دانشگاه آزاد اسلامی- واحد علوم و تحقیقات قزوین
دانشکده ی مهندسی کامپیوتر

سمینار دوره کارشناسي ارشد مهندسي کامپیوتر-گرایش نرم افزار

موضوع:

بررسی روشهای داده کاوی در تشخیص بیماری عروق کرونری

دانشجو:
الهام بیرجندیان
استاد راهنما:
جناب آقای دکتر صنیعی آباده
استاد مشاور:
جناب آقای دکتر معصومی

زمستان ۱۳۲۹


بسم الله الرحمن الرحیم

تقدیم به

برترینی که به ما فرصت بودن داده به ماه و خورشید آسمان زندگیم که هیچ گاه افول نخواهند کرد، پدر و مادر مهربانم.

به آنان که با عشق از عصاره وجود خود محبت را به ما ارزانی داشتند.

تقدیم به استاد ارجمندم که معلم روح بود و سازنده آینده من و به تمام سال‌های درس و تحصیل.

قدردانی و سپاسگزاری

اساتید فرهیخته ام جناب آقای دکتر صنیعی و جناب آقای دکتر معصومی

که این تجربه گران بها را مدیون زحمات ارزشمند آنانیم.

مفتخرم که در طول مدت انجام این تحقیق از رهنمودهای علمی و اخلاقی آنها بهره‌مند شدم و درگاه خداوند بزرگ را شاکرم که افتخار شاگردی آنها را نصیبم نمود.

واژه ها و کلمات در بیان عمق سپاس و قدردانیتان عاجز و ناتوانند و در مقام سپاسگذاری تنها می گویم اساتید عزیز و گرامی خسته نباشید.

چکیده

امروزه در دانش پزشکی شاهد جمع آوری داده های فراوان در مورد بیماری های مختلف هستیم. مراکز پزشکی با مقاصد گوناگون به جمع آوری این داده ها می پردازند، تحقیق روی این داده ها و به دست آوردن نتایج و الگوهای مفید در رابطه با بیماری ها یکی از اهداف استفاده از این داده هاست. حجم زیاد این داده ها و سردرگمی حاصل از آن مشکلی است که مانع رسیدن به نتایج قابل توجه می شود. ازاین رو از داده کاوی و روش های آن برای حل این مشکل استفاده می کنیم. بیماری های قلبی یکی از پنج عامل اول دلیل مرگ ومیر در جهان محسوب می گردند، شایعترین نارسایی قلبی، گرفتگی عروق کرونر یا رگهای تغذیه کننده ماهیچه قلب می باشد. در حال حاضر استاندارد طلایی برای تشخیص این بیماری، آنژیوگرافی عروق کرونری می باشند که علاوه بر وقت گیر بودن و داشتن هزینه بالا، روشی تهاجمی بوده و ممکن است منجر به خطراتی برای بیمار گردد. لذا پژوهشگران در تلاشند تا با به کارگیری روش های داده کاوی، از طریق روش های غیرتهاجمی نظیر ارزیابی نتایج حاصل از تست ورزش، نوار قلبی، آزمایشات خون و ... وجود یا عدم وجود این بیماری را تشخیص دهند. این گزارش، مقالات تخصصی و تجدید نظرات متنوع درباره پیش بینی و تشخیص بیماری عروق کرونری را جمع بندی می کند. هدف اصلی از این گزارش یک پروژه تحقیقاتی به منظورمقایسه تکنیک های مختلف طبقه بندی در داده کاوی از طریق مقایسه حساسیت، ویژگی و دقت بین آنها، جهت انتخاب دقیق ترین مدل برای پیش بینی بیماری عروق کرونر در افراد مبتلا می باشد.

کلید واژه: داده کاوی، طبقه بندی، بیماری قلبی عروقی، بیماری عروق کرونر، روش تهاجمی.

فهرست مطالب

فصل اول: مقدمه

فصل دوم: مفاهیم و تعاریف اولیه

فصل سوم:بررسی تکنیک های طبقه بندی داده کاوی در تشخیص بیماری عروق کرونری

فصل چهارم: نتیجه گیری و فعالیت های آتی

فهرست جدول ها

فهرست شکل ها

فـصل اول:

مـقـدمـه

۱-۱ مقدمه

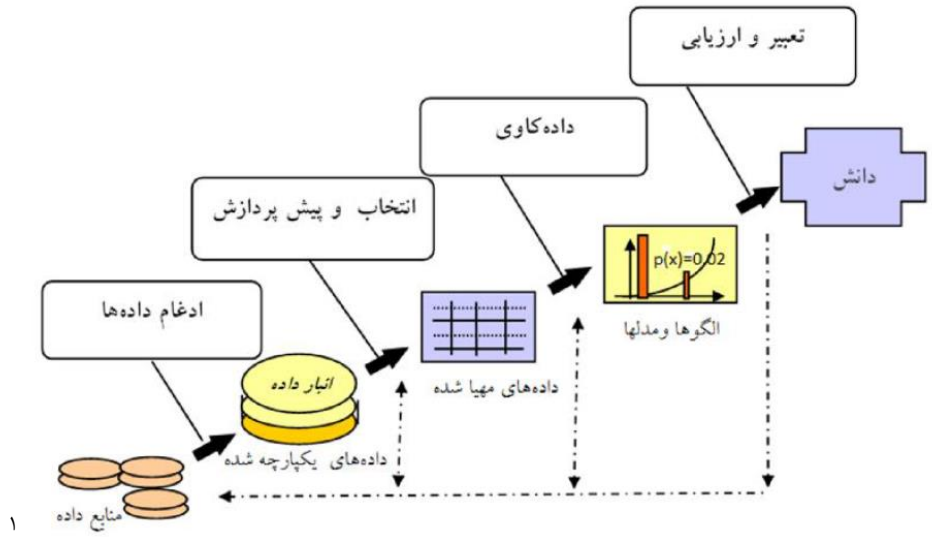
بیماری های قلبی عروقی یکی از شایعترین انواع بیماری ها است که با به خطر انداختن سلامت فرد، آمار بسیار بالایی از مرگ و میر را به خود اختصاص می دهد. در ابتدای قرن بیستم ۱۰٪ کل مرگ و میرها به علت بیماری های قلبی عروقی بود. در انتهای همین قرن موارد مرگ و میر ناشی از بیماریهای قلبی به ۲۵٪ افزایش یافت و امروزه سالانه حدود ۵۰۰۰۰۰ نفر بر اثر این بیماری جان خود را از دست می دهند[۱]. پیش بینی می شود با توجه به روند کنونی تا سال ۲۰۲۵ میلادی بیشتر از ۳۵ تا ۶۰ درصد موارد مرگ و میر در جهان از بیماری های قلبی عروقی ناشی می شود. رشد چشمگیر این بیماریها و اثرات و عوارض آنها و هزینه های بالای که بر جامعه وارد می کند باعث شده که جامعه پزشکی به دنبال برنامه هایی در جهت بررسی بیشتر، پیشگیری، شناسایی زود هنگام و درمان موثر آن باشد. مراکز تحقیقاتی طرح هایی را برای بررسی عوامل دخیل در این بیماری و کنترل آن انجام داده اند.

شایعترین نارسایی قلبی، گرفتگی عروق کرونر[1] یا رگهای تغذیه کننده ماهیچه قلب می باشد. بیماری CAD یک بیماری مزمن است که در آن شریانهای کرونری به تدریج سخت و تنگ می شوند. این بیماری رایجترین شکل بیماری قلبی عروقی در دنیا و علت اصلی حملات قلبی می باشد[۱].

توجه به خطرناک بودن بیماری واینکه اینکه حدود یک سوم تمام مرگها در دنیا در اثر این عارضه رخ میدهد[۲]، تشخیص زودهنگام و پیشگیری به موقع از آن از اهمیت فراوانی برخوردار بوده و از مهمترین زمینه های تحقیقات پزشکی می باشد.

از چند دهه پیش علوم کامپیوتر با علوم دیگر پیوند محکمی خورده است،به گونه ای که با استفاده از شبیه سازی های کامپیوتری بسیاری از مسائل پیچیده دنیای واقعی حل شده است. یکی از زیر شاخه های علوم کامپیوتر داده کاوی و کشف دانش است که روند کلی آن را می توان در شکل ۱-۱ دید.

[1] Coronary Artery Disease (CAD)



شکل ۱-۱: اکتشاف دانش در پایگاه داده ها [۱]

در دهه اخیر محققین توجه خاصی به مسائل پزشکی داشتند و از داده های حاصل از تحقیقات پزشکی که منابع ارزشمندی در جهت بدست آوردن الگوهای مفید در رابطه با بیماری ها می باشند، استفاده کردند. در این گزارش ما تمرکز خود را بر روی استفاده از روش های داده کاوی در بررسی CAD قرار داده ایم. در بخش بعد مقدمه ای از روش های ارائه شده توسط محققین را در این حوزه خواهیم آورد و به شرح هدف اصلی این گزارش می پردازیم.

۱-۲ هدف گزارش

همکاری متخصصان در زمینه کامپیوتر و پزشکی راه حل جدیدی را در تحلیل داده های پزشکی و علی الخصوص بیماری های قلبی عروقی و به دست آوردن الگوهای مفید و کاربردی ارائه می دهد؛ که همان داده کاوی است. در داده کاوی بر خلاف علم آمار به دنبال پیشگویی هستند نه کشف یا اثبات. بدین معنا که با استفاده از روش های داده کاوی به دنبال تایید آنچه از قبل وجود دارد نیستند، بلکه به دنبال مشخص کردن الگوهای از قبل شناخته نشده هستند. همچنین در این کاربرد به دنبال این نیستند که تعیین کنند چه کسانی دارای بیماریهای قلبی عروقی هستند، بلکه به دنبال این مورد هستند که چه عواملی ممکن است در بروز این بیماری نقش بیشتری داشته باشند. به کارگیری روش های داده کاوی در دانش پزشکی در کشور ایران سابقه ی کمی دارد و توانائی های این روش می تواند در به دست آوردن الگوهای مفید کارساز باشد. این دانش کشف شده می‌تواند باعث بهبود کیفیت سرویس به وسیله مدیران مرکزهای درمانی شود و

1

همچنین می‌تواند به وسیله پزشکان استفاده شود تا رفتار آینده بیماران قلبی و عروقی را از روی سابقه داده شده پیش بینی کنند.

برای رسیدن به این هدف سه فاکتور اساسی را باید در نظر بگیریم که این سه فاکتور از بایدهای هر مسئله ی داده کاوی است. این فاکتورها عبارتند از : (۱) گردآوری مجموعه داده ها، (۲) استخراج ویژگی های موثر که روند تشخیص بیماری قلبی از روی آن ها آسان تر می شود و (۳) طبقه بندی ویژگی‌ها که به تشخیص و ارزیابی فاکتورهای ریسکی که باعث افزایش حمله قلبی می شود،کمک می کند.

۱-۳ فرضیات

فرضیاتی که برای حل مسئله مورد نظر می توان در نظر گرفت را در موارد زیر خلاصه می نماییم:

۱. گردآوری داده ها : هر اندازه مجموعه داده ها، که نتیجه تحلیل پزشکان قلب می باشد، بیشتر باشد می تواند در هدف نهایی که تشخیص و دسته بندی بیماری های قلبی عروقی از جمله بیماری عروق کرونری است، بیشتر به ما کمک کند. به علاوه افزایش مجموعه داده ها باعث تعمیم[1] در الگوریتم های طبقه بندی می شود.

۲. استخراج ویژگی ها: مفهوم ویژگی یک مفهوم عام است و انتخاب آن به ماهیت مسئله بر می گردد. در پایگاه داده هایی که برای حل این مسئله ارائه شده است معمولا ویژگی هایی نیز مشخص شده اند. این ویژگی ها توسط یک پزشک خبره استخراج شده اند که بهترین حالت ممکن در مقایسه با روش هایی مانند روش های آماری می باشد.

۳. انتخاب ویژگی ها : انتخاب ویژگی ها فرآیند انتخاب یک زیر مجموعه از ویژگی های مرتبط است که در ساخت مدل های یاد گیری های ماشین مورد استفاده قرار می گیرد. فرض ما در این فرآیند حذف ویژگی های افزونه و غیر مرتبط می باشد تا به این ترتیب در روند یادگیری ماشین سهولت ایجاد شود و ماشین سریعتر به همگرایی برسد. هدف از این مرحله در کار ما این است که در کنار ویژگی های بهینه ای که یک متخصص قلب و عروق استخراج کرده، به یک زیر مجموعه بهینه و به عبارتی بهتر به یک بردار ویژگی کوچکتر برسیم.

۴. طبقه بندی: روشهایی نظیر شبکه های عصبی مصنوعی، ماشین بردار پشتیبان و درختهای تصمیم و... در حل مساله تشخیص CAD استفاده شده است که ما به دنبال روشی هستیم که با توجه به بردار ویژگی های استخراج شده بالاترین عمومیت، دقت و تعمیم را داشته باشد.

[1] Generalization

۱-۴ جمع بندی و ساختار

هدف استفاده از روشهای داده کاوی در تشخیص بیماری های قلبی عروقی را می توان به صورت زیر بیان کرد:

۱. بررسی میزان تاثیر دارو بر بیماری و اثرات جانبی آن.

۲. تشخیص و پیش بینی بیماری.

۳. تعیین روش درمان بیماری .

۴. پیش بینی میزان موفقیت اقدامات پزشکی مانند اعمال جراحی.

۵. تجزیه و تحلیل داده های موجود در سیستم های اطلاعات سلامت(HIS) .

۶. تحلیل عکس های پزشکی .

در ادامه این گزارش کل مطالب به صورت زیر بیان می گردد:

این گزارش حاوی ۴ فصل می‌باشد. در فصل اول مقدمه‌ای کلی از چارچوب و مفاهیم آورده شده است.

در فصل دوم به معرفی و مرور اجمالی داده کاوی، روشهای آن ، تعاریف دسته بندی شامل روش ها، مزایا و معایب می پردازیم.

در فصل سوم به مرور مقالاتی می پردازیم که به پیش بینی عوامل خطرزای قلبی عروقی و تشخیص آن پرداخته اند و این روش ها را مقایسه می نماییم.

فصل چهارم به نتیجه‌گیری و پیشنهاد فعالیت‌های آتی اختصاص خواهد داشت.

فصل دوم:

مفاهیم و تعاریف اولیه

۲-۱ مقدمه

در دنیای امروز با استفاده از داده کاوی می توان به دانشی دست یافت که خود انسان قرن ها بعد این دانش را کسب خواهد کرد. در حقیقت آزمایش هایی از این دسته در زمینه های متعددی نظیر پزشکی، بورس، اوراق بهادار، هواشناسی، بازاریابی، تشخیص کلاهبرداری های بانکی و بیمه ای، تجارت الکترونیک، بیوانفورماتیک و.... وجود دارد. در واقع علم داده کاوی انسان را قادر می سازد که حجم عظیمی از داده ها را مورد پردازش عمیق قرار داده و کلیه نظم هایی را که در عمق داده وجود دارند، همچون طلا در یک معدن طلا به صورت دانشی با ارزش کشف کرده و جهت استفاده عرضه نمایند.

داده کاوی کاربردهای گوناگونی دارد که یکی از آنها پزشکی است:

- تعیین نوع رفتار با بیماران و پیشگویی میزان موفقیت اعمال جراحی .
- تعیین میزان موفقیت روشهای درمانی در برخورد با بیماریهای سخت .
- تشخیص بیماریها براساس انواع اطلاعات (تصاویر پزشکی، مشخصات بیمار احتمالی) .
- تشخیص ناهنجاری هایی که توسط انسان به سختی قابل تشخیص خواهند بود .

