مقاله پژوهشی

شناسایی بیماران سیروز در معرض ابتلا به سرطان کبد به کمک رویکردهای داده‌کاوی: گزارش مقدماتی

ملینا ابراهیمی خامنه[۱]، محمد مهدی سپهری[۲]، مهدی صابری فیروزی[۳]

[۱] دانشجوی کارشناسی ارشد، دانشکده مهندسی صنایع، دانشگاه آزاد اسلامی واحد تهران جنوب، تهران، ایران
[۲] دانشیار، بخش مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه تربیت مدرس، تهران، ایران
[۳] استاد، مرکز تحقیقات گوارش و کبد بیمارستان دکتر شریعتی، دانشگاه علوم پزشکی تهران، تهران، ایران

چکیده

زمینه و هدف:
کشف دانش و داده‌کاوی به دنبال یافتن الگوها و یا مدل‌های موجود در پایگاه داده است که در میان حجم عظیمی از داده‌های مخفی هستند. اعمال روش های داده کاوی بر داده‌های پزشکی می تواند به عنوان سیستم های تصمیم یار، در تصمیم گیری برای انتخاب نوع درمان و یا تشخیص بیماری‌ها، به متخصصان کمک نماید. هپاتوسلولار کارسینوما شایعترین نوع سرطان کبد است و با توجه به پیش آگهی ضعیف آن، چهارمین علت مرگ و میر مربوط به سرطان نیز می‌باشد. هدف این پژوهش استفاده از ابزارهای داده کاوی است تا بتوان به کمک آن سیستم کمک تصمیمی طراحی کرد که به پزشکان در شناسایی بیماران با سیروز کبدی و شانس بالاتر ابتلا به سرطان کبد کمک کند.

روش بررسی:
داده‌ی مورد نیاز برای این پژوهش از پژوهشکده گوارش و کبد بیمارستان شریعتی جمع آوری‌شده است. تعداد ۲۵۸ بیمار سیروتیک که وضعیت بیماری آنها برای ۴ سال پیگیری شده بود، برای مطالعه انتخاب شدند. از بین الگوریتم‌های داده‌کاوی، الگوریتم درخت تصمیم که یکی از معروف ترین و ساده ترین روش های قابل فهم در داده‌کاوی است، برای ساخت مدل دسته بند مورد استفاده قرار گرفت.

یافته‌ها:
با بررسی‌های انجام شده بر روی بیماران تحت بررسی و استفاده از الگوریتم های داده‌کاوی، ساختار درخت تصمیم علاوه بر آلفا فیتوپروتئین اهمیت مشخصه‌های مانند شاخص توده ی بدنی (BMI)، کراتینین، پلاکت، بیلی روبین توتال، INR و آلبومین را در پیش بینی سرطان در بیماران سیروتیک مشخص کرد. درخت تصمیم توانست به طور متوسط با دقت ۸۸٪ برای بیماران با اتیولوژی ویرال و دقت ۹۲٪ برای بیماران با اتیولوژی غیر ویرال به پیش بینی سرطان در بیماران سیروتیک بپردازد.

نتیجه گیری:
با توجه به نتایج دسته بند درخت تصمیم که نشان می‌دهد عواملی مانند اتیولوژی، سن، شاخص توده بدنی (BMI)، پلاکت، بیلی روبین توتال، INR، کراتینین، آلفا فیتوپروتئین و آلبومین می‌تواند احتمال ابتلای بیماران سیروتیک به سرطان کبد را پیش‌بینی کند، پیشنهاد می‌شود در مطالعات طولی با تعداد نمونه بیشتر نتایج مورد بررسی قرار گیرد.

کلید واژه: داده کاوی، هپاتوسلولار کارسینوما، سیروز کبدی، پیش بینی، دسته بندی، درخت تصمیم

گوارش/ دوره ۱۹، شماره۴/ زمستان ۱۳۹۳/ ۲۷۴-۲۶۵

زمینه و هدف:

کشف دانش و داده کاوی[1] 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 در حجم بزرگی از داده ها را استخراج نماید. هدف داده کاوی، یافتن الگوها و یا مدل های موجود در پایگاه داده ها اســت 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 انســانی با ارزش ترین و حســاس ترین داده‌ها برای کاوش و تحلیل هستند، تحلیل و کسب دانش از آنها می‌باید با درجه بالایی از دقت و حساسیت صورت گیرد. داده کاوی فرآیند استخراج دانش و یا الگوهای غیر بدیهی، مفید، قبلا ناشناخته و بالقوه مفید، از درون داده های با حجم زیاد تعریف شده است.(۱)

1. Data mining

نویسنده مسئول: محمد مهدی سپهری
تهران، بزرگراه جلال آل احمد، پل نصر، دانشــگاه تربیت مدرس، کد پستی ۱۴۱۱۷۱۳۱۱۴
تلفن و نمابر: ۰۲۱-۸۲۸۸۳۳۷۹
پست الکترونیک : mehdi.sepehri@modares.ac.ir
تاریخ دریافت: ۹۳/۵/۱۵
تاریخ اصلاح نهایی: ۹۳/۸/۲۰
تاریخ پذیرش: ۹۳/۸/۲۱

هپاتوسلولار کارسینوما شایع ترین نوع سرطان کبد است و ششمین سرطان شایع در سراسر جهان به حساب می‌آید.(۲) سرطان کبد اغلب در بیماران سیروز کبدی و به خصوص بیماران با عفونت هپاتیت مزمن B یا C به وجود می‌آید.(۳) سیروز کبدی در حدود ۸۰ تا ۹۰ درصد موارد زمینه ساز ابتلا به سرطان کبد در جهان است.(۴) این سرطان با نرخ سالانه ۱٪ تا ۴٪ بعد از ابتلا بیمار به سیروز کبدی رخ می‌دهد (۵)، ولی راهی برای پیش بینی شانس ابتلا به سرطان کبد در این بیماران پیشنهاد نشده است.

بنابراین در صورتی که بتوانیم از داده های پزشکی بیماران و با استفاده از الگوریتم های داده کاوی به پیش بینی ابتلای بیماران سیروز کبدی به سرطان کبد بپردازیم، در واقع به تصمیم گیری پزشکان در خصوص این بیماران کمک شایانی خواهیم کرد. روش های پیشگویی داده کاوی در پزشکی بالینی ساخت یک مدل پیشگویانه است که به پزشکان کمک می‌کند تا روش‌های پیشگیری، تشخیص و برنامه های درمانی خود را بهبود بخشند.(۶) تا به حال تحقیقات و مطالعات زیادی در زمینه کاربرد داده کاوی در پزشکی انجام شده است.(۹- ۷) در زمینه ی بیماری های کبدی، الگوریتم های داده کاوی بسیار مورد توجه محققان قرار گرفته است.(۱۳- ۱۰) کوروساکی[1] و همکاران در سال ۲۰۱۲ به پیش بینی سرطان در بیماران مبتلا به هپاتیت c پرداختند. آنها برای پیش بینی سرطان مشخصه هایی مانند سن، پلاکت، آلبومین و آمینوترانسفر را در نظر گرفتند و به شناسایی بیمارانی که شانس ابتلای بالا به سرطان دارند، پرداختند.(۱۴)

با توجه به این که فاکتورهای آزمایشگاهی هر کدام به تنهایی قادر به پیش‌بینی سرطان نیستند، حتی سطح آلفا فیتوپروتئین سرم نیز به تنهایی برای تشخیص هپاتوسلولار کارسینوما مورد استفاده قرار نمی‌گیرد.(۱۵) بنابراین در تحقیق موجود فرض بر این است که در نظر گرفتن فاکتورهای آزمایشگاهی مختلف در کنار هم می‌تواند باعث ساخته شدن مدل مناسب تری برای پیش بینی سرطان شود و وضعیت بیمار را در آینده بهتر مشخص کند.

