقاله از دیتاست سرطان ریه که شامل 32 نمونه و 56 صفت خاصه و 3 کلاس می باشد، استفاده شده است. همچنین برای ارزیابی نتایج طبقه بندی این الگوریتم ها از از پنج معیار و

[1] Han and Kamber
[2] Bansal etal.
[3] Chauhan and Jaiswal

meature استفاده شده است.

نتایج به دست آمده نشان می دهد که بر اساس معیار دقت، الگوریتم KNN و پس از آن به ترتیب الگوریتم های Naive Bayes، CN2، Random Forest و SVM بهترین کارایی را داشته اند و همچنین نتایج به دست آمده بر اساس خصوصیت سطح زیر نمودار (AUC) نشان می دهد که در ابتدا الگوریتم Random Forest و پس ار آن به ترتیب KNN، Naive bayes، CN2 و SVM دارای بهترین کارایی بوده اند و سپس نتایج به دست آمده بر اساس معیار F-meature که ترکیبی از Recall و Precision می باشد نشان می دهد که الگوریتم KNN نسبت به بقیه الگوریتم ها بهترین کارایی را دارد و پس از آن Naive Bayes، CN2، Random Forest و SVM قرار دارد.

بنابراین طبق مقایسه ها و ارزیابی های صورت گرفته بر روی دیتاست مربوط به سرطان ریه، الگوریتم KNN نسبت به چهار الگوریتم مطرح شده در این مطالعه، بهترین و بالاترین کارایی و الگوریتم ماشین بردار (SVM) بدترین و پایین ترین کارایی را داشته است(پورکودی،2014).

زوبی و همکاران[2] با هدف تشخیص زود هنگام سرطان ریه با استفاده از اعمال چندین تکنیک داده کاوی و الگوریتم شبکه عصبی مصنوعی Artificial Neural Network(ANN) ، بر روی تصاویر گرفته شده از روی قفسه سینه با استفاده از اشعه ی ایکس (x) را معرفی کردند. در این مطالعه از 60 عدد تصویر قفسه سینه استفاده شده است که در ابتدا این تصاویر را به دو دسته تصاویر نرمال و غیر نرمال تقسیم و سپس تصاویر غیر نرمال را باز به دو دسته خوش خیم و بدخیم تقسیم بندی کرده اند که از این 60 تصویر ، 24 نمونه نرمال ،22 نمونه خوش خیم و 14 نمونه بدخیم بوده اند. بسیاری از این تصاویر در شرایط روشنایی متفاوتی اسکن شده اند و به همین دلیل بعضی از این تصاویر بسیار روشن و بعضی دیگر بسیار تاریک هستند و همچنین حدود نیمی از مجموعه تصاویر شامل نویز می باشند.

لذا قبل از هرگونه عملیاتی نیاز به پیش پردازش داده ها شامل : کاهش داده ها ، تبدیل داده ها، متراکم کردن داده ها و تمیز کردن داده ها بوده است . سپس عملیات Image Segmentation بر روی تصاویر اعمال تا بخشی که دارای تومور و غده ی سرطانی است از بقیه بخش ها جدا شود . سپس فازFeature Extraction اجرا شده است، تا مجموعه صفت های خاصه بهینه از درون حجم و تعداد بسیار زیاد ویژگی ها برای تشخیص زود هنگام سرطان ریه، به دست آید که از جمله مهمترین ویژگی ها، اندازه و شکل توده بوده است که بر اساس این دو ویژگی می توانند خوش خیم یا بد خیم بودن توده ها را تشخیص دهند.

هرچه که سلول ها بزرگتر، بی شکل و سوزنی شکل باشند احتمال اینکه سرطانی باشند بیشتر و هرچه که سلول ها و تومورها کوچکتر، صاف، نرم و گردتر باشند احتمال اینکه خوش خیم باشند بیشتر می باشد. در نهایت الگوریتم شبکه عصبی مصنوعی (ANN) که از جمله الگوریتم های طبقه

[1] Porkodi
[2] Zubi

تشخیص داد که داده های ورودی جدید به کدام دسته تعلق دارند. این الگوریتم نمونه های نرمال را با دقتی برابر 100% و نمونه های خوش خیم را با دقتی برابر 95% و نمونه های بدخیم را با دقتی برابر 85% تشخیص داده است(زوبی و همکاران،2014)

احمد و همکاران[1] یک سیستم پیش گویی سرطان ریه پیشنهاد کردند که هم استفاده از آن ساده و کم هزینه بوده و هم باعث صرفه جویی در زمان شده است.