در مسئله تشخیص بیماری گرفتگی عروق کرونری هدف کشف الگوهای داده های موجود در بانک داده ی مربوط به بیماری مذکور با کمک الگوریتم های یادگیری ماشین می باشد. الگوریتم های یادگیری مختلفی نظیر شبکه های عصبی، ماشین بردار پشتیبان و درختهای تصمیم در حل مساله تشخیص CAD استفاده شده است.در این فصل ضمن تعریفی مختصر از بیماری گرفتگی عروق کرونری، مفاهیم پایه ای داده کاوی که در این گزارش مروری مورد استفاده قرار گرفته، توضیح داده می شود.

۲-۲ بیماری گرفتگی عروق کرونری

بیماریهای قلبی عروقی در حال حاضر شایعترین علل مرگ را در اکثر نقاط جهان و ایران تشکیل می دهند. شایعترین بیماری قلبی، گرفتگی عروق کرونر می باشد که یکی از دلایل اصلی مرگ و میر جهان است بطوریکه حدودا یک سوم تمام مرگها در دنیا در اثر این عارضه رخ می دهد و تقریبا تمامی افراد مسن تا حدودی دارای اختلال گردش شریانی کرونر می باشند. از این رو تشخیص زودهنگام و پیشگیری به موقع از این بیماری از اهمیت فراوانی برخوردار می باشد. عروق کرونر رگ های خونی هستند که خون دارای اکسیژن را به عضله قلب حمل می کنند. به منظور پمپ کردن خون به سراسر بدن، قلب باید ذخیره دائمی از اکسیژن را داشته باشد.

شکل۲-۱: گرفتگی رگ توسط املاح [۲]

زمانی که خون از بطن چپ خارج می شود به سرخرگ اصلی بدن یعنی آئورت وارد می شود. در همان ابتدای سرخرگ آئورت، نزدیک بالای قلب، دو سرخرگ کرونر منشعب شده است که تحت عنوان سرخرگ های کرونر چپ و راست شناخته می شوند. سرخرگ های کرونر در سطح قلب قرار گرفته و به شاخه های کوچکتر منشعب می شوند سپس به اعماق عضله قلب وارد شده و اکسیژن را به سلول های قلب می رسانند. داخل دیواره سرخرگ ها به طور طبیعی صاف و انعطاف پذیر است که امکان حرکت جریان خون را به راحتی فراهم می کند. طی سال ها داخل دیواره رگ ممکن است با باقیمانده چربی ها پوشیده شود. زمانی که این روند(که آترواسکلروزیس نام دارد)دو سرخرگ کرونر را دربرگیرد، نتیجه بیماری عروق کرونر خواهد بود. ادامه تشکیل این رسوب باقیمانده های چربی که پلاک نام دارند در دیواره عروق، باعث باریک تر شدن سرخرگ ها شده و جریان خون کاهش می یابد. این پلاک ها می توانند جریان خون سرخرگی را به حدی کاهش دهند که باعث آنژین یا حمله قلبی شوند.

آنژین به صورت درد و ناراحتی در قفسه سینه، بازو، گردن یا فک ظاهر می شود و زمانی روی می دهد که عروق کرونر مسدود شده اجازه عبور خون کافی را به عضله قلب نمی دهند. آنژین به طور مشخص طی فعالیت بدنی یا استرس روحی یعنی زمانی که قلب شدیدتر کار می کند و به اکسیژن بیشتری نیاز دارد، روی می دهد و تنها چند دقیقه طول می کشد و با استراحت فروکش می کند.

در حمله قلبی، یک لخته خون معمولا در قسمت باریک رگ شکل می گیرد و راه عبور خون رامسدود می کند. این قطع جریان خون دارای اکسیژن به عضله قلب منجر به صدمه دائمی به بخشی از عضله قلب می شود. برخلاف درد آنژین، درد ناشی از حمله قلبی معمولا بیش از۱۵ دقیقه طول می کشد و با استراحت برطرف نمی شود.

۲-۲-۱ روشهای پزشکی تشخیص بیماری گرفتگی عروق کرونری

دراین بخش به بیان روش های پزشکی در تشخیص بیماری عروق کرونری می پردازیم، که این روش ها به شرح زیر می باشند:

۰۱ نوار قلب:

متخصصین قلب از روشهای گوناگونی برای تشخیص بیماری قلبی استفاده می‌نمایند. یکی از متداول‌ترین آنان گرفتن نوار قلب است.

۰۲ تست ورزش:

همانطور که می‌دانید، قلب در هنگام کار، فعالیت، بازی،ورزش، انجام کارهای بدنی و ... علائم بیماری احتمالی را از خود بهتر و روشن‌تر نشان می‌دهد.

در یک تست ورزشی می‌توان تا حد قابل ملاحظه‌ای (۷۰٪) از احتمال بیماری قلبی اطمینان یافت. بیمار در یک نوار نقاله شروع به راه رفتن می‌کند و در این حال فشار خون وی اندازه‌گیری شده و نوار قلبی دقیقی از وی تهیه می‌شود.

۰۳ اکو کاردیوگرافی:

اکو کاردیوگرافی، روشی است که در آن از امواج صوتی برای تعیین شکل قلب، دهلیزها، بطنها و سایر اجزاء قلب استفاده می‌گردد. اکوی قلب (اکوکاردیوگرافی) یک روش گسترده و غیر تهاجمی است که در آن با استفاده از امواج صوتی بی ضرر برای انسان تصویر اجزاء قلب و میزان سرعت جریان خون تعیین می‌شود. با استفاده از این روش می‌توان نمای دقیقی از دیواره‌های قلبی ، دریچه‌ها و ابتدای سرخرگهای بزرگ را بدست آورد. غیر تهاجمی بودن این آزمون از امتیازهای خاص آن به شمار می‌رود.

ممکن است در تشخیص بیماری قلبی پزشک لازم بداند که از این روش (اکو) استفاده نماید. این عمل بدون درد و ناراحتی است و به سادگی انجام می‌گیرد. در این روش پزشک وضعیت عملکرد اجزای قلب، شدت جریان خون پمپ شده ضخامت دیواره‌ها و... را به تصویر می‌کشد.

۰۴ سی تی اسکن:

از سی‌تی‌اسکن زمانی استفاده می‌شود که پزشک لازم می‌داند وضعیت عروق داخلی قلب را به روش غیرتهاجمی بررسی نماید.

۰۵ آنژیو گرافی:

درحال حاضر مطمئن ترین روش و در واقع استاندارد طلایی جهت تشخیص گرفتگی در شریانهای کرونری آنژیوگرافی است .در صورتی که فردی دچار آنژین صدری، درد در ناحیه قفسه سینه، تنگی آئورت یا نارسایی

قلبی که علت آن مشخص نیست، شده باشد معمولا متخصص قلب انجام عمل آنژیوگرافی را به او توصیه می کند. به کمک آنژیوگرافی تعداد عروق کرونر مسدود شده، محل انسداد، و میزان آن مشخص می شود. آگاهی از این موارد به پزشک امکان می دهد نحوه درمان را تعیین کند.

در صورتی که فردی دچار آنژین صدری، درد در ناحیه قفسه سینه، تنگی آئورت یا نارسایی قلبی که علت آن مشخص نیست، شده باشد معمولا متخصص قلب انجام عمل آنژیوگرافی را به او توصیه می کند.

اما همانطور که میدانیم آنژیوگرافی یک روش تهاجمی بوده که علاوه بر هزینه و وقتگیر بودن منجر به خطراتی برای بیمار می گردد. از طرفی بسیاری از کسانی که آنژیوگرافی می شوند دارای گرفتگی عروق کرونر نیستند و در واقع نیازی به این عمل برای آنها نمی باشد. لذا عمل آنژیوگرافی برای این دسته از بیماران تنها خطرات ناشی از این روش هجومی و هزینه را در بر دارد. از طرف دیگر ۲۵ درصد از افراد مبتلا به این بیماری بدون هیچ گونه علامت قبلی به صورت مرگ ناگهانی یا سکته حاد از دنیا می روند[۲]. بنابراین تصمیم گیری در مورد انجام شدن عمل آنژیوگرافی بسیار حائز اهمیت بوده وتا حد بسیار زیادی به تجربه پزشک معالج بستگی دارد که این تصمیم گیری را برای پزشک دشوار می سازد.

۲-۳ داده کاوی

دنیای مدرن در حقیقت دنیای داده گرا[1] می باشد و ما در محاصره ی داده ها، چه عددی و چه انواع دیگر قرار گرفته ایم. پیشرفت شگفت انگیز فن آوریهای رایانه ای و مجهز شدن به این ابزار امکان جمع آوری اطلاعات دقیق و کامل در زمینه های مختلف را فراهم ساخته است ومنجر به پیدایش ساختارهای داده بسیار حجیم شده است. همچنین حجم بسیار عظیمی از داده ها در پایگاه دادهی شرکتها، دانشگاهها، مراکز دولتی و سایر موسسات بدون استفاده مانده است. دستیابی به اطلاعات نهفته در دادههای حجیم مستلزم مدیریت کارا است و با به کار بردن سیستمهای سنتی این امر میسر نمی باشد. شدت رقابتها درعرصه های علمی، اجتماعی، اقتصادی، سیاسی و حتی نظامی نیز اهمیت عامل سرعت یا زمان دسترسی به اطلاعات را دو چندان کرده است. بنابراین نیاز به طراحی سیستمهای قادر به اکتشاف سریع اطلاعات مورد علاقه ی کاربران، با تاکید بر حداقل مداخله ی انسانی، به خوبی احساس می شود. داده کاوی فرآیندی است که در آغاز دهه۹۰ پا به عرصه ی ظهور گذاشته و با نگرشی نو به مسئله ی استخراج اطلاعات از پایگاه داده می نگرد. این فرآیند یک مرحله فراتر از بازیابی ساده داده ها است و به کاربران اجازه می دهد که دانش جدید را در داده ها کشف

[1] Data Oriented

کنند. داده کاوی یک علم میان رشتهای است وترکیبی از علومی نظیر هوش مصنوعی، تحلیل آمار، بینایی ماشین و پایگاه داده است.

۲-۳-۱ داده و اطلاع

داده نمایشی از واقیعتها، معلومات، مفاهیم، رویدادها یا پدیده ها، برای برقراری ارتباط، تفسیر ،پردازش توسط انسان یا ماشین است. از طرف دیگر واژهی اطلاعات، به معنی دانشی که از طریق خواندن، مشاهده و آموزش بدست می آید، اطلاق می شود و در حقیقت می توان گفت اطلاعات داده هایی هستند که پس از جمع آوری پردازش شده اند و شکل با مفهومی تولید کرده اند. بین داده ها و اطلاعات همانند خبر و اطلاع رابطه وجود دارد. خبری که دریافت می شود، پس از ارزیابی به اطلاع تبدیل می شود. به بیان دیگر اطلاع حاصل تکامل داده ها است. به عنوان مثال وقتی انسان تصویری را مشاهده می نماید، این داده به عنوان یک داده از طریق ورودی یعنی چشم وارد سیستم مغز می شود. مغز پس از دریافت تصویر آن را با تصاویر موجود در حافظه مقایسه می کند .چنانچه تصویر دریافتی با یکی از تصاویر موجود در حافظه مطابقت داشته باشد، اطلاع حاصل می گردد. به این ترتیب بین داده و اطلاع یک شکاف وجود دارد که اندازهی این شکاف با حجم داده ها ارتباط مستقیم دارد. هرچه داده ها حجیم تر باشند، این شکاف بیشتر خواهد بود و هرچه داده ها کمتر و روشها و ابزارها کاراتر باشند، فاصله ی بین داده و اطلاع کمتر خواهد بود. امروزه افزایش سریع حجم پایگاه های داده به گونه ای است که توانایی انسان برای درک این داده ها بدون ابزار کارا کفایت نمی کند.

در این وضیعت تصمیم گیری ها به جای تکیه بر اطلاعات، بر درک مدیران و کاربران تکیه دارند. چرا که تصمیم گیرندگان ابزارکارا برای استخراج اطلاعات با ارزش را در دست ندارند. در واقع شرایط فعلی توصیف کننده شرایطی است که ما از لحاظ داده قوی ولی از لحاظ اطلاعات ضعیف هستیم.

۲-۳-۲ استخراج دانش و داده کاوی

به طور کلی داده کاوی به استخراج دانش[1] از پایگاه های داده ی بزرگ اشاره دارد و در واقع یک نام بی مسمی است که بنا بر مصالحی مورد استفاده قرار می گیرد. بی مسمی بودن از این جهت که به طور معمول استخراج طلا از صخره ها یا شنها طلاکاوی نامیده می شود نه صخره کاوی یا شن کاوی. بنابراین در واقع به داده کاوی باید نام مناسبتر اکتشاف دانش از پایگاه داده را نسبت داد که متاسفانه کمی طولانی است. از طرفی دانش کاوی فاقد تاکید لازم بر کاوش مقادیر بزرگ داده ها است، لذا داده کاوی تنها عبارتی است که

[1] Knowledge Discovery

بر هر دو کلمه داده و کاوش تاکید دارد و از این جهت یک انتخاب عمومی است. گرچه عبارات معادل زیادی در این زمینه وجود دارد که از این جمله می توان دانش کاوی در پایگاه داده ها، استخراج دانش و لاروبی داده ها را نام برد .برخی مولفین مانند چتفیلد، داده کاوی را مترادف عبارت استخراج دانش از پایگاه داده می دانند و برخی دیگر مانند فایاد به داده کاوی به عنوان یک مرحله ضروری از فرآیند بزرگتر استخراج دانش می نگرند که شامل مراحل زیر است:

۰۱ پاکسازی داده ها [1]: حذف داده های نا ایستا و مزاحم.

۰۲ یکپارچه سازی داده ها [2]: ترکیب منابع داده متعدد، پراکنده و ناهمگن.

۰۳ انتخاب صفات [3]: انتخاب صفات مهمی از داده ها.

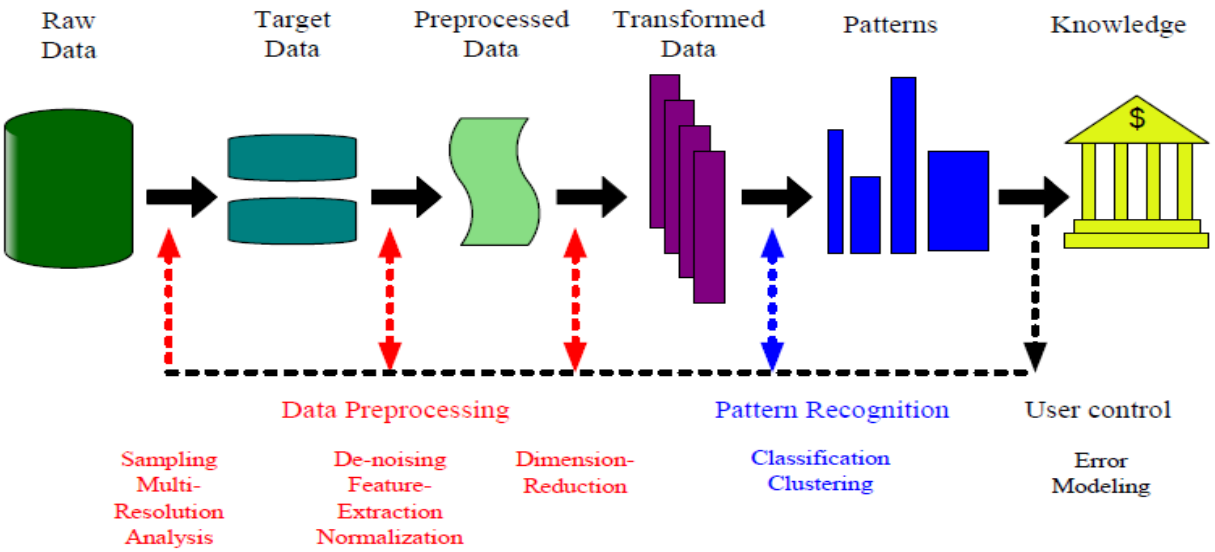
۰۴ تبدیل داده ها : تبدیل یا ترکیب داده ها به اشکالی مناسب برای به کار بردن روشهای مختلف آماری.