روش بررسی:

داده‌ی مورد نیاز برای این پژوهش از پژوهشکده گوارش و کبد بیمارستان شریعتی جمع آوری‌شده است. تعداد ۲۵۸ بیمار سیروتیک که وضعیت بیماری آنها برای ۴ سال پیگیری شده بود، برای مطالعه انتخاب شدند. خصوصیات دموگرافیک بیماران، اتیولوژی و نتایج آزمایش ها در جدول ۱ خلاصه شده است. داده های مربوط به وضعیت بیماران، در یک بازه‌ی زمانی ۴ ساله، برای ساخت مدل های دسته بند مورد استفاده قرار گرفتند. از ۲۵۸ بیمار تحت بررسی ۱۸۰ نفر مرد (۷۰٪) و ۸۷ نفر زن (۳۰٪) بودند. تعداد ۳۶ نفر در طی این ۴ سال مبتلا به سرطان کبد شدند، که بر چسب کلاس مثبت به آنها اختصاص داده شد و ۲۲۲ نفر دیگر برچسب کلاس منفی گرفتند.

1. Kurosaki

جدول ۱: نام و نوع مشخصه های جمع آوری شده از بیماران

نام مشخصه	نوع مشخصه	نام مشخصه در مدل
جنسیت	دسته ای (زن، مرد)	Gender
سن	عددی	Age
اتیولوژی	دسته ای (ویروسی، غیر ویروسی)	Etiology
شاخص توده بدنی (BMI)	عددی	BMI
پلاکت (متوسط)	عددی	Platelet
البومین (متوسط)	عددی	SerumAlbumin
بیلی روبین توتال (متوسط)	عددی	Total Bilirubin
کراتینین (متوسط)	عددی	Creatinine
آلفا فیتو پروتئین	عددی	AlfaFP
زمان پروترومبین	عددی	INR
ابتلا به سرطان	دسته ای (بله، خیر)	HCC

تشخیص سیروز کبدی بر اساس پاتولوژی، فیبرواسکن، یافته‌های رادیولوژیک و علایم بالینی توسط متخصصین فوق تخصص گوارش و کبد داده شده بود. از بیماران آزمایش های پلاکت، بیلی روبین توتال، INR، کراتینین، آلفا فیتوپروتئین و آلبومین انجام شده بود. که در هر مورد میانگین حسابی آنها در نظر گرفته شد. سرطان کبد بر اساس یافته های رادیولوژیک و یا افزایش بیش از ۴۰۰ نانوگرم در سی سی آلفا فیتوپروتئین و یا پاتولوژی بر اساس معیار های پیشنهادی AASLD[2] در مدیریت HCC (۱۶) داده شد.

از بین الگوریتم های داده‌کاوی، الگوریتم درخت تصمیم که یکی از معروف ترین و ساده ترین روش های قابل فهم در داده‌کاوی است، برای ساخت مدل دسته بند مورد استفاده قرار گرفت. در شکل ۱ گام های اجرای پژوهش نشان داده شده است.که در دو بخش داده های پژوهش و داده کاوی مراحل به طور کامل توضیح داده می‌شوند.

داده های پژوهش

جمع آوری داده اولین مرحله در فرآیند داده کاوی به حساب می‌آید. تعداد و کیفیت داده های جمع آوری شده در دقت مدل پیش بینی کننده نقش مهمی ایفا می‌کند.(۱۷) داده ی مورد نیاز این پژوهش از پژوهشکده گوارش و کبد بیمارستان شریعتی جمع آوری شده است. تعداد ۲۵۸ بیمار سیروتیک برای مطالعه انتخاب شدند. وجود سیروز کبدی در بیماران از طریق نتایج بیوپسی یا سونوگرافی، جواب‌های آزمایش و علایم سیروز پیشرفته مانند آسیت، آنسفالوپاتی کبدی و یا واریس تشخیص داده شده است. بیمارانی که در نتایج سونوگرافی و سی تی اسکن آنان، وجود توده با خصوصیات کارسینوم هپاتوسلولار مشخص شده بود و یا آلفا فیتوپروتئین بالای ۴۰۰ و توده کبدی داشتند و علل دیگر افزایش آلفا فیتوپروتئین رد

2. American Association for the Study of Liver Diseases

شکل ۱: مراحل اجرای پژوهش

شده بود به عنوان سرطان کبد در نظر گرفته شدند.

داده کاوی

یکی از مهم ترین تکنیک های داده کاوی روش دسته بندی است که یک روش دو مر حله ای می‌باشد.(۱) مرحله اول، مرحله‌ی یادگیری می‌باشد. در این مرحله یک دسته بند بر اساس یک مجموعه از داده ها با برچسب کلاس مشخص ساخته می‌شود. در مرحله دوم، دسته بند ساخته شده روی مجموعه ای از داده ها به نام داده های آزمایشی اعمال می‌گردد تا دقت دسته بند آزمایش شود. با توجه به این که یکی از هدف های پژوهش حاضر قابلیت درک دانش توسط متخصصان و کارشناسان می‌باشد، از الگوریتم دسته بند درخت تصمیم استفاده می‌کنیم. درخت تصمیم ابزاری قدرتمند برای دسته بندی و پیش بینی می‌باشد که در آن فرآیند دسته بندی هر نمونه با امتحان ویژگی بیان شده در گره ریشه شروع شده و سپس از شاخه های درخت با توجه به مقدار آن ویژگی پایین می‌آید. سپس این فرآیند با امتحان گره بعدی که در انتهای شاخه ی انتخاب شده قرار دارد، ادامه می یابد تا در نهایت به یک برگ برسد .درخت تصمیم از جمله تکنیک هایی در داده کاوی به شمار می‌آید که برای تجزیه و تحلیل داده ها بسیار مورد استفاده قرار می‌گیرد.(۱۸) در زمینه بیماری های کبدی نیز استفاده از درخت تصمیم بسیار مورد توجه محققان قرار گرفته است. (۱۹و۲۰) از ۷۰٪ داده ها برای مرحله ی آموزش و از ۳۰٪ باقی مانده که در ساخت مدل شرکت نداشتند برای آزمایش مدل استفاده می‌شود. داده های موجود در دنیای واقعی ممکن است کیفیت لازم برای شروع داده کاوی را نداشته باشند. به عنوان مثال وجود نویز[1]، نمونه های پرت[2]، مقادیر از دست رفته[3] و داده های تکراری[4] در داده ها، اجرای مرحله پیش پردازش را ضروری می‌کند.