تمام عملیات مورد نظر روی 400 داده که شامل 200 نفر زن و 200 نفر مرد که سن آنها بین 20 تا 80 سال می باشد و از مراکز مختلف تشخیص سرطان در بنگلادش جمع آوری شده است ، صورت گرفته است.

به دلیل وجود مقادیر تکراری، مفقود شده (Miss Values) در منبع داده ها، ناگزیر عملیات پیش پردازش بر روی داده ها انجام شده تا بدین طریق داده های تکراری حذف و مقادیر مفقود شده با داده های نمونه های قبلی جایگزین شوند. همچنین عملیات پیش پردازش باعث کاهش حافظه و نرمال کردن داده نیز می شود. سپس با استفاده از الگوریتم k-means که از جمله الگوریتم های معروف خوشه بندی می باشد، داده ها را به خوشه سرطانی و غیر سرطانی تقسیم کرده اند. در اینجا k یک مقدار مثبت و صحیح می باشد که نشان دهنده تعداد کلاسترها (خوشه ها) می باشد و در این مطالعه برابر با دو در نظر گرفته شده است. بعد از خوشه بندی، الگوریتم های درخت تصمیم گیری وAprioriTid مورد استفاده قرار می گیرد تا الگوهای تکراری را کشف و بدست آورند، چون این الگوریتم ها، روش های موثری در زمینه بیرون کشیدن الگوهای تکراری از دیتاست های خوشه بندی شده می باشند، سپس با استفاده از فرمول 1 الگوهای مهم انتخاب می شوند:

$$Sw(i) = \sum (Wi * Fi) \qquad (1)$$

که Wi وزن عددی هر ویژگی و Fi تعداد تکرار هر الگو (قانون) را نشان می دهد و در صورتیکه $Sw(n) \geq \emptyset$ برای همه مقادیر n برقرار باشد ، الگوهای تکراری مهم انتخاب می شوند.

الگوهای تکراری در واقع مجموعه داده هایی هستند که درون مخزن داده ها مکررا تکرار شده اند و الگوهای تکراری مهم، مجموعه داده هایی هستند که نقش مهم تری در سرطان ریه دارند. در نهایت با استفاده از این الگوهای تکراری مهم ابزار ساده و موثر پیش گویی سرطان ریه پیاده سازی شده است(احمد و همکاران،2013).

کریشنا و همکاران[2] دو الگوریتم طبقه بندی NB(Naive Bayes) و One Dependency Augmented Naïve Bayes(ODANB) را با هم مقایسه کردند.

[1] Ahmed et al.
[2] Krishnaiah et al.

Cancer –H و Lung Cancer –Statlog که تکنیک های داده کاوی بر روی داده های موجود در این دیتاست ها اعمال شده اند.

برای آماده سازی داده ها، بر روی آنها عملیات پیش پردازش داده ها صورت گرفته است و مقادیر مفقود شده را با میانگین و مد مقادیر موجود جایگزین کرده اند. چون هم داده هایی از نوع اسمی چندمقداری (Nominal) وعددی (Numerical) در دیتاست ها وجود داشته است در ابتدای کار، صفت های خاصه عددی را به نوع اسمی چند مقداری تبدیل و سپس دو الگوریتم NB و ODANB را بر روی داده ها اعمال کرده اند. به طوریکه دقت به دست آمده برای الگوریتم NB بر روی دیتاست هایی که در بالا مطرح شد به ترتیب برابر با 84.14% ، 84.05% ، 83.70% و همچنین دقت های به دست آمده برای الگوریتم ODANB بر روی دیتاست های مذکور به ترتیب برابر با 80.46% ، 79.66% ، 80.00% بوده است. این نتایج حاکی از کارایی بهتر الگوریتم NB نسبت به الگوریتم ODANB می باشد. بنابراین برای تشخیص سرطان ریه الگوریتمNB مناسبتر می باشد و در واقع هدف از این مطالعه، پیشنهاد یک مدل برای تشخیص درست و زود هنگام سرطان ریه می باشد به طوری که بتوان به پزشکان در نجات جان انسان ها کمک موثری کرد(کریشنا و همکاران،2013).