۰۵ داده کاوی : مرحله ای ضروری از فرآیند استخراج دانش است که در آن از روشهای مختلف آماری برای استخراج الگوها استفاده می شود.

۰۶ ارزیابی الگوها :شناسایی الگوهای جذاب ارائه دانش بر اساس معیارهای جذابیت.

۰۷ ارائه دانش :ارائه دانش استخراج شده با استفاده از تکنیک های نمایش اطلاعات.

شکل ۲-۲ فرآیند کشف دانش از پایگاه داده را نشان می دهد:



شکل ۲-۲: فرآیند داده کاوی [۳]

[1] Data Cleaning
[2] Data Integration
[3] Feature Selection

اختلاف نظرهایی درباره تعریف دقیق داده کاوی وجود دارد، متخصصین و صاحبنظران بر این باورند که داده کاوی یک گرایش چندمنظوره می باشد که مفاهیمی از یادگیری ماشین، هوش مصنوعی، آمار، محاسبات با کارایی بالا، پردازش سیگنال و تصویر، بهینه سازی ریاضی، تشخیص الگو را در بردارد[۴]. آنچه که تازه است، ترکیب و تلفیق این تکنولوژی ها به منظور تحلیل مجموعه داده های بزرگ می باشد. با کاربرد یافتن داده کاوی در مسائل و حوزه های جدید، پیچیدگی آن نیز رشد می یابد. به عنوان مثال، رشد و توسعه اینترنت و وب گسترده جهانی باعث ایجاد وظایفی نظیر خوشه بندی مستندات متنی، جستجوهای چندرسانه ای و کاوش الگوهای کاربران وب به منظور پیش بینی گشت و گذارهای آینده کاربران و یافتن محل های مناسب برای تبلیغات شده است. تعاریف مختلفی از داده کاوی در مراجع گوناگون وجود دارد. اما شاید جامع ترین تعریف از داده کاوی در مرجع[۵] ارائه شده است:

داده کاوی به فرآیند جستجو و بکارگیری تکنیکهای کامپیوتری به منظور تحلیل داده ها، کشف مدلهای جدید، مختصرسازی داده ها و اخذ مقادیر جدید از مجموعه ای از داده های معلوم اطلاق می شود.

به صورت کلی روش های داده کاوی به دو گروه با ناظر و بدون ناظر تقسیم بندی می شوند[۶,۷]. درروش های بدون ناظر متغیر هدفی تعریف نمی شود و الگوریتم داده کاوی همبستگی ها و ساختارهای بین تمام متغیرها را جستجو می کند. از مهمترین روش های داده کاوی بدون ناظر، خوشه بندی را می توان نام برد. اکثر روش های داده کاوی روش های باناظر می باشند که در این روش ها یک متغیر هدف از قبل تعریف شده وجود دارد. در این روش ها مثال های زیادی وجود دارند که مقدار متغیر هدف برای آنها مشخص می باشد، بنابراین الگوریتم می تواند به کمک آنها آموزش ببیند که کدام متغیر هدف با کدام مقادیر متغیرهای پیش بینی کننده متناظر می باشد. بیشتر روش های داده کاوی با ناظر متدولوژی زیر را برای ایجاد و ارزیابی مدل مورد استفاده قرارمی دهند. ابتدا یک مجموعه آموزشی از داده ها برای الگوریتم ایجاد می شود که شامل مقادیر از پیش دسته بندی شده برای متغیر هدف می باشد. دسته هر یک از نمونه های مجموعه آموزشی باید از پیش مشخص شده باشند. سپس به کمک نمونه های آموزشی که در مجموعه داده آموزشی وجود داشتند یک مدل داده کاوی موقت ایجاد می شود. با این حال مجموعه آموزشی ناکامل است. بدین معنی که داده های جدیدی که مدل سازان تمایل به دسته بندی آنها دارند را در بر ندارد. بنابراین لازم است که از الگوریتم در مقابل مجموعه آموزشی مراقبت کنیم که به صورت کورکورانه، همبستگی های یافته را برای نمونه های جدید به کار نبرد. لذا قدم بعدی در متدولوژی داده کاوی باناظر ارزیابی مدل داده کاوی موقت بر روی یک مجموعه آزمایشی از نمونه ها می باشد. صحت و کارایی مدل داده کاوی موقت با مقایسه

مقادیر صحیح متغیر هدف با مقادیر بدست آمده از مدل موقت ارزیابی می شود و مدل داده کاوی موقت، طوری تنظیم می شود که نرخ خطا در مجموعه آزمایشی کمینه شود .سپس مدل داده کاوی تنظیم شده بر روی یک مجموعه داده اعتباری اعمال می شود که در این مرحله مدل تنظیم شده دقیق تر می گردد تا نرخ خطا بر روی مجموعه اعتباری کمینه گردد.

۲-۳-۳ پیش پردازش ها

عملیات مختلفی نظیر پاکسازی داده ها، یکپارچه سازی و تبدیل داده ها در پیش پردازش داده ها وجود دارد[۸]. در تبدیل داده ها گاهی لازم است برای قرار دادن داده ها در محدوده خاص از روشهای گوناگونی استفاده شود. یکی از این روش ها نرمالسازی می باشد. در این روش هدف، قرار دادن داده ها در یک محدوده مشخص است. فایده اصلی نرمال سازی این است که از چیره شدن صفاتی که در محدوده عددی بزرگتری هستند بر صفات با محدوده عددی کوچکتر جلوگیری می کند. فایده دیگر آن جلوگیری از ایجاد مشکلات عددی در حین محاسبات است. نرمال سازی با روشهای مختلفی قابل انجام است. یکی از این روش ها بکار گیری معادله زیر می باشد:

$$X_{nor} = \frac{X - X_{avg}}{\sqrt{X_{var}}} \qquad (۱-۲)$$

در معادله بالا، X بردار ورودی، X_{avg} بردار میانگین هر بعد بردار ورودی و X_{var} بردار پراکندگی هر بعد بردار ورودی می باشد.

بعد از مرحله پیش پردازش، داده ها برای کشف و یافتن الگو آماده می باشند که این امر به کمک الگوریتم هایی نظیر دسته بندی، خوشه بندی و غیره امکانپذیر می باشد.

۲-۴ انواع روشهای داده کاوی

در این قسمت به چند نمونه از روشهای داده کاوی بطور خلاصه اشاره خواهیم کرد:

۰۱ دسته بندی (Classification):

در این روش یک نمونه به یکی از چند دسته از پیش تعریف شده دسته بندی می شود.

۰۲ رگرسیون (Regression):

پیش بینی یک مقدار متغیر مبنی بر متغیرهای دیگر .

۰۳ خوشه بندی (Clustering):

یک دسته داده را به یکی از چند خوشه نگاشت می کند. خوشه ها گروه بندیهای دسته های داده ای هستند که بر اساس شباهت برخی از معیارها بوجود می آیند.

۰۴ کشف قواعد هم باشی (Association Rule Discovery):

روابط وابستگی بین خصیصه های مختلف را بیان می کند.

۰۵ تحلیل دنباله :

الگوهای دنباله ای همچون سریهای زمانی را مدل می کند.

حال به توضیح کامل دو مورد از مهمترین روشهای داده کاوی می پردازیم:

۲-۴-۱ خوشه بندی

خوشه بندی[1] یکی از مهمترین روش های داده کاوی بدون نظارت است که سعی دارد داده های مشابه را در قالب دسته هایی تا حد امکان متفاوت، گروه بندی کند. نکته ی مهم در این گروه بندی این است که داده های ورودی فاقد متغیر هدف می باشند و رویهی خوشه بندی، علاوه بر تشخیص خوشه های داده باید متغیر هدف نمونه ها را نیز بیان کند. این گروه بندی بر اساس معیاری از شباهت انجام خواهد شد. شباهت یا عدم شباهت میان داده ها بر اساس یک معیار ریاضی بیان می شود. معیارهای گوناگونی برای ارزیابی کارایی الگوریتمهای خوشه بندی مورد استفاده قرار می گیرد. از جمله ی این معیارهای می توان به میزان خطای هر خوشه، تعداد خوشه های تولید شده را نام برد. الگوریتم های خوشه بندی به دو دسته کلی تقسیم می شوند:

۱. خوشه بندی سلسله مراتبی
۲. خوشه بندی افرازی

در خوشه بندی سلسله مراتبی، تمام نمونه ها در یک سلسله مراتبی سازمان دهی می شوند که درجه مشابهت بین این نمونه ها توصیف می شود. خوشه بندی افرازی یک افراز از داده ها ایجاد می کند به طوری که هر نمونه در خوشه ای قرار می گیرد. بنابراین در این نوع خوشه بندی اطلاعات کمتری بدست می آید اما توانایی تعامل با تعداد زیادی از نمونه ها بهبود می یابد. در ادامه یک روش مهم خوشه بندی یعنی K-Means شرح داده خواهد شد.

[1] Clustering

۲-۴-۱-۱ روش K-Means

هدف این روش، قرار دادن مجموعه ای از داده ها X=(x1, x2, …, xn) در تعدادی خوشه می باشد به طوری که داده های موجود در یک خوشه تا حد ممکن مشابه و داده های موجود در خوشه های متفاوت تا حد ممکن متمایز باشند. تعداد خوشه ها، K ، یک عدد مثبت صحیح است که باید در ابتدای اجرای الگوریتم مقدار دهی اولیه شود. خوشه بندی براساس کمینه کردن مجموع مربعات فاصله بین نمونه ها نسبت به مرکز ثقل نمونه ها صورت می گیرد. در مرحله اول تعداد خوشه ها مشخص می شود. بعد از مشخص شدن تعداد خوشه ها مراکز ثقل آنها تعیین می گردد. نزدیکی بین دو نمونه توسط توابع مختلفی قابل محاسبه است که از معروف ترین آنها می توان به تابع فاصله اقلیدسی اشاره کرد. سپس مرکز ثقل خوشه با استفاده ازبردار میانه محاسبه می شود و با مرکز فعلی قیاس می گردد در صورتی که این دو مرکز متفاوت باشند مرکز ثقل جدید جایگزین مرکز فعلی می شود. شرط خاتمه الگوریتم این است که تفاضل میانه جدید و میانه فعلی از یک مقداری کوچکتر شود. از مزایای این الگوریتم به سادگی آن می توان اشاره کرد همچنین این الگوریتم همگرائی سریعی در رسیدن به جواب نهایی دارد. از معایب آن میتوان به این اشاره کرد که چون در ابتدا تعداد خوشه ها به الگوریتم داده می شود بهینه بودن کلی جواب بدست آمده را نمیتوان تضمین کرد چرا که تعیین تعداد خوشه ها به صورت دستی است و اشتباه در تعیین این عدد در جواب نهایی تاثیر زیادی دارد. همچنین کارایی این الگوریتم بسیار وابسته به نحوه تعیین مراکز اولیه برای خوشه ها است.

۲-۴-۲ دسته بندی

رایج ترین عمل داده کاوی با ناظر دسته بندی[1] می باشد. در دسته بندی یک متغیر هدف گروهی وجود دارد که به دسته ها و گروه های از قبل تعیین شده افراز می گردد.

هدف در عمل دسته بندی، ایجاد یک دسته بند توانا برای تخصیص متغیر هدف متناظر با هر نمونه ورودی، بر اساس مقادیر متغیرهای آن نمونه می باشد. برای این منظور ابتدا داده های ورودی به دو دسته تقسیم می شوند:

۱. داده آموزش
۲. داده های تست

[1] Classification

الگوریتم با استفاده از داده های که برچسب آموزش خورده اند، یک دسته بند ایجاد می کند سپس کارایی این دسته بند با استفاده از نمونه های تست ارزیابی می شود.

۲-۴-۲-۱ کارایی الگوریتم های دسته بندی

دو معیار عمده برای ارزیابی کارایی الگوریتم های دسته بندی عبارتند از:

نرخ دسته بندی و قابلیت تفسیر:

نرخ دسته بندی میزان دقت الگوریتم را برای دسته بندی درست نمونه های ورودی را می سنجد در حالی که قابلیت تفسیر به میزان سادگی و قابلیت توسعه روش دسته بندی اشاره دارد. الگوریتم های مختلفی برای عمل دسته بندی ارائه شده اند که از مهمترین آنها می توان به رگرسیون خطی، درخت تصمیم، شبکه های عصبی، روشهای فرااکتشافی، سیستمهای فازی و ماشین بردار پشتیبان اشاره کرد. از بین این روشها، دو روشهای ماشین بردار پشتیبان و شبکه های عصبی دارای دقت بهتری نسبت به سایر روشها هستند. اما قابلیت فهم این روشها (ماشین بردار پشتیبان با درجه ی غیر خطی) به دلیل استفاده از توابع پیچیده ریاضی بسیار پایین است[۹]. بنابراین در کاربردهایی که هم دقت و قابلیت تفسیر اهمیت دارند، روشهای ماشین بردار پشتیبان(با توابع غیر خطی)و شبکه های عصبی نمی توانند گزینه های مناسبی باشند.روشهای فرااکتشافی، درخت تصمیم و سیستم های فازی دارای قابلیت تفسیر بسیار مناسبی هستند. ضمن اینکه این روشها دارای دقت مناسبی نیز می باشند.

۲-۴-۲-۲ گروه بندی الگوریتم های دسته بند

به طورکلی الگوریتم های دسته بندی را می توان به چهار گروه مختلف تقسیم کرد[۳] که عبارتند از:

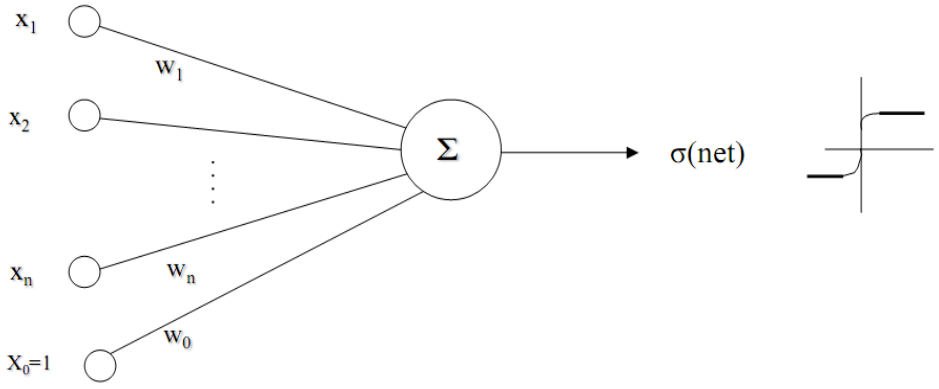
دسته بندی منطقی یا نمادین، دسته بندهای مبتنی بر پرسپترون، دسته بندهای آماری و دسته بندهای مبتنی بر نمونه.

۰۱ دسته بند های منطقی یا نمادین:

الگوریتم های یادگیری نمادین به دستهای از الگوریتم ها اطلاق می شود که داده های ورودی را گرفته و یک مجموعه قواعد که نشان دهنده ی ارتباط بین صفتها و کلاس ها می باشد را برمی گردانند. درخت تصمیم از جمله الگوریتم های نمادین می باشد.