وقتی تکنیک های مختلف پیش پردازش روی داده ها انجام می‌شود در هر مرحله کیفیت داده ها و درنتیجه کیفیت کل فرآیند کشف دانش بهبود پیدا می‌کند به دلیل وجود داده‌های از دست رفته در مشخصه‌های تحت بررسی بر آن شدیم که از روش K نزدیک ترین همسایه[5] برای پرکردن مقادیر از دست رفته استفاده کنیم. عملکرد این الگوریتم بدین صورت است که از k نزدیک ترین همسایه‌ی نمونه، برای پرکردن مقادیر مفقود استفاده می‌شود. برای هر مقدار از دست رفته ۳ نزدیک ترین همسایه انتخاب شد. به هر همسایه با توجه به فاصله اش با داده ی مورد نظر وزنی اختصاص یافت. برای این منظور از پکیج DMwR در نرم افزار R استفاده شد و میانگین وزن دار همسایه ها محاسبه و برای پر کردن مقادیر از دست رفته به کار گرفته شد. وزن های مورد نیاز طبق معادله ۱ به دست می‌آید که در آن dist(k,x) فاصله ی اقلیدسی بین نمونه x با مقادیر از دست رفته و همسایه اش (k) می‌باشد.

$$w=e^{dist(k,x)}$$

پژوهش حاضر از الگوریتم خوشه بندی K-Means برای شناسایی نقاط پرت استفاده شده است که از جمله محبوب‌ترین الگوریتم های یادگیری بدون نظارت است و در آن مجموعه داده ها به تعداد خوشه های از پیش تعیین شده تقسیم می‌شوند. بعد از شناسایی نقاط پرت، پرونده این بیماران مجدد بررسی شد تا مشخص شود آیا داده‌ها اشتباه وارد شده اند یا نه. بعد از این بررسی بعضی مقادیر اصلاح و بقیه‌ی نقاط پرت حذف شدند.

برای تهیه یک مدل کاربردی و توسعه درخت از الگوریتم CART(Classification And Regression Tree) استفاده شده است. دراین روش برای جداسازی و شاخه زدن از معیار p-value استفاده می‌شود و چنانچه اختلاف بین دو دسته به صورت آماری معنی دار باشد، عمل شاخه زنی در درخت انجام می‌گیرد و در غیر این صورت در درخت

1. Noise
2. Outliers
3. Missing values
4. Duplicate Data
5. K _Nearest Neighbor(KNN)

شـــاخه ای ایجاد نمی‌شود. برای ســـاخت درخت تصمیم ۷۰ درصد داده‌ها برای مرحله‌ی آموزش و ۳۰ درصد دیگر برای تســـت مدل ســـاخته شده ، استفاده شدند. دسته بند درخت تصمیم ساخته شده ۱۰۰ بار اجرا و میزان دقت و شـــاخص recall هر درخت محاسبه شـــد. درنهایت با توجه به این دو معیار درخت های با دقت بالا انتخاب شدند و نتایج آنها برای تفسیر در اختیار خبره قرار گرفت.

با توجه به اینکه احتمال ابتلا به ســـرطان در بیماران با اتیولوژی ویرال بیشـــتر از بیماران با اتیولوژی غیرویرال می‌باشـــد (۱۹) بیماران براســـاس اتیولوژی به دو گروه (ویرال ـ غیر ویرال) تقسیم شدند و از نرم افزار R برای ساخت درخت تصمیم استفاده شد. طبق نتایج حاصل از اجرای مدل‌ها در نرم افزار R درخت تصمیم پیشـــنهادی بـــرای بیماران با اتیولوژی ویرال در شکل ۲ و درخت تصمیم پیشنهادی برای بیماران با اتیولوژی غیرویرال در شکل ۳ آورده شده است.

یافته‌ها:

درختان تصمیم رســـم شـــده، در هر گره درصد بیماران سرطان کبد با برچســـب مثبت و بیماران با ســـیروز و بدون سرطان کبد در طی مطالعه با برچســـب منفی و همچنین درصد بیمارانی که شرایط آنها در گره‌ی تحت بررســـی صدق می کند را مشـــخص می‌کند. برای مثال در گره شـــماره ۳ درشـــکل ۲، ۸۷ درصد بیماران برچسب کلاس مثبت و ۱۳ درصد بیماران برچسب کلاس منفی دارند. همچنین عدد ۲۵٪ در این گره نشان می‌دهد که ۲۵ درصد کل بیماران در این گره قرار دارند.

طبق نتایج درخت تصمیم که برای بیماران با اتیولوژی ویرال در شکل ۲ پیشنهاد شده است، برای پیش بینی سرطان در بیماران سیروتیک ابتدا مقدار مشخصه‌ی آلفا فیتوپروتئین مورد بررسی قرار می‌گیرد. درصورتی که ســـطح آلفا فیتوپروتئین بیش از ۵۲ باشد، سن بیمار مورد بررسی می‌شود که عمل شـــاخه زنی در این گره در سن ۴۶ سال است. بنابراین ۲۱ درصد بیمارانی که در طی ۴ سال به سرطان مبتلا شده اند، سطح آلفا فیتوپروتئین بالای ۵۲ و ســـن بالای ۴۶ ســـال داشته اند که به احتمال ۹۴٪ به سرطان مبتلا می‌شوند. در بیماران با سطح آلفا فیتوپروتئین کمتر از ۵۲ و سن کمتر از ۵۹ ســـال که تصور می‌شـــود به احتمال کمتری به سرطان مبتلا شوند، نتایج درخت تصمیم نشـــان می‌دهد که اگر این بیماران ســـطح کراتینین بیشـــتر از ۱، BMI بالای ۲۵ داشته باشـــند و سطح سرم آلبومین آنها در بازه‌ی ۲/۶ تا ۳/۳ باشد، در معرض بیشتری برای ابتلا به سرطان قرار دارند. درخت تصمیم نشـــان می‌دهد اگر متوسط سطح آلفا فیتوپروتئین بیمار با اتیولوژی ویرال بین ۶/۵ تا ۵۲ و ســـن آنها بیش از ۵۹ سال باشد، این افراد اگر BMI بزرگتر از ۲۷ داشته باشند به احتمال ۹۵ درصد در طی ۴ سال به سرطان مبتلا می‌شوند. به همین ترتیب می توان تمام شاخه های درخت را پیمایش کرد و قوانین به دســـت آمده از آن را برای پیش بینی دقیق تر سرطان در بیماران سیروتیک به کار گرفت. با توجه به نتایج درخت تصمیم می‌توان بیمارانی که در معرض خطر کمتری برای سرطان کبد قرار دارند را نیز شناسایی کرد. به عنوان مثال بیماران با سطح آلفا فیتوپروتئین کمتر از ۵۲ و سن کمتر از ۵۹ سال در صورتیکه سطح کراتی نین کمتر از ۱ داشته باشند تنها به احتمال ۴٪ در معرض خطر ابتلا به سرطان کبد قرار دارند.