بنای دزفولی و همکاران تحقیقاتی پیرامون و کشف ارتباطات و الگوها با استفاده از داده های بالینی در داده کاوی انجام دادند که می تواند در پیش بینی بقای بیماران مبتلا به سرطان ریه مورد استفاده قرار گیرد. این مقاله از داده های پزشکی که در ارتباط با این بیماری است و توسط سایت SEER طی سال های 1998تا 2011 جمع آوری شده است، استفاده می کند.

در روش پیشنهادی ویژگی ها با استفاده از سه الگوریتم داده کاوی درخت تصمیم ، شبکه بیزیان و شبکه های عصبی مورد بررسی گرفته و میزان تاثیر هر ویژگی، در پیش آگهی بقای بیماران سرطان ریه مورد مطالعه قرار گرفت. با استفاده از الگوریتم ها، مدل هایی ایجاد شده است که میزان خطر مرگ و میر را در پنج دسته ی شش ماه، نه ماه، یک سال، دو سال و پنج سال تشخیص می دهند. نتایج آزمایشات نشان می دهد بالاترین دقت مربوط به الگوریتم $C_{5.0}$ درخت تصمیم می باشد.

پس از اعمال روش پیشنهادی، انتخاب ویژگی به طوری که ویژگی هایی که نقش مهم تری در تشخیص سرطان ریه دارند انتخاب شده و ویژگی هایی که مهم نبوده اند حذف شده، انجام گردید ودقت مربوط به الگوریتم$C_{5.0}$ معادل 97.93% بدست آمد. همچنین روش کاهش ابعاد، با استفاده از الگوریتم PCA، به کار برده شده است که با توجه به مقایسه نتایج ، روش انتخاب ویژگی دقت بالاتری را نسبت به روش کاهش ابعاد نشان می دهد.

در این مطالعه با استفاده از سه الگوریتم درخت تصمیم، شبکه عصبی و شبکه بیزیان، دقت بالاتری نسبت به روش های پیشین حاصل شد(بنای دزفولی و همکاران،1393).

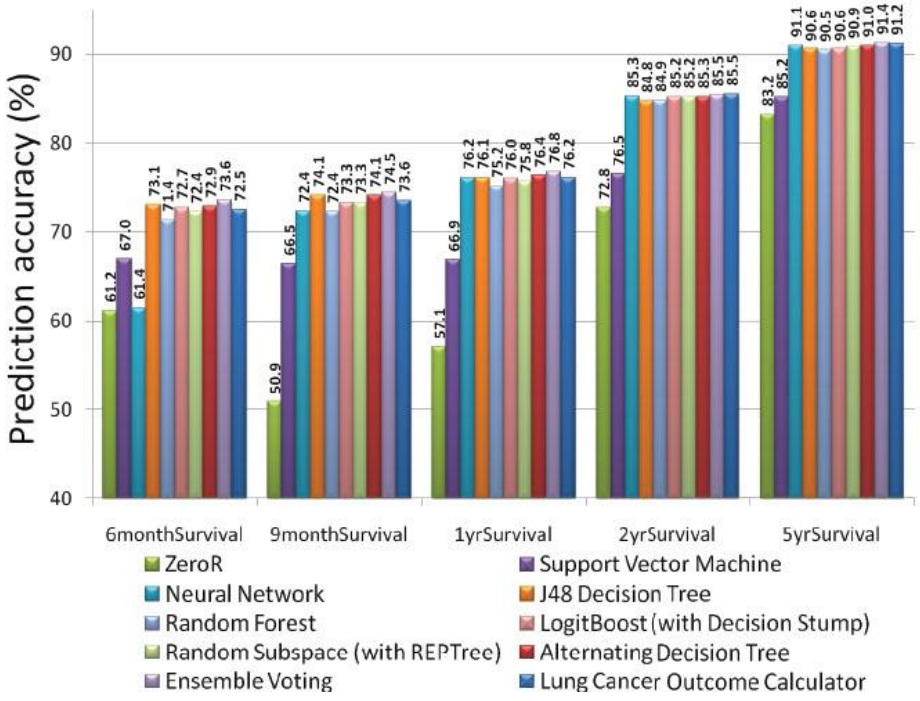
کاوی ارائه کردند. در این تحقیق از مجموعه داده هایی SEER استفاده شده است که از محدوده داده 9 دفتر ثبت (از جمله : آتلانتا، کانکتیکات، دیترویت، هاوایی، آیووا، نیومکزیکو، سان فرانسیسکو، پرشین سیاتل، یوتا) جمع آوری شده است.