۰۲ دسته بند های مبتنی بر شبکه عصبی

الگویی برای پردازش اطلاعات می باشند که با تقلید از شبکه های عصبی بیولوژیکی مثل مغز انسان ساخته شده اند. عنصر کلیدی این الگوساختار جدید سیستم پردازش اطلاعات آن می باشد و از تعداد زیادی عناصر (نرون) با ارتباطات قوی داخلی که هماهنگ با هم برای حل مسائل مخصوص کار می کنند تشکیل شده اند.



شکل ۲-۳: یک واحد سلولی از پروسپترون چند لایه [۳]

۰۳ دسته بند های آماری

این گروه از دسته بندها بر خلاف سایر دسته بندها، میزان عضویت یک نمونه به هر کلاس را با یک احتمال نشان می دهند[۱۰]. به عبارت دیگر، یک الگوریتم یادگیری آماری به صورت احتمالی میزان تعلق هر نمونه را به هر کلاس محاسبه می کند .شبکه های بیزین[1] از رایجترین روشهای متعلق به این دسته می باشند.

۰۴ دسته بندهای مبتنی بر نمونه

در روشهایی که تاکنون بررسی شد، سعی بر این بود که با استفاده از مثالهای آموزشی تابعی پیداکنیم که بتواند توصیف کننده داده ها باشد. در روش دسته بندی مبتنی بر نمونه فقط مثالها را ذخیره می کنیم و هرگونه تعمیم تا مشاهده مثال جدید به تعویق می افتد. به همین دلیل این روش گاهی روش تنبل[2] نیز نامیده می شود. با مشاهده مثال جدید رابطه آن با نمونه های ذخیره شده بررسی شده و یک مقدار برای تابع

[1] Bayesian
[2] Lazy

هدف آن نسبت داده می شود. در این روش یک فرضیه عمومی مشخص برای داده ها به دست نخواهد آمد بلکه دسته بندی هر نمونه جدید هنگام مشاهده آن و بر اساس نزدیکترین مثالهای ذخیره شده انجام خواهد شد.

از مزایای این روش این است که می تواند توابع پیچیده را مدل کند و اطلاعات موجود در مثالهای آموزشی از بین نمی رود. از مشکلات این روش این است که دسته بندی داده جدید می تواند پرهزینه باشد. زیرا در مرحله آموزش عملی صورت نمی پذیرد و تمامی محاسبات در هنگام دسته بندی انجام می گردند. تعیین یک تابع فاصله مناسب مشکل است و ویژگی های نامرتبط تاثیر منفی در معیار فاصله دارند.

۲-۴-۳ دسته بندی قوانین "اگر-آنگاه" فازی

در داده کاوی دانش کشف شده غالبا به شکل قوانین پیشگویانه یا قوانین دسته بندی اگر- آنگاه به صورت زیر نمایش داده می شوند:

IF<condition>THEN<class>

بخش <condition> یا مقدم قانون، شامل ترکیب منطقی از ویژگی های پیشگویی کننده و به شکل Term1 AND Term2 AND… .

هر term یا عبارت شامل یک سه تایی به شکل < attribute, operator, value > است، بخش < class > یا نتیجه قانون شامل کلاس پیش بینی شده برای نمونه هایی است که ویژگیهای پیش بینی کننده آن، بخش < condition > قانون را برآورده می کند.

در سالهای اخیر سیستم های فازی مبتنی بر قانون برای دسته بندی مورد استفاده قرار گرفته اند که در اینگونه مسائل بردارهای ورودی غیر فازی در دسته های داده شده قرار می گیرند. قوانین اگر - آنگاه فازی برای یک مسئله دسته بندی فازی با C دسته و n خاصیت به صورت زیر نوشته می شوند[۱۱].

Rule R_j: if x_1 is A_{j1} and…and x_n is A_{jn} Then class C_j with CF_j (معادله ۲-۲)

که $X=(X_1….X_n)$ یک بردار الگوی n بعدی و یک مقدار زبانی مقدم نظیر Small ، Large، Don’t Care وغیره ... ، هستند. C_j دسته نتیجه، $CF_j=[0,1]$ درجه قطعیت قانون فازی R_j و N تعداد قوانین" اگر- آنگاه " فازی می باشند.برای به دست آوردن دسته نتیجه و درجه قطعیت هر قانون" اگر- آنگاه " فازی از روال زیر استفاده می شود.

۰۱ محاسبه درجه سازگاری هر الگوی آموزشی با قانون if-then فازی که به کمک عمل حاصل حاصلضرب به دست می آید.

۰۲ محاسبه مجموع درجه های ناسازگاری برای هر دسته.

معادله(۲-۳)

$$\beta_{class\ h}(R_j) = \sum_{x_p \in class\ h} \mu_{R_j}(x_p) \qquad h=1, 2,..,c$$

۰۳ محاسبه دسته نتیجه C_j که بیشترین مقدار $\beta_{class\ h}(R_j)$ را در بین c دارد.

اگر بیش از یک دسته دارای بیشترین مقدار باشند آنگاه دسته نتیجه به صورت یکتا مشخص نمی شود و در این حالت مقدار تهی را به عنوان دسته نتیجه قانون در نظر میگیریم.

۰۴ پس از تولید قوانین فازی، نوبت به تعیین کلاس برای یک الگوی جدید ($X_p =(X_{P1},\ldots,X_{Pn})$ می رسد. معمولا از وزندهی قوانین به عنوان یک مکانیسم ساده برای هماهنگ کردن قوانین استفاده می گردد[۱۲]. روشهای مختلفی برای تعیین کردن وزن برای قوانین فازی وجود دارد [۱۳,۱۵]. مشهورترین آنها ضریب اطمینان یا CF می باشد که در بعضی از مقاله ها اطمینان [۱] نامیده می شود. هنگامی که دسته نتیجه C_j تعیین شد، به کمک رابطه درجه قطعیت مشخص می شود.

به کمک روال فوق برای هر قانون دسته نتیجه و درجه قطعیت هر قانون مشخص می شود. پس از وزندهی قوانین باید به طریقی قوانین را به سوی یک نتیجه هماهنگ نمود. در [۱۶] دو تک قانون روش برای استدلال قوانین فازی توضیح داده شده است. این دو روش عبارتند از «رای گیری[۲] » و « تک قانون برنده[۳] ».

۲-۵ نتیجه گیری

به عنوان نتیجه گیری این فصل می توان بیان کرد که روش های طبقه بندی که مورد هدف ما هستند می توانند با بهره گیری از دانش پزشکان که در قالب بردارهای ویژگی بیان می شوند ما را به نتایجی برسانند که رسیدن به این نتایج توسط یک فرد خبره نیز زمان بر و در برخی از موارد غیر قابل دسترسی می باشد. چرا که روابط بین ویژگی ها، روابط خطی نیستند که انسان بتواند به راحتی از پس حل آنها بر آید.

[1] Confidence
[2] Single winner Rule
[3] Voting

در کل ما به دنبال یک سیستم داده کاوی برای استخراج قوانین برای وقایع CAD هستیم. قوانین استخراج شده می تواند بنا به درخواست کاربر نهایی به گروه هایی تقسیم شود که از آن جمله می توان به دو گروه از عوامل خطرزا یعنی عوامل پرخطر و کم خطر اشاره کرد [۱۷].

انتظار می رود که روش های داده کاوی بر اساس تکنیک های طبقه بندی می توانند در شناسایی زیر گروه هایی از افراد در معرض خطر برای توسعه رویدادهای آینده کمک کند و ممکن است یک عامل تعیین کننده برای انتخاب درمان هایی مانند آنژیوپلاستی و یا عمل جراحی باشد.

فصل سوم

بررسی تکنیک های طبقه بندی داده کاوی در تشخیص بیماری عروق کرونری

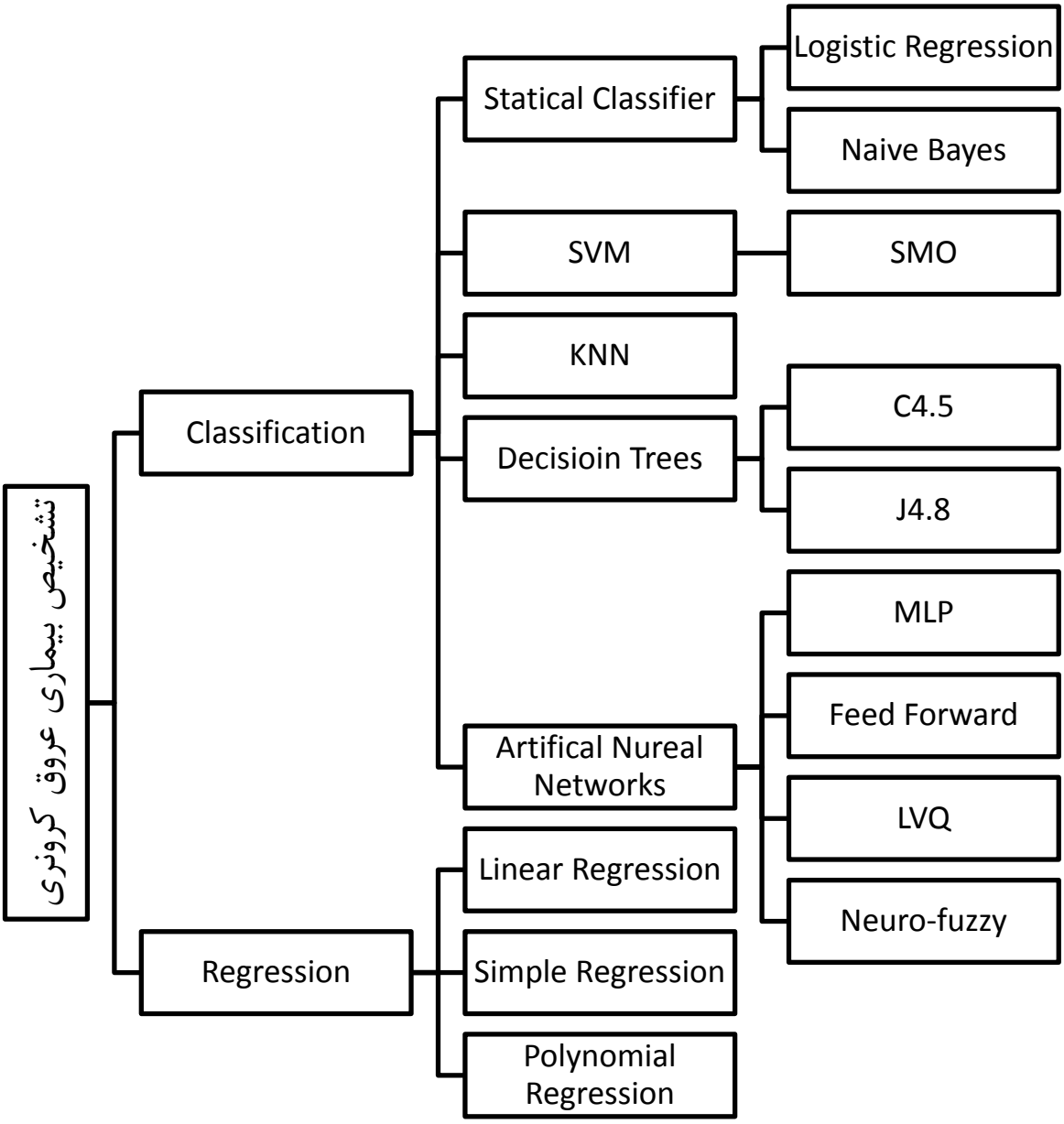
۳-۱ مقدمه

در سال های اخیر پژوهشگران زیادی تلاش نموده اند تا سیستم هایی ارائه دهند که در زمینه های مختلف پزشکان را برای تشخیص هرچه سریع تر و دقیق تر یاری رسانند، بطوری که امروزه داده کاوی پزشکی[1] یکی از شاخه های مهم در زمینه داده کاوی محسوب می گردد. حوزه کار داده کاوی پزشکی بسیار وسیع بوده و زمینه های مختلفی را در بر می گیرد.

یکی از حوزه های مهم و پرکاربرد داده کاوی پزشکی، تشخیص بیماری می باشد .بیماری های مختلفی از جمله آلزایمر[۱۸,۱۹]، مشکلات کبدی[۲۰]، تیروئید[۲۱]، دیابت[۲۲] و ... موضوع تحقیق پژوهشگران در سال های اخیر بوده اند .بیماری های قلبی عروقی نیز به عنوان یکی از شایع ترین و مهم ترین انواع بیماری ها که بنا به گزارش انجمن قلب آمریکا[2] یکی از پنج عامل اصلی مرگ در جهان است، در تحقیقات بسیاری موضوع کار پژوهشگران قرار گرفته است. در مقالات مختلف انواع مختلف بیماری های قلبی مورد توجه واقع شده اند که به تعدادی از آن ها اشاره می گردد.

در ادامه به مرور ادبیات تحقیق به صورت مطالعه مروری خواهیم پرداخت. سپس به بررسی دقیق تر مقالاتی می پردازیم که نقش موثرتری در این حوزه تحقیقاتی داشتند. به این منظور دسته بندی بر اساس روش های طبقه بندی و رگرسیون که در مقالات ارائه شده است در شکل ۳-۱ نمایش می دهیم.

[1] Medical Data Mining
[2] American Heart Association (AHA)



شکل۳-۱: درختواره مقالات

۲-۳ روشهای داده کاوی در تشخیص بیماری های قلبی عروقی

در این قسمت به مطالعه مروری راه کارهای تشخیص بیماری های قلبی عروقی با روش های داده کاوی می پردازیم و در بخش بعدی به توضیح جامع تر تکنیک های طبقه بندی داده کاوی انجام شده در تشخیص بیماری عروق کرونر خواهیم پرداخت.

یکی از نارسایی های مهم قلب مشکلات مربوط به دریچه های قلبی است .زنجیره ای از تحقیقات در زمینه تشخیص این نارسایی، صورت گرفته است که در سال ۲۰۰۲ توسط آقای تورکگلو[1] و همکارانشان با پیشنهاد یک سیستم خبره شروع شد[۲۳]. این سیستم خبره از یک شبکه عصبی استفاده می کرد. نرخ طبقه بندی صحیح در این سیستم در مورد افراد سالم برابر ۰.۹۴ و در مورد افراد بیمار ۰.۹۵ حاصل شد.

آقای سنگور[2] در سال ۲۰۰۸ استفاده از آنالیز عناصر اصلی[3] و سیستم ایمنی مصنوعی[4] را برای تشخیص نرمال و غیر نرمال بودن دریچه های قلبی با استفاده از سیگنال های داپلر بررسی کرد .اعتبار روش پیشنهادی با استفاده از پارامترهای حساسیت و اختصاصی بودن ارزیابی شد که برای حساسیت مقدار ۹۵,۹٪ و برای اختصاصی بودن میزان ۹۶٪ حاصل گردید[۲۴].

استفاده از آنالیز جدا کننده های خطی[5] و سیستم استنتاجی تطبیقی فازی- عصبی[6] برای تشخیص بی نظمی های دریچه های قلبی در[۲۶] بررسی شد و به ترتیب حساسیت و اختصاصی بودن ٪۹۵.۹ و٪۹۴ حاصل گردید.

سیستم هوشمند بعدی از ترکیبی از الگوریتم ژنتیک و ماشین بردار پشتیبان استفاده نمود[۲۶] نرخ دسته بندی بندی این سیستم که GSVM8[7] نامیده شد، برابر ۰.۹۵ گزارش شده است.