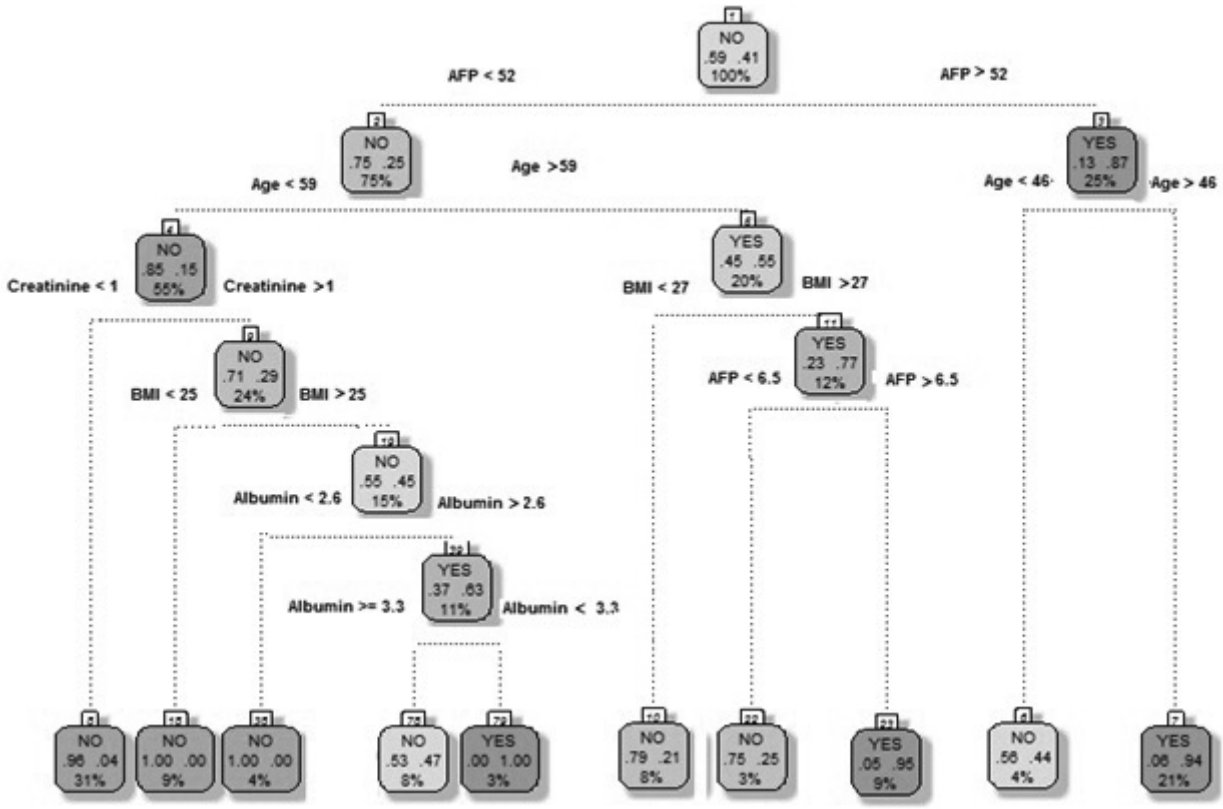
درخت تصمیم پیشـــنهادی رسم شده (شـــکل۳) برای کمک به پیش بینی ســـرطان در بیماران با اتیولوژی غیرویرال نشان دهنده‌ی آن است که بیمارانی که سطح آلفا فیتوپروتئین بالای ۲۶۰، بیلی روبین توتال کمتر از ۴/۸ و ســـن بیشتر از ۶۲ سال داشـــته اند با داشتن احتمال ابتلای ۱۰۰٪، محتمـــل ترین گروه برای ابتلا به ســـرطان در بین بیماران با اتیولوژی غیر ویرال می‌باشند. بیمارانی که سطح آلفافیتوپروتئین بین ۱۷ تا ۲۶۰ ، بیلی روبین توتال کمتر از ۲ و ســـن بیشـــتر از ۶۲ سال دارند به احتمال ۸۳٪ به سرطان مبتلا می‌شوند.

با توجه به این که یکی از هدف های پژوهش حاضر قابلیت درک دانش توسط متخصصان و کارشناسان می‌باشد، از این رو دانش حاصل شده توسط قوانین اگر ـ آنگاه برای بیماران ســـیروتیک که در معرض ابتلا به ســـرطان هستند در جدول ۲ آورده شده است.

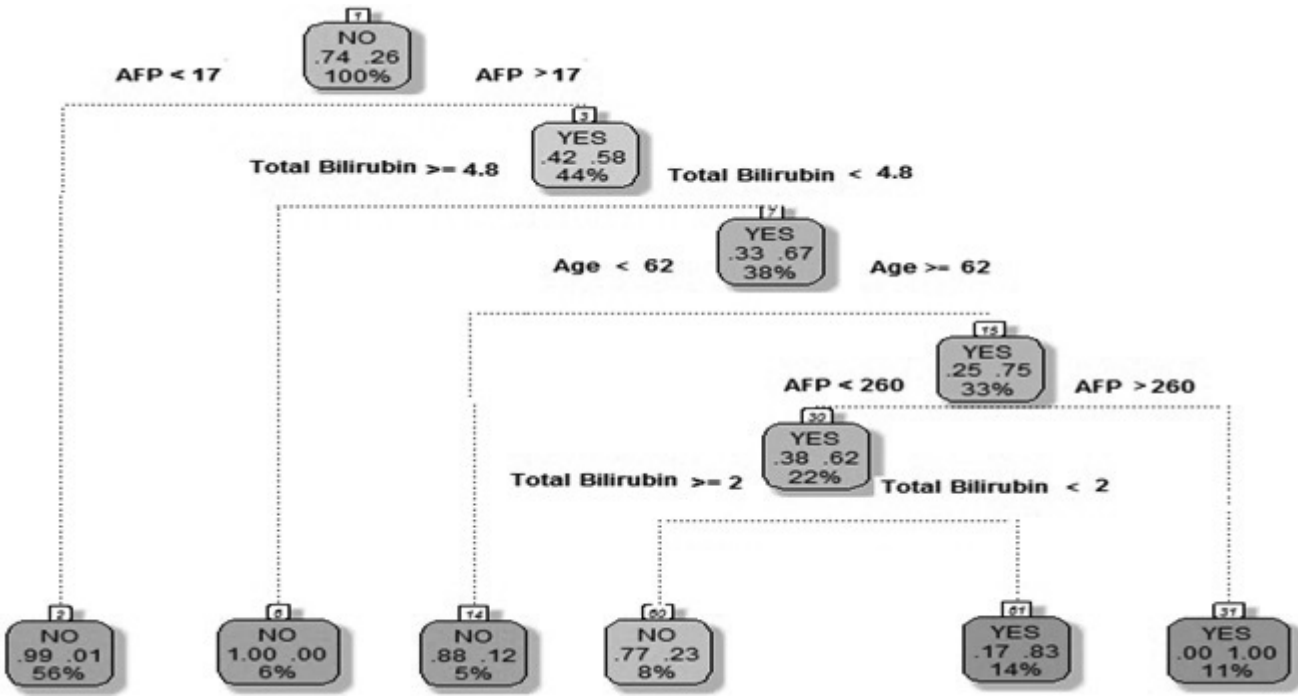
بعد از رسم درخت تصمیم به تفکیک اتیولوژی، در قدم بعدی دو گروه بیماران به صورت همزمان برای ســـاخت مدل مورد استفاده قرار گرفتند و نتایج حاصل از اجرای مدل در شکل ۴ آورده شده است. طبق نتایج درخت تصمیم ابتدا مقدار مشخصه‌ی آلفا فیتوپروتئین مورد بررسی قرار می‌گیرد. درصورتی که سطح آلفا فیتوپروتئین بیش از ۲۱ باشد، کراتین ین به عنوان مشـــخصه ی بعدی در نظر گرفته می‌شـــود که عمل شاخه زنی در این گره در مقدار ۰/۸ اســـت. بنابراین بیمارانی که آلفا فیتوپروتئین بیشـــتر از ۲۱ و کراتینین کمتر از ۰/۸ دارند به احتمال بسیار کم به سرطان مبتلا می‌شوند (۱۱٪). در بیماران با مقدار آلفا فیتوپروتئین بیش از ۲۱ و سطح کراتینین بالای ۰/۸ و سطح بیلی روبین توتال کمتر از ۰/۸۳ احتمال ابتلا به سرطان تنها ۱۴٪ می‌باشـــد. همانطور که در شکل ۴ مشاهده می‌کنید بیمارانی که ســـطح آلفا فیتوپروتئین آنها بالای ۲۱، کراتینین بـــالای ۰/۸، بیلی روبین توتال بالای۰/۸۳ و آلبومین بالای ۳ دارند و شـــاخص توده‌ی بدنی (BMI) آنها بالای ۲۲ است با داشتن احتمال ابتلای ۹۵٪، محتمل ترین گروه برای ابتلا به سرطان می‌باشند.