این مجموعه داده شامل 70132 رکورد و 118 ویژگی که پس از فیلتر و حذف آنها به 57254 رکورد و 63 ویژگی کاهش یافت که از 63 ویژگی 13 ویژگی مهم آنها را انتخاب کرده اند.

برای آزمایشات از 5 دسته بندی درخت تصمیم اصلی و5 درخت فرا دسته بندی استفاده شده است که عبارتند از :

SVM,J48 Decision Tree,Random Forest,Logitboost,Decision Stump,Random Subspace Alternating Decision Tree,Lung Cancer Outcome Calulator,Voting,ZeroR که دقت آنها در شکل1 نشان داده شده است. نتیجه بدست آمده از این مقاله، یک مدل برای تخمین خطر مرگ سرطان ریه پس از 6 ماه، 9ماه، 1سال، 2سال و 5 سال ارائه داده شده است که توزیع کلاس ها در این مقاله به این صورت است:

در 6 ماه 38/85% زنده در 9 ماه 49/12% زنده در 1سال 57/04% زنده در 2 سال 72/79% زنده و در 5 سال 83/23% زنده بوده است. برای اینکه 13 ویژگی با دقت انتخاب شود باید از روش های انتخاب ویژگی استفاده شود. دقت های پیش بینی با مقدار 73.61%و 74.45%و 76.80%و 85.45%و91.35%را به ترتیب برای 6 ماه، 9ماه، 1سال، 2سال و 5سال را برای پیش بینی بقاء(وجود) سرطان ریه بدست آورده اند(آگراوال و همکاران،2011).



شکل 1- دقت الگوریتم ها(آگراوال و همکاران،2011)

[1] Agrawal et al

زیبا و همکاران[1] برای حل مشکل داده های غیر متوازن، SVM تقویت شده را ارائه دادند. راه حل پیشنهادی، مزایای استفاده از Classifier دسته جمعی برای داده نا هموار را با ماشین بردار حساس به هزینه را ترکیب کرده است. همچنین رویکردی جدید برای استخراج قوانین تصمیم گیری از SVM تقویت شده را ارائه کردند. در نهایت کیفیت متد ارائه شده توسط مقایسه کارایی با الگوریتم های دیگر با وجود داده های نا متوازن آزمایش گردید که بهبود نزدیک به 7% را نشان می دهد. سپس SVM تقویت شده برای برنامه های کاربردی پزشکی جهت پیش بینی امید به زندگی در بیماران سرطان ریه به کار گرفته شده بود(زیبا و همکاران،2014).

داس و همکاران[2] این واقعیت را در نظر گرفتند که جهش ژن ها اساس توسعه سرطان است. مجموعه های داده Genomic و Proteomic از سلول های NSCLC(سلول کوچک) و دو زیر رده اصلی آن SCC و ADC در این مطالعه تجزیه و تحلیل گردیده است. در روش پیشنهادی یک الگوریتم استنتاج درخت تصمیم گیری روی نشانگرهای سرطان های NSCLC برای ایجاد پیش بینی ها اعمال شده است. الگوریتم پیشنهادی دقت کلاس بندی بالای دارد و دارای قابلیت پیش بینی نوع سرطان ریه است. تکنیک اعتبارسنجی پیوند نیز جهت بهبود دقت کلاس بندی الگوریتم J48 اعمال گردیده است. در این روش درخت تصمیم گیری توسط ابزار وکا برای زیر رده سرطان ریه استفاده شده است و نوع سرطان ریه های دارای کلاس ناشناخته را پیش بینی می کند. در مرحله دوم خروجی بدست آمده توسط الگوریتم J48 را با درخت تصمیم بهبود یافته (J48) مقایسه گردیده است. بر اساس ساختمان درخت تصمیم گیری کل 10 قوانین کلاس بندی اول (بالا) توسط الگوریتم Apriori ابزار وکا جهت پیش بینی سرطان ریه بدست آمده است. میانگین دقت صحت کلاس بندی در این مقاله حدود 99.7%می باشد. (داس و همکاران،2014)