در سال ۲۰۰۹ آقایان سنگور، تورکگلو و داس تحقیقات خود را با پیشنهاد یک رویکرد تجمیعی ادامه دادند[۲۷]. این سیستم که از سه شبکه عصبی تشکیل می شد، به خوبی توانست به افزایشی در نتیجه کلی دست یابد به طوریکه دقت دسته بندی به ٪۹۷.۴ رسید و حساسیت و اختصاصی بودن به ترتیب برابر ٪۹۶ و ٪۱۰۰شدند.

دیگر بیماری قلبی بررسی شده در تعدادی از پژوهش ها، بیماری میوکاردیال قلبی می باشد .آقای تاسای در زمینه تشخیص این بیماری از روی تصاویر اکوکاردیوگرافی یک سلسله تحقیقات انجام داده ند. در ابتدا یک

---

[1] Turkoglu
[2] Sengur
[3] Principle Component Analysis (PCA)
[4] Artificial Immune System (AIS)
[5] Linear Discriminate Analysis (LDA)
6 Adaptive Neuro– Fuzzy Inference System (ANFIS)
[7] Generic Support vector Machine

شبکه عصبی سه لایه طراحی شد که ضرایب توزین این شبکه ازطریق یک الگوریتم پس انتشار مشخص می شدند .دقت به دست آمده از این روش، ۸۲,۱٪ گزارش شد[۲۸].

در تحقیق بعدی[۲۹]. یک شبکه عصبی طراحی شد که در آن به جای استفاده از الگوریتم پس انتشار، از ژنتیک الگوریتم استفاده گردید و دقت را تا ۸۸,۱ ٪ افزایش داد .سپس یک روش دسته بندی فازی به منظور بهبود عملکرد تشخیص بیماری میوکاردیال پیشنهاد گردید که دقت را به ۹۱,۴ ٪ رساند[۳۰].

نهایتا در[۳۱]، یک الگوریتم ژنتیک جهت بهینه سازی پارامترهای قوانین فازی پیشنهاد شد .در این روش دقتی برابر ۹۵,۹ ٪ همراه با حساسیت ۹۱,۷ ٪ و اختصاصی بودن ۱۰۰٪ گزارش گردید.

در تحقیقی که در سال ۲۰۰۳ توسط آقای یان[1] و همکارانشان صورت گرفت[۳۲]، یک مدل کامپیوتری براساس شبکه عصبی پرسپترون چند لایه، به منظور توسعه یک سیستم پشتیبان تصمیم، برای تشخیص پنج بیماری اصلی قلب(فشارخون، گرفتگی عروق کرونری، مشکلات دریچه های قلبی ناشی از روماتیسم، بیماری های قلبی مادرزادی و نارسایی قلبی ریوی) به کار برده شد. از خصوصیات این روش یکی سرعت یادگیری انطباقی و دیگری توانایی حل مشکل داده های گم شده است(با استفاده از جایگزینی). دقت این سیستم برابر ۶۳,۳ ٪ محاسبه شد.

در مطالعه بعدی[۳۳]، آن ها سیستمی طراحی نمودند تا بتواند ویژگی های اساسی و لازم برای بیماری قلبی را انتخاب نماید .این کار با استفاده از الگوریتم ژنتیک صورت پذیرفت .همچنین برای هریک از این ویژگی ها، با توجه به میزان اهمیت شان برای هریک از پنج بیماری اصلی قلب، وزنی اختصاص یافت. دقت این روش ۹۲ ٪، همراه با اختصاصی بودن ٪ ۹۶ و حساسیت ٪ ۹۸ گزارش شده است.

۳-۳ تکنیک های طبقه بندی در تشخیص بیماری عروق کرونری

۳-۳-۱ درخت تصمیم (Decision Tree)

درخت های تصمیم گیری الگوریتم های طبقه بندی قدرتمندی می باشند. الگوریتم های محبوب درخت تصمیم گیری شامل ID3، C4.5، وCART ی Breiman و همکاران می باشد[۳۴]. همانطور که از نامش پیداست، این روش به صورت بازگشتی مشاهدات را به شاخه های جدا برای ساخت یک درخت به منظور بهبود دقت پیش بینی جدا می کند. در انجام این کار، الگوریتم های ریاضی متغیر و آستانه مربوط به آن متغیر را که مشاهدات ورودی را به دو یا چند زیر گروه تجزیه می کند را شناسایی می کند. این مرحله در هر

[1] Yan

گره برگ تکرار می شود تا زمانی که درخت کامل ساخته شود. هدف الگوریتم تقسیم ، پیدا کردن یک جفت متغیر آستانه است که حداکثر همگن بودن نتیجه دو یا چند زیر گروه از نمونه ها را داشته باشد[۳۴].

۳-۳-۱-۱ الگوریتم C4.5

C4.5 یکی از الگوریتم های درخت تصمیم گیری است. با استفاده از هرس[1]، معیار نرخ بهره[2]، قادر به مدیریت داده های مداوم است. C4.5 برای بهبود الگوریتم ID3 ارائه شد.

الگوریتم c4.5 به شرح زیر می باشد[۱۷]:

۰۱ ایجاد یک گره بنام Nd.

۰۲ اگر همه مشاهدات در مجموعه داده های آموزشی دارای مقدار خروجی همان کلاس یعنی C، بود سپسReturn کن Nd به عنوان یک گره برگ با برچسب C.

۰۳ اگر لیست ویژگی خالی است، سپس Nd را به عنوان گره برگ با برچسب مقدار خروجی کلاس اکثریت در مجموعه داده های آموزشی برگردان.

۰۴ اعمال روش انتخاب معیارهای تقسیم[3](جداسازی) به مجموعه داده های آموزشی در جهت پیدا کردن "بهترین " معیار تقسیم ویژگی.

۰۵ برچسب کن گره ND را با معیار تقسیم ویژگی.

۰۶حذف معیار تقسیم ویژگی از لیست ویژگی.

۷) به ازای هر مقدار J در معیار تقسیم ویژگی :

الف) D_j را مشاهدات در مجموعه داده آموزشی با مقدار ویژگی J در نظر بگیرید.

ب) اگر D_j خالی (بدون مشاهدات) است، سپس یک گره برگ با برچسب مقدار خروجی کلاس به گره ND Attach ، کنید.

[1] Purnuning
[2] Gain ratio
[3] Seprating Criteria

ج) در غیر اینصورت Attach کنید گره بازگردانده شده توسط درخت تصمیم تولید شده را (D_j لیست ویژگی، روش انتخاب معیارهای تقسیم) به گره ND.

۰۸پایان For.

۰۹ Return کن گره Nd.

با توجه به اینکه بیماری گرفتگی شریان های کرونر، شایع ترین بیماری قلبی بوده و مرگ و میر بسیاری را سالانه به خود اختصاص می دهد، لذا در بین سایر بیماری های قلبی، بیشترین آمار به کارگیری داده کاوی پزشکی را به خود اختصاص می دهد .عموما تحقیقات به صورت مجزا و با مجموعه داده های متنوع که از روش های مختلف غیرتهاجمی، به دست آمده اند، انجام شده است.

در اولین مطالعه Minas A. Karaoli و همکاران[۱۷]، محققین بیماری عروق کرونری را در منطقه پافوس در قبرس مورد بررسی قرار دادند و در مجموع ۵۲۸ نفر را به عنوان نمونه در نظر گرفته و بعد از استخراج ویژگی ها، الگوریتم درخت تصمیم C4.5 را با استفاده از پنج معیار تقسیم متفاوت پیاده سازی نمودند.

در این مطالعه مهمترین عوامل خطر، که از تجزیه و تحلیل قوانین طبقه بندی استخراج شدند برای سه رویداد زیر در نظر گرفته شدند، عبارت بودند از:

۱) برای MI(سکته قلبی): سن، سیگار کشیدن، و سابقه فشار خون بالا؛
۲) برای PCI(مداخله کرونری از راه پوست): سابقه خانوادگی، سابقه فشار خون بالا و سابقه دیابت؛
۳) برای CABG(جراحی بای پاس عروق کرونر): سن، سابقه فشار خون بالا و سیگار کشیدن.

در واقع عوامل خطرزای در نظر گرفته شده برای CAD عبارت بودند از: سیگار کشیدن، فشار خون بالا، دیابت، کلسترول، لیپوپروتئین با چگالی بالا، لیپوپروتئین کم چگالی، تری گلیسیرید. هدف توسعه یک سیستم داده کاوی برای ارزیابی عوامل خطرزای بیماری قلبی با هدف کاهش حوادث CAD است.

تجزیه و تحلیل داده کاوی با استفاده از الگوریتم درخت تصمیم C4.5 با استفاده از پنج معیار های مختلف برای تقسیم استخراج قوانین بر اساس عوامل خطر فوق الذکر انجام شد و همچنین برای جلوگیری از Overfitting در درخت تصمیم از هرس استفاده کردیم که الگوریتم هرس پایین به بالا را با استفاده از برآورد خطای Laplace اجرا کردیم. در حالی که درخت تصمیم گیری ساخته شده است و یک گره برگ ایجاد می شود، پس از آن خطا لاپلاس به شرح زیر برآورد می شود:

$$E(D) = \frac{N - n + m - 1}{N + m} \qquad (۱-۳)$$

ابتدا داده ها از ۱۵۰۰ فرد مبتلا به CAD بین سال های ۲۰۰۳-۲۰۰۶ و ۲۰۰۹ (۳۰۰ نفر در هر سال) با توجه به یک پروتکل مشخص، تحت نظارت متخصص قلب(دکتر J Moutiris، نویسنده دوم این مقاله) در گروه قلب و عروق، در بیمارستان عمومی پافوس در قبرس جمع آوری شد. سپس روی داده ها پیش پردازش هایی نظیر تمیز کردن داده ها،پر کردن مقادیر از دست رفته صورت گرفت و بعد با استفاده از معیارهای تقسیم مانند بهره اطلاعات[۱]، شاخص جینی[۲]، نرخ احتمال آمار مجذور مربع[۳]، نرخ بهره[۴] و اندازه گیری فاصله[۵][۳۵] و الگوریتم C4.5 درخت تصمیم و ۳ مدل طبقه بندی زیر ایجاد شد:

۱) MI: MI در مقابل غیر MI. افراد با سکته قلبی علامت دار و بقیه را به صورت بدون علامت مشخص شدند.

۲) PCI: PCI در مقابل غیر PCI . افرادی با داشتن تنها PCI علامت دار و بقیه به صورت بدون علامت مشخص شد. افراد دارای هم PCI و هم MI از مطالعه حذف شدند.

۳) CABG: CABG در مقابل غیر CABG است. افراد با داشتن تنها CABG علامت دار و بقیه به صورت بدون علامت مشخص شد. افراد دارای هم CABG و هم MI حذف شدند.

آزمون مجموع رتبه Wilcoxon نیز برای بررسی اینکه اختلاف قابل توجهی بین پنج معیار تقسیم استفاده شده در مدلهای درخت تصمیم در سطح p کمتراز ۰,۵ صورت گرفته است یا نه انجام میشود.

بالاترین درصد طبقه بندی صحیح به دست آمده به ترتیب: ۶۶٪، ۷۵٪ و ۷۵٪ برای MI، PCI، مدل CABG بود که می توان بطور کامل در جدول زیر مشاهده کرد.

[1] Information Gain
[2] Gini Index
[3] LikeliHood Chi Square Static
[4] Gain Ratio
[5] Distance Measure

جدول۳-۱ نتایج طبقه بندی مدلهای CABG،PCI،MI

	%CC			%TP			%FP			Sensitivity			Specificity		
	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A	B	A	B+A
	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me(m,M)	Me	Me	Me	Me	Me	Me
MI															
IG	58(57,64)	61(60,63)	62(61,65)	64(60,76)	68(61,73)	67(53,68)	48(44,55)	45(41,49)	37(25,47)	58	60	63	60	64	63
GI	61(59,63)	61(59,63)	63(61,66)	67(55,71)	59(55,71)	63(57,76)	47(41,59)	36(33,48)	39(25,51)	59	60	62	61	62	64
X2	58(57,60)	61(59,63)	63(62,65)	65(63,73)	63(59,76)	64(59,72)	49(47,53)	39(35,59)	36(35,47)	57	62	64	59	61	64
GR	60(58,61)	59(59,59)	62(61,64)	65(53,72)	59(55,67)	65(53,67)	45(37,53)	41(36,49)	41(38,45)	59	59	62	61	59	62
DM	60(58,62)	59(58,62)	63(61,65)	71(57,67)	61(57,69)	65(57,71)	47(39,54)	43(40,45)	40(27,45)	59	59	65	63	59	64
PCI															
IG	63(61,65)	67(64,75)	67(65,70)	64(53,72)	72(67,78)	58(56,64)	36(31,42)	39(28,50)	22(22,31)	63	65	71	63	69	65
GI	61(61,64)	67(65,68)	67(63,70)	67(50,86)	69(50,75)	67(56,69)	39(28,64)	42(14,50)	31(22,42)	63	64	69	65	64	64
X2	63(60,64)	65(63,72)	65(63,65)	69(56,69)	72(58,78)	72(58,78)	36(33,44)	36(33,42)	42(28,53)	61	64	63	65	65	68
GR	63(61,70)	64(64,65)	65(64,67)	67(56,82)	67(53,83)	72(53,72)	44(31,50)	39(25,56)	39(22,44)	65	63	65	63	65	67
DM	64(63,65)	65(61,71)	65(64,68)	69(64,78)	72(67,78)	69(64,75)	42(33,47)	42(36,56)	39(33,47)	63	62	64	66	67	67
CABG															
IG	69(67,73)	66(63,69)	70(70,71)	70(63,77)	74(65,79)	65(63,65)	35(23,40)	42(33,47)	23(11,26)	67	67	73	70	68	68
GI	69(69,71)	63(61,65)	69(67,71)	67(58,74)	67(56,72)	74(72,74)	28(21,35)	42(30,42)	37(33,40)	70	63	67	68	64	70
X2	69(67,73)	63(61,65)	69(67,72)	72(63,81)	72(63,79)	74(72,77)	33(21,44)	47(42,58)	37(30,42)	67	61	67	69	66	71
GR	69(66,71)	63(61,66)	69(69,75)	67(65,74)	70(61,74)	74(65,77)	35(26,37)	44(28,49)	30(26,40)	67	62	69	68	65	71
DM	71(70,72)	61(59,67)	69(69,71)	67(63,72)	77(58,81)	70(58,74)	28(19,30)	49(40,58)	33(21,35)	73	59	70	71	67	70

در این مطالعه محققین اظهار داشته اند که انتظار می رود که داده کاوی می تواند در شناسایی گروه پرخطر و کم خطر از افراد، یک عامل تعیین کننده در انتخاب درمان باشد، به عنوان مثال، طبی یا جراحی. .با این حال، تحقیقات بیشتر با مجموعه داده های بزرگ هنوز هم مورد نیاز است.

Ordoñez [۱۴,۱۵]، درختان تصمیم و را برای پیش بینی CAD بر اساس عوامل خطر جنسیت، سیگار کشیدن، کلسترول، و سن بررسی کرد.

همچنین تسین و همکاران[۳۸]، از درختان طبقه بندی برای ساخت سه مدل مختلف برای سکته قلبی (MI) استفاده کردند.

علاوه بر این، Polat و همکاران[۳۹]، مدل های مبتنی بر درخت تصمیم را برای طبقه بندی CAD، دستیابی به طبقه بندی صحیح با ۸۲٪ را توسعه دادند.

علاوه بر این، Pavlopoulos و همکاران[۴۰]، با استفاده از درخت های تصمیم گیری الگوریتم C4.5 به تجزیه و تحلیل ویژگی های صدای قلب های مختلف، که به پزشکان برای تشخیص بهتر CAD کمک می کند، پرداختند.