در ســـه درخت تصمیم رســـم شده مشاهده شـــد که مدل مشخصه‌ی آلفافیتوپروتئیـــن را به عنـــوان اولین فاکتور در پیش بینی ســـرطان مورد بررسی قرار می‌دهد. با حذف مشخصه‌ی آلفافیتوپروتئین و اجرای دوباره‌ی مدل، درخت ترســـیم شده در شکل ۵ به دست می‌آید که اهمیت مشخصه هایی مانند بیلی روبین توتال و کراتینین را در پیش بینی سرطان مشخص می‌کند. نتایج حاصل از این درخت گویا این است که بالا بودن سن و داشتن شاخص توده بدنی بیشتر از ۲۳ احتمال ابتلا به سرطان را افزایش می‌دهد.

ماتریـــس توافقـــی(confusion matrix) ابزار مناســـبی برای آنالیز چگونگی عملکرد یک دســـته بند اســـت. این ماتریـــس چگونگی عملکرد



شکل ۲: درخت تصمیم به دست آمده برای کمک به پزشکان در پیش بینی سرطان در بیماران با اتیولوژی ویرال



شکل ۳: درخت تصمیم به دست آمده برای کمک به پزشکان در پیش بینی سرطان در بیماران با اتیولوژی غیرویرال

جدول ۲: قوانین به دست آمده از درخت تصمیم

بیماران با اتیولوژی ویرال	•Rule1: If (AFP< 52.29) AND (Age< 59) AND (Creatinine>=1.019) AND (BMI>=25.48) AND (2.6<=Albumin< 3.285) THEN (HCC=YES, accuracy =100%)
	•Rule2: If (6.462 <= AFP< 52.29) AND (Age>=59) AND (BMI>=26.56) THEN (HCC=YES, accuracy =95%)
	•Rule3: If (AFP>=52.29) AND (Age>=46.5) THEN (HCC=YES, accuracy =94%)
بیماران با اتیولوژی غیرویرال	•Rule1: If (AFP>=259.8) AND (Total Bilirubin< 4.75) AND (Age> 62.5) THEN (HCC=YES, accuracy =100%)
	•Rule2: If (17.39<=AFP < 259.8) AND (Total Bilirubin< 4.75) AND (Age> 63) AND (Total Bilirubin< 1.992) THEN (HCC=YES, accuracy =83%)

الگوریتم دسته بند را با توجه به مجموعه داده ورودی به تفکیک انواع برچسب کلاس ها نشان می‌دهد. نمونه ای از این ماتریس در شکل ۶ آورده شده که در آن (TN) تعداد مقادیر منفی درست پیش بینی شده، (FP) تعداد مقادیر منفی که به اشتباه مثبت پیش بینی شده، (FN) تعداد مقادیر مثبت که به اشباه منفی پیش بینی شده و (TP) تعداد مقادیر مثبت درست پیش بینی شده می‌باشد.

از آنجا که معیار دقت دسته بندی ارزش رکوردهای دسته های مختلف را یکسان در نظر می‌گیرد، در مسائل واقعی نمی‌تواند به عنوان معیار مناسبی برای ارزیابی عملکرد استفاده شود. چرا که در دنیای واقعی کلاس ها ارزش یکسانی ندارند و پیش بینی درست و یا غلط برچسب رکورد های آن دسته نفع و یا ضرر متفاوتی در مقایسه با رکورد های سایر دسته بند ها دارد. بنابراین علاوه بر دقت، از شاخص صحت که به معنی درصد پیش بینی درست بیماران مبتلا به سرطان می‌باشد، استفاده می‌شود. برای اعتبار سنجی مدل از روش Random Subsampling استفاده شد به این صورت که مدل دسته بند درخت تصمیم ۱۰۰ بار اجرا و مقادیر دقت و صحت در هر اجرا محاسبه شد. با استفاده از نتایج ماتریس در هم ریختگی، مقادیر میانگین دقت و صحت برای مدل‌های دسته بند درخت تصمیم محاسبه و در جدول ۳ نشان داده شده است.

یکی از معیار های سنجش تناسب مدل های دسته بند، نمودار ROC می‌باشد. در صورتی که مدل به صورت کاملا تصادفی عمل کند، مساحت سطح زیر نمودار برابر ۰/۵ خواهد بود هر چه دقت مدل در شناسایی موارد صحیح بهتر باشد، مساحت زیر نمودار به سمت عدد ۱ میل خواهد کرد. نمودار ROC دسته بند درخت تصمیم به تفکیک اتیولوژی در شکل ۷ نشان داده شده است که مطابق آن سطح زیر نمودار برای بیماران با اتیولوژی ویرال برابر ۸۶/۰۲ و برای بیماران با اتیولوژی غیر ویرال برابر ۹۱/۰۸ می‌باشد.

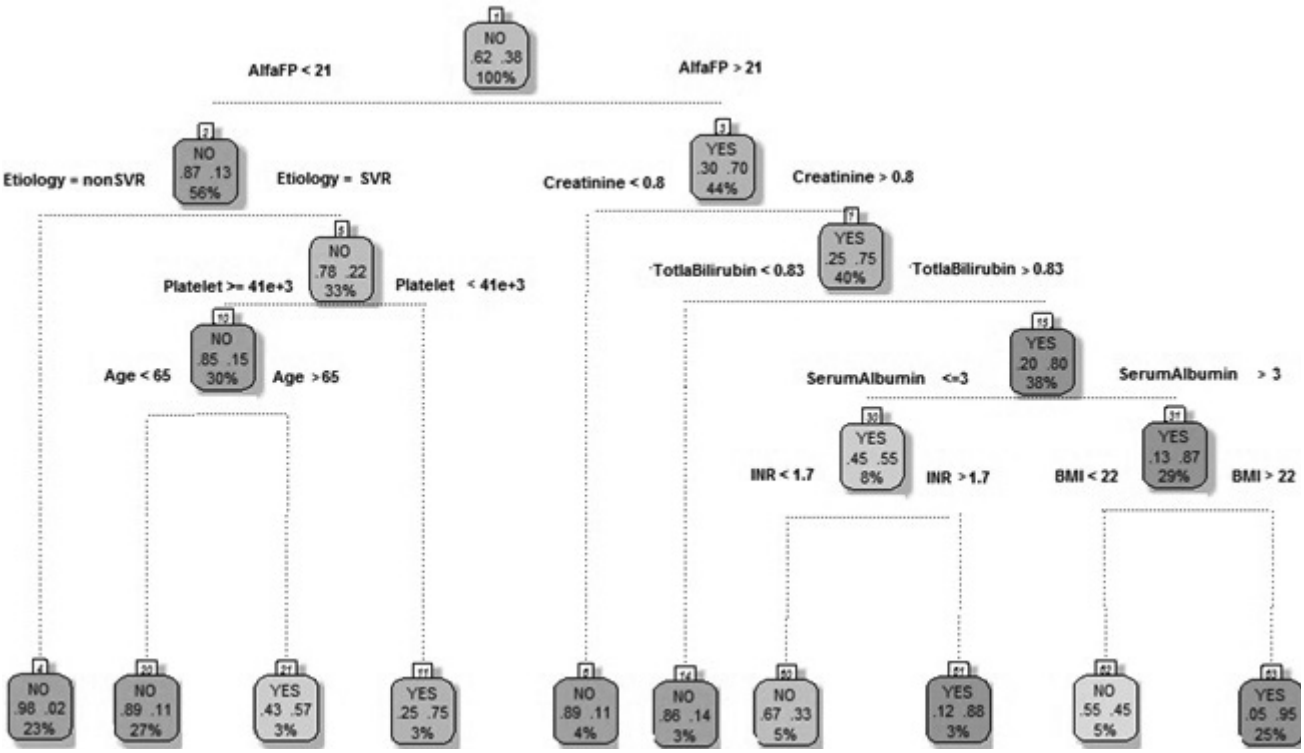
بحث:

داده کاوی یکی از زمینه های چند تخصصی برآمده از علومی چون آمار، ریاضی و هوش مصنوعی است که کاربردهای آن در حیطه های گوناگون پژوهشی، مدیریتی، اجرای سلامت و درمان رو به گسترش است. داده کاوی پزشکی علاوه بر مسائل فنی که در زمینه های دیگر داده کاوی هم وجود دارد، با پاره ای از مشکلات غیر فنی هم رو به رو است. محدودیت‌های اخلاقی، حقوقی و اجتماعی روی جمع آوری و توزیع داده های پزشکی وجود دارد که در مورد داده های دیگر وجود ندارند و این مطلب موجب محدود شدن نتیجه گیری‌ها می‌شود. به بیان دیگر داده‌های پزشکی دارای ویژگی‌های منحصر به فردی از جمله ناهمگن بودن، موارد اخلاقی، قانونی و اجتماعی می‌باشند که در نتایج به دست امده تاثیر گذار خواهد بود.

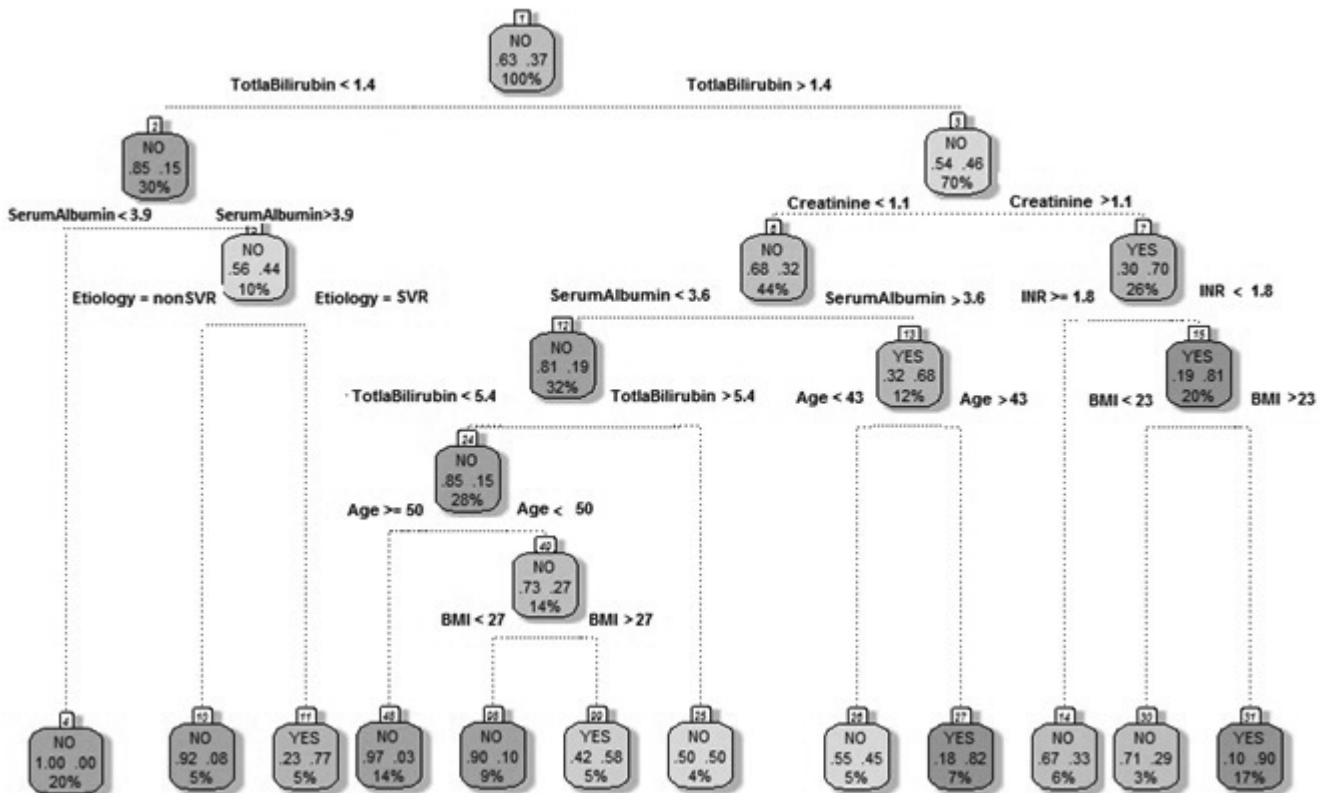
تشخیص زود هنگام یا پیش بینی افراد در معرض خطر بالاتر برای سرطان کبد و تدوین و توسعه مدلی که بتواند این سرطان را پیش بینی کند بسیار مورد نیاز است.(۱۲) درصورتی که سرطان دیر تشخیص داده شود ممکن است به نواحی دیگر بدن منتشر شود. در صورت انتشار سرطان کبد، گاهی سلول‌های سرطانی در ریه‌ها مشاهده می‌شوند.(۲۲) بنابراین در صورت به کارگیری نتایج به دست آمده این پژوهش توسط پزشکان، افرادی که در معرض ابتلا به سرطان قرار دارند به موقع شناسایی می‌شوند و این امر بهبود قابل توجه‌ای را در روند درمانی آنها به وجود خواهد آورد. تشخیص و درمان زود هنگام علاوه بر این که هزینه های اضافی را حذف می‌کند، به پزشکان کمک می‌کند که اقدامات مناسب تری را در زمان مناسب انجام دهند. به دلیل این که بیماری پیشرفت چندانی نداشته است، اقدامات درمانی ممکن است موثر تر واقع شوند.

چنانچه بتوانیم از آزمایش هایی که به طور توسط پزشکان آزمون می‌شوند برای پیش بینی و شناسایی بیماران در معرض خطر بالاتر استفاده کنیم می‌توانیم دفعات یا روش های غربالگری را تغییر دهیم. پژوهش حاضر برای پیش بینی سرطان کبد در بیماران سیروتیک از راه کار داده کاوی استفاده کرد. دانش حاصل شده از درخت تصمیم می‌توان به عنوان سیستم کمک تصمیم جهت پیش‌بینی سرطان در بیماران سیروتیک در بیمارستان‌ها استفاده شود.