رامانی و جکب[3] یک استراتژی محاسباتی جهت پیش بینی رده تومورهای سرطان ریه، از طریق ساختار و ویژگی های فیزیکی و شیمیایی که شامل 1497 ویژگی، از ترتیب های پروتئین بدست آمده از ژن ها، به کمک تجزیه و تحلیل ریز آرایه طراحی نمودند. در متد پیشنهادی از تکنیک های انتخاب ویژگی پیوندی بر پایه پیش بینی شبکه بیزین برای تمایز بین تومورهای سرطانیSCLC، NSCLC و رده های معمولی استفاده شده است. همچنین از طریق خوشه بندی نظارت شده، خوشه های ممکن در داده تومور ریه را پیش بینی کردند. نتایج های بدست آمده نشان می دهد که الگوریتم های خوشه بندی نظارت شده کارایی ضعیفی را در تمایز رده های تومور ریه نشان می دهد(رامانی و جکب،2013)

[1] Zięba et al.
[2] Dass et al.
[3] Ramani and Jacob

مرحله پیش پردازش مربوط به دیتابیس SEER ،پیش پردازش مشکلات مشخص، مدل پیش گویی و ارزیابی می باشد.

برای آزمایشات از10 درخت تصمیم اصلی و فرا دسته بندی Random Subspac,Random Forest , Logitboost ,Alternating Decision Tree,J48 Decision Tree,Ensemble Voting,Neural Network,Lun Cancer Outcome Calculator , SVM استفاده شده است. در تشخیص زود هنگام سرطان به موقع، برای تخمین ریسک زنده ماندن بیمار را بعد از6ماه، 9ماه،1سال ،2سال و 5 به وسیله 13 ویژگی بررسی کردند. این ویژگی ها شامل داده های آماری جمعیت شناسی(جنس،سن،...)بوده اند.

نتایج بدست آمده از دست بندی ها نشان می دهد که SVM بی دقت تر و ناسازگارتر در مقایسه با سایر دسته بندی ها بوده است.و بهترین درخت تصمیم Ensemble Voting با دقت 73/6، 74/5، 76/8، 85/5، 91/4 به ترتیب برای از 6ماه، 9ماه، 1سال،2سال و 5 بوده است(آگراوال و همکاران،2012)

دزفولی و ساجدی[1] یک مدل کارآمد برای پیش بینی میزان بقای افراد مبتلا به سرطان ریه پیشنهاد داده بودند. در این مقاله هشت الگوریتم chaid,c5,C&R,Quest,TAN Bayes,Mark or Bayes,NN ouick,NN Dyna mic استفاده شده است.

داده ها استفاده شده در این مقاله از دیتاست SEER جمع آوری شده است. در روش پیشنهادی داده ها زنده ماندن بیماران را بعد از 6ماه،9ماه،1سال،2سال و 5 به وسیله4ویژگی بررسی کردند. در بین الگوریتم بیشترین دقت را C5 با دقت98/12%،96/88%،96/76%،95/23%،90/63%به ترتیب برای از 6ماه، 9ماه، 1سال،2سال و 5 بوده است.