Shouman و Turner[۴۱]، از درخت تصمیم C4.5 برای تشخیص CAD استفاده کردند و با استفاده از هرس خطای کاهشی [1]، در نتیجه به دقت ۸۴,۱٪ رسیدند.

۳-۳-۱-۲ الگوریتم J4.8

در تحقیق انجام گرفته توسط آقای شومان و همکارانشان[۴۲]، برای جستجوی بهتر درختهای تصمیم مختلف، تکنیکهای متفاوتی پیشنهاد گردید. در این پژوهش مدلی ارائه می شود که درخت تصمیم J4.8 و الگوریتم بگینگ[2] را در تشخیص بیماران قلبی به کار میگیرد .در این تحقیق سه نوع درخت تصمیم مورد آزمایش قرار گرفته است .بهره اطلاعات، ضریب جینی، نسبت بهره و نهایتا هرس خطای کاهش یافته، روی تمام قوانین استخراج شده توسط درخت تصمیم اعمال می گردد. مجموعه داده مورد استفاده در این تحقیق، کلیولند[3] می باشد. بیشترین دقتی که به آن دست یافته شد برابر ۸۴,۱٪ و توسط درخت تصمیم نسبت بهره می باشد.

۳-۳-۲ شبکه های عصبی مصنوعی(ANN)

یک شبکه عصبی مصنوعی[۴۳]، روشی برای پردازش اطلاعات است که از سیستم های عصبی زیستی الهام گرفته شده و مانند مغز پردازش اطلاعات انجام می گیرد. عنصرکلیدی این ایده، ساختار جدید سیستم پردازش اطلاعات است. این سیستم از شمار زیادی عناصر پردازشی فوق العاده به هم پیوسته تشکیل شده که برای حل یک مسأله با هم هماهنگ عمل می کند. شبکه های عصبی، نظیر انسانها، با مثال یاد می گیرند. یک شبکه عصبی مصنوعی برای انجام وظیفه ای مشخص، مانند شناسایی الگوها و دسته بندی اطلاعات، در طول یک پروسه یادگیری، تنظیم می شود. در سیستمهای زیستی یادگیری با تنظیماتی دراتصالات سیناپسی که بین اعصاب قرار دارد همراه است.این روش شبکه عصبی مصنوعی هم می باشد. این شبکه ها قادر به مدل سازی توابع غیر خطی است. شبکه های عصبی مصنوعی تکنیکهای تحلیلی هستند که قادر به پیش بینی مشاهدات جدید(متغیرهای یکسان و یا سایر) پس از اجرای یک فرایند یادگیری به اصطلاح ازداده های موجود هستند[۴۳].

درتحقیق در سال ۱۹۷۰ دریافتند که رابطه اساسی بین ریزدانه های درون خون و بعضی بیماریها نظیر گرفتگی عروق کرونری وجود دارد و زمانی که میزان ریزدانه های درون خون تغییر می کند ممکن است بیماری رخ دهد[۴۴]. بر این اساس در[۴۵]، سیستمی توسعه داده شد که بیماری گرفتگی عروق کرونری را

[1] Reduce Error Pruning
[2] Bagging
[3] Clevlend

با استفاده از یک شبکه عصبی و با توجه به مقدار املاح درون خون(میزان Sr, Cu, Mg, Zn درون خون) تشخیص می دهد.

در[۴۶]، Chen Tian-hua و همکاران، سیگنالهای صوتی قلب از طریق یک سیستم کامپیوتری که سیگنال های صوتی قلب را جمع آوری و مرتب سازی می کند، جمع آوری شده و با استفاده از متدهای دیجیتال نظیر تبدیل موج باکامپیوتر پردازش می شوند، پارامترهای مشخصه محاسبه شده و در نهایت از یک شبکه عصبی با الگوریتم یادگیری پس انتشار[1] سه لایه برای تشخیص انواع سیگنالهای صوتی استفاده می شود.

بسیاری از مطالعات براساس ترکیبی از داده های استخراج شده از تستها و آزمایشات مختلف، سعی در تشخیص گرفتگی شریانهای کرونری نموده اند. به عنوان مثال می توان به استفاده از مجموعه داده کلیولند نام برد. این مجموعه داده شامل مواردی از قبیل :سن، جنس، میزان درد قفسه سینه، فشارخون، میزان کلسترول خون، اطلاعات مربوط به قطعه(ST قسمتی از نوار قلب)و ... می باشد.در تحقیقی که در سال ۱۹۹۵ توسط آقای کافمن و همکارانشان حاصل شد[۴۷]، گرفتگی عروق کرونر به وسیله یک شبکه عصبی از نوع پس انتشار تشخیص داده شد .این شبکه روی ۲۳ متغیر غیرتهاجمی نظیر داده های تست ورزش آموزش یافت .دقت این روش حدود ۸۶ درصد می باشد.

در[۴۸,۴۹]، سیگنالهای داپلر را جهت تشخیص بیماری گرفتگی عروق کرونری انتخاب نموده و مورد استفاده قرار دادند. در[۴۸]، از PCA برای انتخاب ویژگی استفاده شده، سپس با استفاده از یک شبکه عصبی و با استفاده از الگوریتم پس انتشار خطا دسته بندی صورت گرفته است .در این تحقیق به حساسیت ۹۷,۷٪ و اختصاصی بودن ۹۸,۱٪ رسیدند.

۳-۳-۲-۱ شبکه عصبی مصنوعی MLP [2]

پرسپترون چند لایه (MLP) مدل شبکه عصبی مصنوعی پیشخور است که مجموعه ای از داده های ورودی را به مجموعه ای از خروجی های مناسب نگاشت می کند. MLP از چندین لایه ازگره در یک گراف جهت دارتشکیل شده است که هر لایه به طور کامل به لایه دیگر متصل می شوند، به جز گره های ورودی، هر گره یک نورون[3] یا جزء پردازشی با یک تابع فعالیت غیر خطی است. MLP با بهره گیری از تکنیک یادگیری

[1] Back Propagtion
[2] MultiLayer Perceptron
[3] Nureon

نظارت شده[1] به نام پس انتشاربرای آموزش است. شبکه MLP یک پرسپترون خطی استاندارد اصلاح شده است و می تواند داده هایی را که به صورت خطی از هم جدا نشده اند تشخیص دهند.

در اولین مطالعه[۵۰]، Oleg Yu و همکاران، یک مدل تشخیصی برای بیماری عروق کرونر قلب (CAD) مبتنی بر شبکه های عصبی مصنوعی (ANN) با استفاده از مجموعه ای از عوامل ژنتیکی و پارامترهای بالینی در این بیماری ارائه شد.

امروز، شبکه های عصبی مصنوعی (ANN) در تحقیقات بالینی و ژنتیکی استفاده می شود. تلاش هایی برای ایجاد مدل های تشخیصی برای بیماری های مختلف با استفاده از شبکه های عصبی مصنوعی از توپولوژی های مختلف وجود دارد.فرض بر این است که استفاده از مجموعه ای از نشانه های CAD اجازه می دهد شبکه های عصبی مصنوعی نه تنها برای تشخیص، بلکه برای پیش بینی حوادث از نظر بالینی، سکته قلبی استفاده شود.

پایگاه داده اصلی برای شبکه های عصبی مصنوعی (ANN) ویژگی های مربوط به ۴۸۷ نفر شامل علائم بالینی، آزمایشگاهی، کاربردی، آنژیوگرافی عروق کرونر و ژنتیک [پلی مورفیسم تک نوکلئوتیدی (SNP ها)] (۳۲۷ نفر مبتلا به CAD ناشی از آترواسکلروز[2] (گرفتگی رگ ها) عروق کرونر، ۱۶۰ نفر بدون ابتلا به CAD) جمع آوری شد. پیشرفت های اخیر تمرکز بر نشانگرهای ژنتیکی برای پیش بینی CAD و توصیه پلی مورفیسم تک نوکلئوتیدی (SNP ها) برای ارزیابی ریسک است.

عوامل خطرزایCAD شامل: سن، جنسیت،کلسترول تام، کلسترول لیپوپروتئین با چگالی بالا (HDL)، کلسترول لیپوپروتئین با چگالی کم (LDL)، کلسترول لیپوپروتئین با چگالی بسیار کم (VLDL)، تریگلیسیرید و نرخ کلسترول، پلاسما در حالت قند ناشتا (قند خون)، فشار خون شریانی، دیابت، وضعیت فعلی دخانیات، چاقی و یک سابقه خانوادگی از CAD تجزیه و تحلیل شد.

مدل شبکه عصبی با استفاده از شبکه پرسپترون [3] چند لایه، شبکه عصبی پیشخور [4] چند سطحی توسط انتشار خطای آماری آموزش داده شده ایجاد شده است.مجموعه ای از پارامترهای متغیر به منظور تنظیم مدل ANN با ارتباط دو به دو بین پارامترهای پایگاه داده ها و تشخیص CAD انتخاب شدند.

[1] Supervised Learning
[2] Atherosclerosis
[3] perceptron
[4] Feed forward

دقت مدل توسط الگوریتم ژنتیک با پارامترهای بهینه سازی های مختلف از جمله تعداد نورون در لایه پنهان، تعداد ورودی ها به شبکه های عصبی و ضریب شیب توابع فعال سازی [1] بهبود یافته است.

ANN با خواندن عوامل متغیر مشخص شده ایجاد شده است. و دو کلاس طبقه بندی شد:" ۱" – CAD

" ۰" - سالم. در مجموع از ۲۸۷ نمونه برای آموزش، ۱۰۰ تا برای اعتبار متقاطع و ۱۰۰ تا برای تست مورد استفاده قرار گرفت.

.به عنوان نتیجه با تجزیه و تحلیل همبستگی بین عوامل متغیر مشخص و CAD، ۳۲ عامل خطر مرتبط با این بیماری را انتخاب کردیم که این عوامل توسط شبکه های عصبی مصنوعی استفاده و آنالیز شد. روش با استفاده از شبکه های عصبی مصنوعی مختلف و تعداد متغیری از عوامل ورودی (از ۵ تا ۱۰) انجام شدو به مدل هایی با ۶۴-۹۴٪ دقت تشخیص رسید.

بهترین نتیجه (۹۴٪) در یک مدل پرسپترون چند لایه (MLP) با دو لایه پنهان و ۱۰ عامل (حرفه ، LDL، HDL، تری گلیسرید، میزان کلسترول، شاخص SCORE، شکاف تخلیه ای بطن چپ[2] ، سابقه CAD خانواده، داده آرتریوگرافی کرونری[3] (داده گرافی شریانی کرونر)، ژن PAI) به دست آمد. در واقع هر دو عوامل خطرزای CHD ژنتیکی و غیر ژنتیکی به دست آمد. بر اساس نتایج حاصل از این پژوهش ابزاری بالینی، CAD ناشی از تصلب شریان عروق کرونر در ۳۲۷ نفر (۶۷,۲٪) تشخیص داده شد. در مجموع ۱۶۰ نفر (۳۲,۸٪) دیوار سالم عروق کرونر داشتند و شواهدی از بیماری CAD موجود نبود.

از سوی دیگر، در همان نوع ANN با ۵ عامل (داده آرتریوگرافی کرونری، میزان کلسترول، شاخص SCORE، شکاف تخلیه ای بطن چپ، ژن PAI) دقت تشخیص کمتر (۷۸٪) بود. با این حال، همان ۱۰ عامل با انواع دیگر ANN تجزیه و تحلیل شد و نتیجه ۷۹٪ به دست آمد. این نشان می دهد که دقت تشخیصی بستگی به نوع ANN و تعداد عوامل متغیر دارد.

گام بعدی شامل تجزیه و تحلیل ۱۰ مدل پیش بینی CAD تشکیل شده از ترکیبات مختلف از بلوک های آزمون (مشخصات جمعیتی، سابقه CAD، تست های آزمایشگاهی ، اکوکاردیوگرافی و داده آنژیوگرافی عروق کرونر ،ژن) است که توسط MLP با دو لایه ۴ عصبی پنهان [4] ایجاد شده است.

[1] slope coefficient of activation functions
[2] left ventricular ejection fraction
[3] coronary arteriography data
[4] buried

تعدادی از عوامل ورودی متغیر در بازه زمانی از ۸ (مدل های I و V) به ۱۵ (مدل IV) رسید. نتایج در جدول زیر نشان داده شده است:

جدول ۳-۲: دقت مدل های ANN برای تشخیص CAD

Model	Factors	Accuracy (%)
I	Age, profession, diabetes, arterial hypertension, smoking, obesity, family anamnesis of CHD, glucose, cholesterol	64
II	Age, profession, diabetes, arterial hypertension, smoking, obesity, family anamnesis of CHD, glucose, cholesterol	77
III	Age, profession, diabetes, arterial hypertension, smoking, obesity, family anamnesis of CHD, glucose, total cholesterol, HDL, LDL, VLDL, triglycerides, cholesterol ratio	83
IV	Age, profession, diabetes, arterial hypertension, smoking, obesity, heredity, glucose, total cholesterol, HDL, LDL, VLDL, triglycerides, cholesterol ratio, coronary angiography data	91
V	*NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1, CRP-2, CRP-3* SNPs	90
VI	*NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1, CRP-2, CRP-3* SNPs, SCORE index	83
VII	*NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1, CRP-2, CRP-3* SNPs, coronary angiography data	89
VIII	*NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1, CRP-2, CRP-3* SNPs, SCORE index, coronary angiography data	93
IX	*NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1, CRP-2, CRP-3* SNPs, HDL, LDL, glucose	90
X	*NOS, ACE, AGT-235, AGT-174, AGTR, CRP-1, CRP-2, CRP-3* SNPs, age, smoking, obesity, family anamnesis of CHD, HDL, LDL	88

دقت مینیمم ۶۴ ٪ در مدل I به دست آمد، که شامل ۸ عامل غیر ژنتیکی بود. مدل IV، V، VIII، IX و ، که مجموعه های مختلف مشخصی از متغیرها را دارد ، دقت تشخیصی بیش از ۹۰٪ را نشان داد. مدل IV تنها شامل عوامل غیر ژنتیکی بود ، مدل V تنها هشت SNP بود (NOS، ACE، AGT -235، AGT -174، AGTR، CRP - 1، CRP - 2 و CRP - 3) ، مدل VIII شامل همان SNP همراه با داده های آنژیوگرافی عروق کرونر و شاخص SCORE بود، مدل IX شامل همان SNP و HDL، LDL و گلوکز می باشد. در نتیجه دقت مطلوب در اغلب موارد وقتی که یک مدل تشخیصی شامل عوامل ژنتیکی وجود داشت به دست آمد.

بنابراین یکی از روش های مدرن برای حل مشکلات طبقه بندی ، پردازش داده های هوشمند مبتنی بر حل و فصل کارهای بهینه سازی با استفاده از شبکه های عصبی مصنوعی یعنی ANN است.نتایج ما نشان می دهد که شبکه های عصبی مصنوعی ANN، نیز ممکن است برای ایجاد یک مدل بسیار دقیق و موثر برای پیش بینی CAD مورد استفاده قرار گیرد.

دقت تشخیصی را می توان نه تنها با افزایش تعداد نشانگر ژنتیکی، بلکه با انتخاب دقیق آنها بهبود داد. ممکن است کاندیداهای خوب برای در نظر گرفتن ژن های دخیل در رشد سلول عروق، مرگ سلولی و التهاب، و یا دیگر عوامل مرتبط با CAD، به عنوان مثال ژن ABCA1، CYP1A2 (سیتوکروم P450) باشد.