در این پژوهش دسته بند درخت تصمیم برای پیش بینی سرطان در بیماران سیروتیک مورد استفاده قرار گرفت و نتایج برای هر دو گروه بیماران با اتیولوژی ویرال و غیر ویرال ارائه شد. در بیماران با اتیولوژی غیر ویرال علاوه بر مشخصه‌ی آلفا فیتوپروتئین، فاکتورهای سن و بیلی روبین توتال نیز مهم شناخته شد. درخت تصمیم رسم شده در شکل ۵ نشان می‌دهد احتمال ابتلا به سرطان در بیماران با اتیولوژی ویرال بسیار بیشتر از بیماران با اتیولوژی غیر ویرال است که تایید کننده تحقیقات قبلی می‌باشد.(۲۳) نتایج درخت تصمیم نشان می دهد احتمال ابتلای سرطان در بیماران با سن بالای ۶۵ سال بیشتر از بیماران با سن کمتر از ۶۵ سال می‌باشد، که تایید کننده نتیجه تحقیقات گذشته (۲۴و۲۵) می‌باشد. همچنین در



شکل ۴: درخت تصمیم به دست آمده برای کمک به پزشکان در پیش بینی سرطان در بیماران با هرنوع اتیولوژی



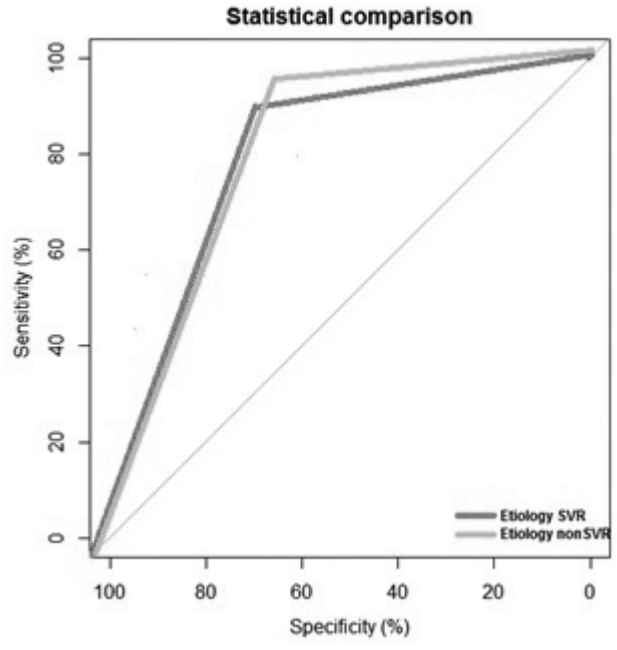
شکل ۵: درخت تصمیم در بیماران با هرنوع اتیولوژی بدون در نظر گرفتن مشخصه آلفافیتوپروتئین

		Predicted	
		Negative	Positive
Actual	Negative	TN	FP
	Positive	FN	TP

شکل ۶: ساختار ماتریس توافقی

جدول ۳: دقت و صحت به دست آمده از دسته بند درخت تصمیم

صحت	دقت	توضیح درخت
٪۷۵	٪۸۸	بیماران با اتیولوژی ویرال
٪۷۹	٪۹۲	بیماران با اتیولوژی غیرویرال
٪ ۷۰	٪۸۱	در نظر گرفتن کل بیماران به صورت یک گروه
٪۶۹	٪۷۳	حذف مشخصه‌ی آلفا فیتوپروتئین



شکل ۷: نمودار ROC

شکل ۵ نشان داد شده است، در بیماران با سرم البومین بالای ۳ که تصور می‌شود شانس کمتری در ابتلا به سرطان داشته باشند، مشاهده می‌شود که مقدار مشخصه‌ی BMI می‌تواند بسیار تاثیر گذار باشد. به طوری که در بیماران با BMI بالای ۲۲، احتمال ابتلا ۹۵٪ و در بیماران با BMI کمتر از ۲۲، احتمال ابتلا به سرطان ۴۵٪ می‌باشد. در پژوهش کوروساکی و همکاران (۱۴) نیز نشان داده شده است که برخلاف انتظار پزشکان، بیماران با سرم البومین بالای ۴ نیز احتمال ابتلا به سرطان بالاتری دارند.

که البته در تحقیق آنها مشخصه‌ی AST و در تحقیق حاضر مشخصه‌ی BMI به عنوان مشخصه‌ی شاخه زنی انتخاب شده است. از جمله نو آوری های پژوهش حاضر اضافه کردن مشخصه های مانند INR، BMI، کراتی نین و بیلی‌روبین به مرور ادبیات قبلی می‌باشد. مدل ساخته شده به دلیل قابلیت تفسیر بالا می‌تواند به سادگی توسط پزشکان با جایگزین کردن نتایج آزمایش بیماران در درخت تصمیم و پیمودن شاخه های مربوط به وضعیت بیمار، مورد استفاده قرار گیرد و احتمال ابتلای فرد به سرطان را مشخص کند. ساختار درخت تصمیم اهمیت مشخصه هایی مانند اتیولوژی، سن، شاخص توده بدنی (BMI)، پلاکت ، بیلی روبین توتال، INR، کراتی نین، آلفافیتوپروتئین و آلبومین را در پیش بینی سرطان در بیماران سیروتیک مشخص کرد.

نتایج به دست آمده از درخت تصمیم در پیش بینی سرطان پیشنهاد می‌کند افرادی که طبق نتایج درخت تصمیم شانس بالایی برای ابتلا به سرطان دارند، به جای این که هر ۶ ماه یکبار طبق معیار های پیشنهادی AASLD سونوگرافی شود (۲۶) در فواصل زمانی سه ماهه سونوگرافی شوند. که این کار هزینه‌های ناشی از تشخیص دیر هنگام سرطان را از بین می‌برد.

علاوه بر نقاط قوتی که قبلا اشاره شد، مانند اضافه کردن مشخصه‌های INR، BMI، کراتی نین به تحقیقات گذشته، تحقیق حاضر دارای نقاط ضعفی هم می‌باشد. پژوهش حاضر یک گزارش ابتدایی است که در آن تعداد نمونه ها کم است. بنابراین پیشنهاد می‌شود در مراحل بعدی پژوهش بر تعداد نمونه‌ها افزوده شود. همچنین مدت پیگیری بیماران کوتاه بوده و پیشنهاد می‌شود این پیش بینی ها برای مدت طولانی تری در بیماران مورد بررسی قرار گیرند.

استفاده از دسته بندهای دیگر مانند ماشین های بردار پشتیبان می‌تواند در کارهای آینده مورد نظر قرار گرفته و نتایج حاصل از آنها با نتایج این مقاله مورد مقایسه قرار گیرد. با توجه به اینکه هر چه میزان داده های وارد شده به مدل دسته بند بیشتر باشد، نتایج بهتری حاصل می‌گردد، استفاده از مجموعه داده کاملتر در تحقیقات آینده سودمند خواهد بود.

نتیجه گیری:

دسته بند درخت تصمیم که با در نظر گرفتن فاکتور هایی مانند اتیولوژی، سن، شاخص توده بدنی (BMI)، پلاکت، بیلی روبین توتال، INR، کراتی نین، آلفا فیتوپروتئین و آلبومین ساخته شد، توانست احتمال ابتلای بیماران سیروتیک به سرطان کبد را پیش بینی کند.