در این مقاله یک مکانیزم جدید برای انتخاب ویژگی ارائه شده است که نتیجه بهبود آن برای الگوریتم C5 بر روی داده های تست، از دقت97/51%،96/87%،96/47%،94/76%،90/75%به ترتیب برای از 6ماه، 9ماه،1 سال،2سال و 5به97/93%،96/94%،96/91%،96/32%،93/12%افزایش یافته است(دزفولی و ساجدی ،2015)

یاداو و همکاران برای بهینه حذف کردن داده های پرت نسبت به الگوریتم k-means الگوریتم جدیدی برای خوشه بندی داده های سرطان ریه ارائه کردند. داده های استفاده شده در این مقاله از مرکز Sanjay Gandhi Post Graduate Institute Of Medical Sciences(SGPGI) در هند جمع آوری شده است. عملیات مورد نظر روی سوابق 177 بیمار انجام شده است. الگوریتم ارئه شده به نام Foggy K-means نسبت به روش سنتی k-means از خوشه بندی بهتری و حذف بهینه تر داده های پرت برخوردار است. در خوشه بندی جدید عواملی همچون اثرات بد سیگار کشیدن، اشعه تولید شده توسط صنایع مختلف و یا مواد رادیو اکتیو در نظر گرفته شده است(یاداو و همکاران،2013).

[1] Dezfuly and Sajedi

رضا کرمانشاه جمع آوری شده است برای پیش بینی سرطان ریه استفاده کردند. در این مجموعه داده، بیش از 68 % نمونه ها مربوط به بیماران مرد و 32% به بیماران زن تعلق داشته است. بیش از 65 %مبتلایان به سرطان ریه در دهه ششم زندگی قرار دارند، همچنین بیش از 68 %مبتلایان به سرطان ریه سابقه مصرف سیگار داشتند. در این مجموعه داده از شایعترین علائم بالینی افراد مبتلا به سرطان ریه می توان به درد قفسه سینه، خلط خونی، تنگی نفس و سرفه اشاره کرد. ویژگیهای مطرح شده در این مجموعه داده با استفاده از الگوریتمهایMLP و RBFدر نرم افزار وکا و نروسلوشن مورد بررسی قرار گرفتند. نتایج آزمایشات نشان میدهد که شبکه عصبی RBF به ترتیب با دقت 96.57% و 90.74%در هردو نرم افزار وکا و نروسلوشن بالاترین دقت پیشبینی را در تشخیص اولیه سرطان ریه را دارد. همچنین نتایج پیاده سازی با یک نمونه مجموعه داده از پایگاهTCGAمورد آزمایش قرار گرفته است. با توجه به نتایج این بررسی، موفقیت آموزش انجام شده توسط مجموعه داده بیمارستان امام رضا استان کرمانشاه در تشخیص اولیه سرطان ریه بر روی پایگاه داده TCGA را نشان داده شده است(فتحی و فرضی،1395).

4- بحث و نتیجه‌گیری

سرطان ها امروزه پس از تصادفات و بیماری های قلبی و عروقی دلیل اصلی مرگ زود رس در مردم ایران می‌باشند و سرطان ریه یکی از شایعترین آن در ایران است. موثرترین راه برای کاهش مرگ ناشی از سرطان ریه تشخیص زود هنگام این بیماری می باشد، چرا که پزشکان می توانند براساس برخی از اطلاعات اولیه بالینی، افراد مشکوک به سرطان ریه را تشخیص و درمان نمایند. با بررسی کارهای گذشته مشخص گردید که استفاده از تکنیکهای داده کاوی می تواند پزشک را در تشخیص زود هنگام این بیماری کمک کند و باعث صرفه جویی در منابع پزشکی، افزایش بهره وری از درمانهای پزشکی، تشخیص زود رس، پیشگیری اولیه و کاهش زباله های پزشکی شود.

این مقاله مروری بر روش های تشخیص سرطان ریه می باشد و اغلب مقالات موجود در این حوزه مورد بررسی قرار گرفته است. در این راستا مشخص گردید که الگوریتم های مختلف روی دیتابیس های مختلف دقت جداگانه ای خواهند داشت، پس نمی توان گفت که الگوریتم خاصی برتری به الگوریتم دیگری دارد و برای هر دیتابیس الگوریتم خاصی وجود دارد که دقت بالاتری نسبت به بقیه الگوریتم ها خواهد داشت، همچنین دریافتم که عملیات پیش پردازش در بالا بردن دقت مدل بسیار مهم و تاثیرگذار خواهد بود. همچنین نیاز به یک دیتابیس محلی در کشور بشدت احساس می شود چرا که شرایط آب و هوایی، نوع غذا و کلا نوع زندگی در کشور های مختلف متفاوت می باشد و با به بوجود آوردن یک دیتابیس محلی از بیماران سرطان ریه می توان نتیجه مطلوب تری نسبت به داده های غیر محلی بدست آورد.