در تحقیق دیگر[۵۱]، نیز از داده های به دست آمده از تست ورزش، برای تشخیص گرفتگی عروق کرونر، استفاده کردند. در این تحقیق، برای دسته بندی، شبکه عصبی مصنوعی پرسپترون چند لایه با الگوریتم یادگیری پس انتشار به کار گرفته شده است .نکته قابل توجه در این روش، جداسازی کلاس ها برای گرفتگی عروق مختلف یعنی(رگ کرونری اصلی چپ، رگ نزول کننده چپ، رگ چرخشی چپ و رگ کرونری راست) می باشد؛ برخلاف بیشتر روش ها که تنها دوکلاس نرمال و غیرنرمال را پشتیبانی می کنند. در این تحقیق، دقت روش برای هرکدام از کلاس ها به همراه پارامترهای حساسیت و اختصاصی بودن گزارش شده است. طبق این گزارش ، بیشترین دقت مربوط به تشخیص انسداد رگ کرونری اصلی چپ و برابر ۹۱,۲۱٪ و کمترین دقت مربوط به تشخیص انسداد رگ چرخشی چپ و برابر ۶۴,۸۵٪ می باشد.

در تحقیق دیگر دسته بندی توسط یک شبکه عصبی پرسپترون با الگوریتم پس انتشار، به کمک انتخاب ویژگی در[۵۲]، انجام شده است. مجموعه داده مورد استفاده در این تحقیق نیز کلیولند است .در روش پیشنهادی] ویژگیهای مجموعه دادهی مذکور، با استفاده از بهره اطلاعات، از ۱۳ ویژگی به ۸ ویژگی کاهش داده شده است .این کاهش ویژگی باعث شد، روی مجموعه آموزشی، ۱,۱٪ و روی مجموعه تست، ۰,۸۲٪ افزایش دقت حاصل شود.

کار دیگر توسط یان و همکاران، با استفاده از پرسپترون چند لایه برای ساخت سیستم پشتیبانی تصمیم گیری برای تشخیص پنج بیماری های قلبی اصلی [۵۳] پیشنهاد شد.

۳-۳-۲-۲ شبکه عصبی مصنوعی Feed Forward

شبکه های عصبی پیشخور نوع اول و مسلما ساده ترین شبکه عصبی مصنوعی ابداع شده می باشد. در این شبکه اطلاعات تنها در یک جهت یعنی رو به جلو حرکت می کند، داده های گره های ورودی از طریق گره های مخفی به سمت گره خروجی حرکت می کند. هیچ دوره و یا حلقه ای در شبکه وجود ندارد.

در[۵۴]، تحقیقی صورت گرفته که شامل بررسی سه شبکه عصبی می شود. در این تحقیق، یادگیری بر اساس داده های نوار قلب و سایر اندازه گیری ها در طول تست ورزش می باشد. دو شبکه اول از نوع بازگشتی و شبکه

سوم از نوع پیشخور می باشد. میانگین حساسیت برای شبکه های بیان شده برابر ۸۶,۵ % و ۸۵% و ۸۳,۷۵ %و میانگین اختصاصی بودن برابر ۶۴,۵ % و ۷۸,۵ % و ۷۶,۷۵ % می باشد.

آقای داس و همکاران[۱]، نیز از مجموعه داده معروف کلیولند استفاده نموده و یک روش تجمیعی[1] را پیشنهاد دادند که در آن از سه شبکه عصبی مجزا استفاده میشد .این شبکه های عصبی از نوع پیش خور چند لایه بوده و با الگوریتم پس انتشار آموزش داده شدند .دقت این روش ۸۹,۰۱ درصد، حساسیت۸۰,۹۵ درصد و اختصاصی بودن، ۹۵,۹۱ درصد می باشد.

۳-۳-۲-۳ شبکه عصبی مصنوعی LVQ

چندی سازی برداری یادگیر نیز می تواند به عنوان یک ساختار شبکه عصبی تفسیر شود و توسط Teuvo Kohonen پیشنهاد شده بود، در اصل در LVQ[2]، نمایندگان بارز پارامترهاسازی کلاس ها، همراه با اندازه گیری فاصله مناسب، طرح مبتنی بر فاصله را طبقه بندی می کنند.

در پژوهش آقای مکرینی[3] و همکارانشان[۵۵]، استفاده از یک شبکه عصبی LVQ یا چندی سازی برداری یادگیر پیشنهاد گردید .در این تحقیق در الگوریتم نزدیکترین همسایه مجاور استفاده شده در LVQ معیار فاصله اقلیدسی با یک معیار بر اساس وزن دهی جایگزین می گردد که این وزن دهی با توجه به اهمیت هر متغیر در تصمیمگیری میباشد .در این تحقیق از مجموعه داده مربوط به بیماری گرفتگی عروق کرونری که از جمله مجموعه دادههای موجود در UCI [۵۶]، می باشد، جهت اعتبارسنجی روش استفاده شده و به دقت ۸۰ درصد رسیدند.

۴-۳-۲-۳ شبکه عصبی مصنوعی Neuru-Fuzzy

شبکه عصبی فازی یک سیستم استنتاج فازی در بدنه شبکه عصبی مصنوعی است .بسته به نوع FIS، چندین لایه که شبیه سازی فرآیند استنتاج فازی مثل فازی سازی، استنتاج، تجمیع و غیر فازی سازی وجود دارد . تعبیه FIS در یک ساختار کلی در یک ANN این مزیت را دارد که با استفاده از روش های آموزشی ANN در دسترس، پارامترهای سیستم فازی را پیدا کنیم.

[1] Ensemble Method
[2] Learning Vector Quantization
[3] Macrini

در تحقیقی[۵۷]، سیستم های پشتیبانی تصمیم گیری فازی براساس قانون (DSS) برای تشخیص بیماری عروق کرونر (CAD) ارائه شده اند.

از آنجایی که روشهای تهاجمی در تشخیص بیماری CAD مانند اکوکاردیوگرافی دارای هزینه سنگین می باشد از این رو قادر به ارائه یک روش غیر تهاجمی در پیش بینی وقوع CAD با استفاده از ویژگی هایی که آسان به دست آمده و تصمیم گیری بر اساس آن از ارزش زیادی برخوردار است، صورت می گیرد.

سیستم بطور خودکار تولید شده از یک مجموعه داده اولیه بااستفاده از متدلوزی ۴ مرحله ای:

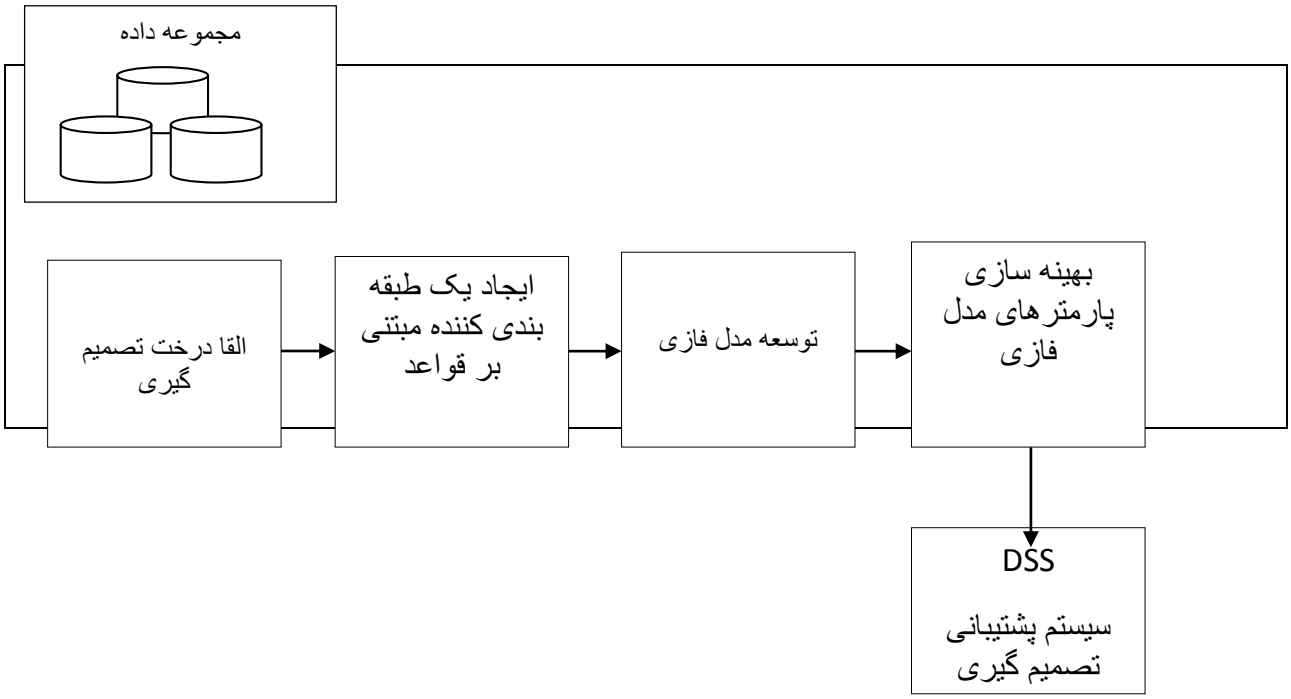
۱) استنتاج یک درخت تصمیم از داده ها

۲) استخراج مجموعه ای از قوانین از درخت تصمیم در شکل نرمال و ساخت یک مدل قطعی [1]

۳) انتقال مجموعه ای از قوانین قطعی به یک مدل فازی.

۴) بهینه سازی پارامترهای مدل فازی

به شکل زیر نگاه کنید:



شکل ۳-۲: مراحل ساخت یک DSS [۵۷]

[1] Crisp

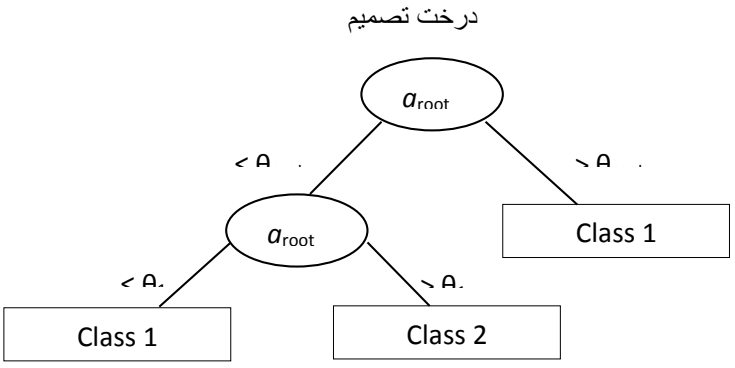
با استفاده از مجموعه داده و ارزیابی ۱۹۹ نفر، توسط ۱۹ ویژگی، از جمله داده های تاریخی همانند سابقه خانوادگی فشارخون، چربی خون، دیابت، سیگار کشیدن و داده های جمعیتی مانند سن و جنسیت و همچنین آزماشهای آزمایشگاه و شاخص سفتی شریانی مانند فشار آئورت و CA برای تمام افراد انجام شد و وضعیت بالینی آنها (وجود یا عدم وجود CAD) توسط دو آنژیوگراف با تجربه تعیین شد، داده های بالینی در بخش قلب و عروق تهاجمی از بیمارستان دانشگاه لوآنینا[1] جمع آوری شد، فرد با دارا بودن حداقل ۵۰٪ تنگی با قطر بیشتر در حداقل یک رگ به عنوان بیمار مبتلا به عروق کرونر تعریف شد. در مجموع از ۱۹۹ نفر، ۸۹ نفر نرمال بودند و ۱۱۰ نفر از افراد دیگر مبتلا به CAD بودند در نهایت سیستم مورد نظر تولید شد.

ساخت درخت تصمیم با استفاده از الگوریتم C4.5 اجرا شده است که یک درخت تصمیم گیری از داده های آموزشی را تولید می کند. پس از القای درخت تصمیم، یک روش هرس بنام postpruning (جایگزین کردن یک زیر درخت با یک گره برگ جدید) را به منظور کاهش حجم و پیچیدگی درخت اعمال می کنیم.

مرحله بعد ساخت یک مدل طبقه بندی مبتنی بر قواعد است، یک طبقه بندی کننده ی مبتنی بر قانون یک تکنیک برای طبقه بندی رکوردها با استفاده از مجموعه ای از قوانین اگر-آنگاه است. قوانین برای مدل، قطعی و به شکل مجموعه ای از قوانین است که در یک DNF، (r1 U r2 U … rk) نشان داده شده است که r1 قوانین طبقه بندی هستند. هر قانون طبقه بندی به صورت زیر بیان می شود:

$$ri: (Condi) \rightarrow y \qquad (۲-۳)$$

که در آن y کلاس پیش بینی شده و سمت چپ پیش شرط است. شکل زیر استخراج مجموعه ای از قوانین قطعی از درخت تصمیم است:



شکل۳-۳: ساخت درخت تصمیم [۵۷]

[1] Leanina

If ($a_{root1} \leq \theta_{root}$ AND $a_1 \leq \theta_1$)

Then class 1

If ($a_{root1} \leq \theta_{root}$ AND $a_1 > \theta_1$)

Then class 2

If ($a_{root1} > \theta_{root}$) Then class 1

در مرحله بعد مدل فازی ساخته شد، مجموعه ای قوانین قطعی با استفاده از یک تابع عضویت فازی به یک مدل فازی تبدیل می شود.

$$g_s^{\text{inc}}(a, \theta_1, \theta_2) = \frac{1}{1 + e^{\theta_1(\theta_2 - a)}} \text{ (increcsing)} \qquad (۳-۳)$$

سپس در مرحله آخر پارامترهای مدل فازی را با الگوریتم تصادفی بر پایه ارتباط چند سطحی MLSL[1] بهینه سازی می کنیم.

ده برابر[2] اعتبار سنجی بکار گرفته شد و میانگین حساسیت و ویژگی با استفاده از مجموعه قوانین استخراج شده از درخت تصمیم (مرحله ۱ و ۲) به ترتیب ۶۲٪ و ۵۴٪ به دست آمد، در حالی که میانگین حساسیت و ویژگی زمانی که مراحل و بهینه سازی فازی استفاده شد به ۸۰٪ و ۶۵٪ افزایش پیدا کرد. سیستم چندین مزیت را زمانیکه بطور خودکار تولید شده ارائه می کند، که تشخیص CAD را به راحتی انجام می دهد و بدون استفاده از روشهای تهاجمی، ویژگی ها را به دست می آورد و قادر به ارائه (مهیا کردن) تفسیر برای تصمیم گیری ساخته شده است.

بنابراین در مقایسه این روش پیشنهادی از ۲ روش طبقه بندی استفاده کردیم:

۱. شبکه های عصبی مصنوعی Feed Forward
۲. سیستم استنتاج فازی-عصبی تطبیقی (ANFIS)

[1] Multilevel Single Linkage
[2] Ten Fold

به این ترتیب در روش پیشنهادی، عملکرد قابل مقایسه با شبکه های عضبی مصنوعی به ترتیب ۷۳/۴٪ در مقابل ۷۳/۹٪ بود و نتایج قابل توجه بهتری از ANFIS گزارش شده است.