در مدل داده کاوی بیماران (درنظر گرفتن اتیولوژی ویرال و غیر ویرال به صورت همزمان) با آلفا فیتوپروتئین بالاتر از ۲۲ نانوگرم درصد، کراتی نین بالای ۰/۸، بیلی روبین توتال بالای۰/۸۳، آلبومین بالای ۳ و شاخص توده ی بدنی (BMI) بالای ۲۲، شانس بالاتری برای ابتلا به سرطان کبد دارند. پیشنهاد می‌شود نتایج این تحقیق در مطالعات بزرگتر و چند مرکزی

مـورد آزمایش بالینی قرار گیرند تا میزان مفید بـودن و قابلیت انجام آنها مشخص شود. درخت تصمیم رسم شده به دلیل قابلیت تفسیر بالا می‌تواند به سادگی توسط پزشکان مورد استفاده قرار گیرد.

REFERENCES

1. Han J, Kamber M, & Pei J. 2006. Data mining: concepts and techniques. Morgan kaufmann.
2. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. *CA Cancer J Clin* 2005;55:74-108.
3. Davila JA, Morgan RO, Shaib Y, McGlynn KA, El-Serag HB. Hepatitis C infection and the increasing incidence of hepatocellular carcinoma: a population-based study. *Gastroenterology* 2004;127:1372-80.
4. Fattovich G, Stroffolini T, Zagni I, Donato F.Hepatocellular carcinoma in cirrhosis: incidence and risk factors. *Gastroenterology* 2004;127:S35-50.
5. Di Bisceglie AM. Hepatitis C and hepatocellular carcinoma. *Hepatology* 1997;26:34S-38S.
6. Bellazzi R, Zupan B. Predictive data mining in clinical medicine: current issues and guidelines. *Int J Med Inform* 2008;77:81-97.
7. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. *Artif Intell Med* 2005;34:113-27.
8. Yeh DY, Cheng CH, Chen YW. A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications* 2011;38:89707.
9. Yeh JY, Wu TH, Tsao CW. Using data mining techniques to predict hospitalization of hemodialysis patients. *Decision Support Systems* 2011;50:439-48.
10. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, et al. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med* 2004;32:71-83.
11. Lin RH. An intelligent model for liver disease diagnosis. *Artif Intell Med* 2009;47:53-62.
12. Rajeswari P, Reena G. Analysis of liver disorder using data mining algorithm. *G J C S T* 2010;10;48-52.
13. Wang Y, Ma L, Liu P. Feature selection and syndrome prediction for liver cirrhosis in traditional Chinese medicine. *Comput Methods Programs Biomed* 2009;95:249-57.
14. Kurosaki M, Hiramatsu N, Sakamoto M, Suzuki Y, Iwasaki M, Tamori A,et al. Data mining model using simple and readily available factors could identify patients at high risk for hepatocellular carcinoma in chronic hepatitis C. *J hepatol* 2012;56:602-8.
15. Hajiani E, Hashemi S, Cheraghi M. Evaluation of serum AFP (α-fetoprotein) level in HBsAg carrier patients for diagnosis of hepatocellular carcinoma. *Yafteh* 2006; 8:101-6.
16. Bruix J, Sherman M; American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. *Hepatology* ;53:1020-2.
17. Burke HB, Goodman PH, Rosen DB, Henson DE, Weinstein JN, Harrell FE Jr, et al. Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* 1997;79:857-62.
18. Cabena P, Hadjinian P, Stadler R, Verhees J, Zanasi A. Discovering data mining: from concept to implementation 1998. Prentice-Hall, Inc..
19. Kawaguchi T, Kakuma T, Yatsuhashi H, Watanabe H, Saitsu H, Nakao K,et al. Data mining reveals complex interactions of risk factors and clinical feature profiling associated with the staging of non-hepatitis B virus/non-hepatitis C virus-related hepatocellular carcinoma. *Hepatol Res* 2011;41:564-71.
20. Luk JM, Lam BY, Lee NP, Ho DW, Sham PC, Chen L, et al. Artificial neural networks and decision tree model analysis of liver cancer proteomes. *Biochem Biophys Res Commun* 2007;361:68-73.
21. Zhang Y, Wang S, Li D, Zhnag J, Gu D, Zhu Y, et al. A systems biology-based classifier for hepatocellular carcinoma diagnosis. *PLoS One* 2011;6:e22426.
22. Ikeda M, Okada S, Ueno H, Okusaka T, Kuriyama H. Spontaneous regression of hepatocellular carcinoma with multiple lung metastases: a case report. *Jpn J Clin Oncol* 2001;31:454-8.
23. Bosch FX, Ribes J, Díaz M, Cléries R. Primary liver cancer: worldwide incidence and trends. *Gastroenterology* 2004;127:S5-S16.
24. Parikh S, Hyman D. Hepatocellular cancer: a guide for the internist. *Am J Med* 2007;120:194-202.
25. El-Serag HB. Epidemiology of hepatocellular carcinoma in USA. *Hepatol Res* 2007;37 Suppl 2:S88-94.
26. Bruix J, Sherman M, American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. *Hepatology* 2011;53:1020-2.

Original Article

Using Data Mining for Identify Patients at High Risk to Hepatocellular Carcinoma in the Cirrhosis Liver: Preliminary Report

Melina Ebrahimi Khameneh[1], Mohammad Mehdi Sepehri[2], Mehdi Saberifiroozi[3]

[1] MSc, Department of Industrial Engineering, South branch Islamic Azad University, Tehran, Iran.
[2] Associate Professor, Department of Industrial Engineering, Tarbiat Modares University, Tehran, Iran
[3] Professor, Digestive Disease Research Center, Shariati Hospital, Tehran University of Medical Sciences, Tehran, Iran.

ABSTRACT

Background:
Data mining has an interdisciplinary field including various scientific disciplines such as: database systems, statistics, machine learning, artificial intelligence and the others. In the field of medical, data mining algorithms can help physicians to diagnose diseases and chose the best type of treatment. Hepatocellular carcinoma has the most common type of liver cancer. Given the poor prognosis, Hepatocellular carcinoma (HCC) has the fourth leading cause of cancer-related deaths. In this article we aimed to build a decision support system which helps physicians for identify patients at risk to liver cancer.

Materials and Methods:
We analyzed 258 patients with cirrhosis liver. Patients have followed up for four years. We have used decision tree as a data mining tool, for identify patient at high risk to Hepatocellular carcinoma.

Results:
Decision tree determined the importance of attributes such as creatinine, INR and BMI which could be useful for prediction of cancer. From decision tree model, cirrhosis disease classification rules were extracted and used to improve the prediction of HCC. Decision tree could identify patients at risk to liver cancer with the accuracy of 88% for patients with Sustained virological response (SVR) and the accuracy of 92% for patients with non SVR found.

Conclusion:
According to decision tree results, attributes such as etiology, age, BMI, Platelet, Total Bilirubin, INR, Creatinine, Alfafetoproteina (AFP), and Serum Albumin can predict HCC in patient with cirrhosis. It is suggest that results examine with a greater number of patient.

Keywords: Data mining; Hepatocellular carcinoma; Cirrhosis liver; Prediction; Classification; Decision tree

please cite this paper as:
Ebrahimi Khameneh M, Sepehri MM, Saberifiroozi M. Using Data Mining for Identify Patients at High Risk to Hepatocellular Carcinoma in Cirrhosis Liver: Preliminary Report. *Govaresh* 2015;19:265-74.

Corresponding author:
Mohammad Mehdi Sepehri, PhD
Department of Industrial Engineering, Tarbiat Modares University (TMU), Jalal-e Al-e Ahmad Highway, Tehran 1411713114, Iran
Telefax: + 98 21 82883379
E-mail: mehdi.sepehri@gmail.com
Received: 06 Aug. 2014
Edited: 12 Nov. 2014
Accepted: 13 Nov. 2014