مراجع

[1]بنای دزفولی ،مهدیس؛ساجدی ، هدیه؛ پیش بینی بقای بیماران مبتلا به سرطان ریه با استفاده از الگوریتم های طبقه بندی در داده کاوی,دومین همایش ملی علوم و مهندسی کامپیوتر,1393

[2]فتحی حاجی آباد ،فیروزه؛ فرضی ، سعید؛ پیش بینی سرطان ریه در استان کرمانشاه براساس شبکه های عصبی مصنوعی, اولین همایش ملی نگرشی نوین در مهندسی برق و کامپیوتر,1395

[3] Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2011, August). ***A lung cancer outcome calculator using ensemble data mining on SEER data***. In Proceedings of the Tenth International Workshop on Data Mining in Bioinformatics (p. 5). ACM.

[4] Agrawal, A., Misra, S., Narayanan, R., Polepeddi, L., & Choudhary, A. (2012). ***Lung cancer survival prediction using ensemble data mining on SEER data***. Scientific Programming, 20(1), 29-42.

[5] Ahmed, Kawsar, Et Al. ***Early Detection Of Lung Cancer Risk Using Data Mining***. Asian Pacific Journal Of Cancer Prevention, 2013, 14.1: 595-598.

[6] Bansal, A., Sharma, M., & Goel, S. (2017). ***Improved k-mean clustering algorithm for prediction analysis using classification technique in data mining***. International Journal of Computer Applications (0975–8887) Volume, 157, 33-40.

[7] Chauhan, D., & Jaiswal, V. (2016, October). ***An efficient data mining classification approach for detecting lung cancer disease***. In Communication and Electronics Systems (ICCES), International Conference on (pp. 1-8). IEEE.

[8] Dass, M. V., Rasheed, M. A., & Ali, M. M. (2014, January). ***Classification of lung cancer subtypes by data mining technique***. In Control, Instrumentation, Energy and Communication (CIEC), 2014 International Conference on (pp. 558-562). IEEE.

[9] Dezfuly, M, Sajedi, H, (2015). ***Predict Survival Of Patients With Lung Cancer Using An Ensemble Feature Selection Algorithm And Classification Methods In Data Mining*** ,Journal of Information, 1(1), 1-11.

[10] Facey, K., Bradbury, I., Laking, G., & Payne, E. (2007). ***Overview of the clinical effectiveness of positron emission tomography imaging in selected cancers***. HEALTH TECHNOLOGY ASSESSMENT-SOUTHAMPTON-, 11(44).

[11] Han, J., Pei, J., & Kamber, M. (2011). ***Data mining: concepts and techniques***. Elsevier. ISO 690

[12] Krishnaiah, V.; Narsimha, Dr G.; Chandra, Dr N. Subhash. ***Diagnosis Of Lung Cancer Prediction System Using Data Mining Classification Techniques***. International Journal Of Computer Science And Information Technologies, 2013, 4.1: 39-45.

[13] Porkodi, ***A Study on Performance Analysis of Data Mining Classification Algorithms over Lung Cancer Dataset*** ,International Journal of Research in Information Technology (IJRIT), 2014, Pg: 49-58

[14] Ramani, R. G., & Jacob, S. G. (2013***). Improved classification of lung cancer tumors Based on structural and physicochemical properties of proteins using data mining models***. PloS one, 8(3), e58772.

[15] Yadav, A. K., Tomar, D., & Agarwal, S. (2013, July). ***Clustering of lung cancer data using Foggy K-means***. In Recent Trends in Information Technology (ICRTIT), 2013 International Conference on (pp. 13-18). IEEE.

[16] Zięba, M., Tomczak, J. M., Lubicz, M., & Świątek, J. (2014). ***Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients***. Applied soft computing, 14, 99-108.

[17] Zubi, Zakaria Suliman; SAAD, Rema Asheibani***. Improves Treatment Programs of Lung Cancer Using Data Mining Techniques***. Journal of Software Engineering and Applications, 2014, 2014.