جدول۳-۳: نتایج مقایسه روش های فازی

		Folds										Overall
		1	2	3	4	5	6	7	8	9	10	
Crisp rule-based classifier	TP	10	6	8	5	7	7	3	10	5	7	68
	TN	6	5	4	7	4	5	4	3	5	5	48
	FP	3	4	5	2	5	4	5	6	4	3	41
	FN	1	5	3	6	4	4	8	1	6	4	42
	Se (%)	90.9	54.5	72.7	45.5	63.6	63.6	27.3	90.9	45.5	63.6	61.8
	Sp (%)	66.7	55.6	44.4	77.8	44.4	55.6	44.4	33.3	55.6	62.5	53.9
	Acc (%)	80.0	55.0	60.0	60.0	55.0	60.0	35.0	65.0	50.0	63.2	58.3
Proposed DSS ***Optimized Fuzzy model***	TP	9	8	11	6	10	9	9	10	8	8	88
	TN	8	5	5	8	7	6	5	4	5	5	58
	FP	1	4	4	1	2	3	4	5	4	3	31
	FN	2	3	0	5	1	2	2	1	3	3	22
	Se (%)	81.8	72.7	100	54.5	90.9	81.8	81.8	90.9	72.7	72.7	80.0
	Sp (%)	88.9	55.6	55.6	88.9	77.8	66.7	55.6	44.4	55.6	62.5	65.2
	Acc (%)	85.0	65.0	80.0	70.0	85.0	75.0	70.0	70.0	65.0	68.4	73.4
Adaptive Neuro-fuzzy Inference System	TP	6	6	7	6	8	5	6	8	6	6	64
	TN	5	7	4	6	5	4	5	2	4	7	49
	FP	4	2	5	3	4	5	4	7	5	1	40
	FN	5	5	4	5	3	6	5	3	5	5	46
	Se (%)	54.5	54.5	63.6	54.5	72.7	45.5	54.5	72.7	54.5	54.5	58.2
	Sp (%)	55.6	77.8	44.4	66.7	55.6	44.4	55.6	22.2	44.4	87.5	55.1
	Acc (%)	55.0	65.0	55.0	60.0	65.0	45.0	55.0	50.0	50.0	68.4	56.8
Artificial Neural Network	TP	9	11	7	8	9	8	8	11	10	7	88
	TN	7	4	7	7	6	6	6	4	5	7	59
	FP	2	5	2	2	3	3	3	5	4	1	30
	FN	2	0	4	3	2	3	3	0	1	4	22
	Se (%)	81.8	100	63.6	72.7	81.8	72.7	72.7	100	90.9	63.6	80.0
	Sp (%)	77.8	44.4	77.8	77.8	66.7	66.7	66.7	44.4	55.6	87.5	66.3
	Acc (%)	80.0	75.0	70.0	75.0	75.0	70.0	70.0	75.0	75.0	73.7	73.9

در تحقیق دیگر، آقای پلات سیستمی را برای تشخیص بیماری قلبی ارائه نمودند[۵۸]، که در آن دسته بندی از طریق سیستم بازشناسی ایمنی مصنوعی با مکانیسم تخصیص مقدار فازی[1]، تشخیص داده می شد. در این تحقیق از یک روش ترکیبی استفاده شد که در آن فرآیند وزندهی جدید بر اساس نزدیکترین همسایه مجاور و در مرحله پیش پردازش روی مجموعه داده، صورت میپذیرد .مجموعه داده استفاده شده در این روش کلیولند می باشد .دقت این روش ۸۷ % و حساسیت و اختصاصی بودن، به ترتیب ۹۲,۳۰ % و ۷۸,۵۷٪ گزارش شد.

[1] Fuzzy-AIRS

در تحقیق دیگر[۵۹]، سیستم پشتیبان تصمیم گیری بالینی، برای تشخیص گرفتگی شریانهای کرونر مطرح شده است،که از دو فاز تشکیل می گردد. فاز اول شامل روشی خودکار برای تولید قوانین فازی است و فاز دوم توسعه یک سیستم پشتیبان تصمیم گیری می باشد که مبتنی بر قوانین فازی است. در فاز اول از روشهای انتخاب ویژگی و وزندهی ویژگی برای به دست آوردن قوانین فازی وزن دهی شده استفاده می شود و در فازدوم سیستم فازی بر اساس قوانین فازی وزن دهی شده و ویژگیهای انتخاب شده ی آن ساخته می شود. دراین تحقیق از مجموعه داده های کلیولند، هانگرین و سوئیتزرلند استفاده شده است. برای هر یک ازمجموعه های مذکور و بر اساس هر یک از کلاسهای بیمار و سالم، دقت، حساسیت و اختصاصی بودن، برای هر دو مجموعه ی آموزشی و تست گزارش شده است.

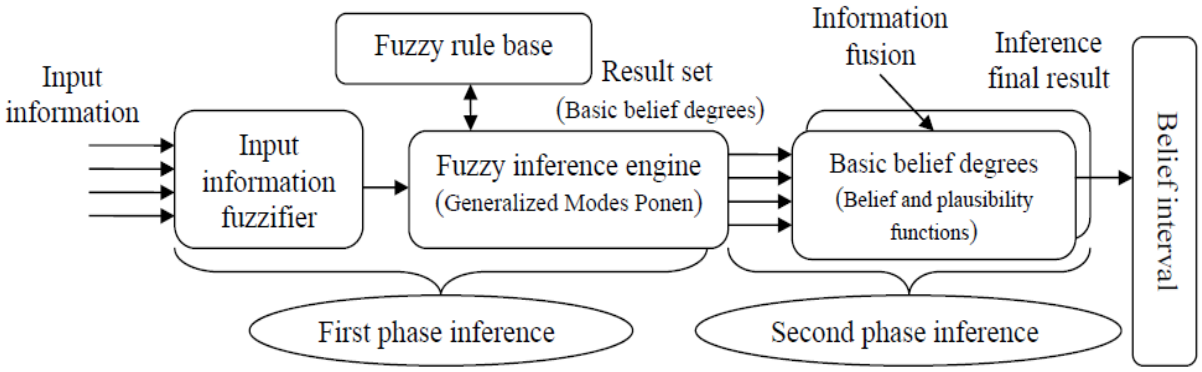
در تحقیق بعدی خطیبی و همکاران[۶۰]، یک موتور استنتاج ترکیبی فازی- شهودی [۱] برای ارزیابی خطر ابتلا به بیماری عروق کرونر قلب در نظر گرفته شد.

در بسیاری از مسائل مهندسی ، ما در رویارویی با اطلاعات و عدم اطمینان در تصمیم گیری با ابهام روبه رو هستیم، به طوری که این پدیده باعث می شود نتوانیم به نتایج خاص برای راه حل پیشنهاد شده برسیم (مشکل است). بنابراین در تحقیق انجام شده، یک موتور استنتاج به نام موتور استنتاج ترکیبی [۲] فازی - شهودی با استفاده از تئوری Dempster-Shafer ، نظریه شواهد وتئوری مجموعه های فازی ارائه شده است . این موتور ترکیبی (پیوندی) در دو مرحله عمل می کند .

در مرحله اول ، اطلاعات ورودی مبهم از طریق مجموعه های فازی مدل می شوند ، مجموعه قانون فازی را برای این مشکل استخراج می کنیم، و قوانین استنتاج فازی را بر روی مجموعه های فازی به دست آورده برای تولید نتایج مرحله اول اعمال می کنیم. در مرحله دوم، نتایج به دست آمده از مرحله قبل را به عنوان باورهای اساسی برای این مسئله در نظر گرفتند، این باور و توابع معقول یک مجموعه هستند.با جمع آوری اطلاعات از منابع مختلف، باورهای اساسی متنوع که باید برای تولید نتیجه یکپارچه ترکیب شوند را آماده می کنند . برای این منظور، قوانین ترکیب شواهد برای انجام ترکیب (همجوشی) [۳] اطلاعات انجام می شود و موتور پیشنهاد شده در ارزیابی خطر ابتلا به بیماری کرونری قلب (CAD) مورد ارزیابی قرار می گیرد، که میزان دقت ۹۱,۵۸ درصد برای پیش بینی درست را به همراه دارد. این مدل موتور ترکیبی ، اطلاعات مبهم و تصمیم گیری های غیر قطعی را از طریق ترکیب اطلاعات دقیق می کند و نتایج دقیق تری را فراهم می کند

[1] Evidential
[2] Hybrid
[3] Fusion

، به طوری که می توان آن را به عنوان سیستم پشتیبانی تصمیم گیری هوشمند در مسائل مهندسی متنوع در نظر گرفت.



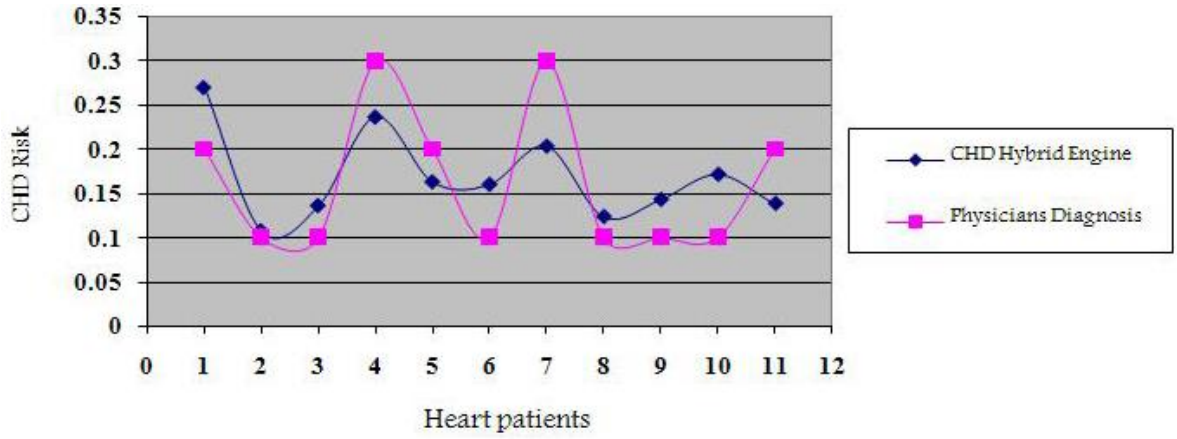
شکل ۳-۴: دو مرحله موتور استنتاج ترکیبی فازی- شهودی [۶۰]

ابهام نشانگر پزشکی و تردید در تعیین خطر CAD می تواند از طریق مجموعه فازی و تئوری شواهد مدل شود. ابتدا، ابهام و عدم قطعیت اطلاعات نشانگرها پزشکی در مجموعه های فازی شان داده شده است و پس از آن، با استفاده از دانش پزشکی مناسب، موتور استنتاج ترکیبی فازی-شواهدی برای ارزیابی خطر CAD استفاده می شود.

برای نشان دادن روابط بین نشانگرها و خطر ابتلا به CAD،از مطالعه قلب فرامینگهام استفاده شده است، بنابراین به عنوان یک پایگاه قانون فازی، متشکل از ۴۷ قانون ، مدل عدم قطعیت در ارزیابی خطر CAD، استخراج شده ، در حال حاضر، دستیابی به یک فازی ساز و قانون فازی اساسی، ممکن است برای تکمیل اولین فاز استنتاج از میان موتور استنتاج فازی پیشنهاد شده استفاده شود. در این موتور، ما برای تعیین درجه هر یک از نشانگرهای پزشکی در بروز CAD در یک بیمار از تعمیم حالت Ponen استفاده می کنیم.

جدول ۳-۴: مقایسه نتایج موتور استنتاج ترکیبی با تشخیص واقعی

No.	Age (Year)	Smoking	Total Cholesterol (Mg/dl)	Diabetes	Blood Pressure (mmHg)	HDL (mg/dl)	Physicians	Proposed System
1	63	0	233	0	145	38	0.11	0.223
2	67	0	286	0	160	42	0.32	0.449
3	67	0	229	0	120	51	0.2	0.270
4	37	1	250	0	130	57	0.12	0.108
5	56	0	236	0	120	52	0.14	0.136
6	63	0	254	0	130	35	0.32	0.237
7	53	1	203	0	140	58	0.21	0.163
8	57	1	192	0	140	60	0.12	0.160
9	54	0	234	0	123	49	0.09	0.136
10	66	0	284	0	158	40	0.28	0.471
11	61	0	228	1	141	37	0.12	0.233
12	51	1	205	0	138	59	0.19	0.165



شکل۳-۵: سازگاری بین نتایج حاصل از موتور ترکیبی با تشخیص واقعی [۶۰]

در مقایسه این روش با روشهای دیگر در تشخیص بیماری عروق کرونر همانند شبکه عصبی(دقت ۸۴٪)، مطالعه ضربان قلب با استفاده از شبکه های عصبی(دقت ۴۸ ٪) روش پیشنهادی دارای دقت بالاتری در تشخیص این بیماری است.در جدول زیر مقایسه دقت روشهای مختلف در تشخیص این بیماری نشان داده شده است:

جدول۳-۵: مقایسه نتیجه سیستم پیشنهادی با مطالعات مشابه

No.	Research	Method(s)	Accuracy rate
1	(Kukar et al., 1999)	Bayesian classification and neural network	80%
2	(Akay, 1992)	Neural network	84%
3	(Haddad et al., 1999)	Neural network	48%
4	(Detrano et al., 1989)	Probability theory (Logistic regression)	77%
5	(Gennari et al., 1989)	Clustering (CLASSIT conceptual system)	78.9%
6	Proposed hybrid system	Fuzzy sets and evidence theories	91.58%

Polat و GUNES از پیش پردازش وزنی فازی و AIRS استفاده کردند و دقت ۹۲,۵۹٪ برای تشخیص CAD به دست آمد و تا آنجا که ما می دانیم این بهترین دقتی است که به آن رسیدند[۶۱].

در مطالعه بعدی Noor Akhmad Setiawan و همکاران [۶۲]، در این پژوهش یک سیستم پشتیبانی تصمیم گیری فازی برای تشخیص بیماری عروق کرونر بر اساس شواهد مورد توسعه قرار گرفت. مجموعه داده بیماری عروق کرونر از دانشگاه کالیفرنیا ایروین (UCI) جمع آوری شد[۶۳]. دانش مبتنی بر سیستم پشتیبانی تصمیم گیری فازی با استفاده از روش استخراج قوانین بر اساس تئوری Rough Set (RST)گرفته شده است. قوانین انتخاب می شوند و بر اساس اطلاعاتی از گسسته سازی ویژگی های عددی، فازی می شوند. وزن قوانین فازی با استفاده از اطلاعات قوانین پشتیبانی استخراج شده ارائه شده است. در نهایت سیسنم با مجموعه داده های بیماری های قلبی از ایالات متحده، سوئیس و مجارستان، بیمارستان تخصصی Ipoh مالزی اعتبارسنجی شد. نتایج نشان می دهد که این سیستم قادر به ارائه درصد گرفتگی عروق کرونر بهتر از متخصصان قلب و آنژیوگرافی است. نتایج سیستم پیشنهاد شده توسط سه متخصص قلب متخصص تایید شده و کارآمد تر و مفید در نظر گرفته می شود.

ویژگی های این مجموعه داده ها مربوط به معاینه فیزیکی، آزمایشگاه های تشخیصی و تست استرس جمع آوری شد. وضعیت CAD با استفاده از این ویژگی ها با استفاده از روش آنژیوگرافی عروق کرونر به دست آمده است. برای ساختن سیستم های پشتیبانی تصمیم گیری، ۶۶۱ نفر از مجموعه داده ها کلیولند، مجارستانی، Long Beach انتخاب شدند. داده های بیماران انتخاب شده در سه ویژگی ورودی که Slope، Ca و thal است داده های گمشده داشتند. در این کار، اطلاعات از دست رفته با استفاده از ANN با RST بر اساس مجموعه داده کامل نسبت داده شدند. همچنین از اطلاعات ۲۲ نفر از بیمارستان تخصصی Ipoh، مالزی استفاده کردند.

سپس ازRST برای کشف دانش از مجموعه داده ها ی نسبت داده شده استفاده می شود. مجموعه داده ها در یک فرم جدولی به عنوان سطر و ستون نمایش داده می شود. هر عنصر از یک ردیف یک شی و یا نمونه است که می تواند یک رویداد و یا یک بیمار، و غیره را نشان دهنده و هر ستون نشان دهنده یک ویژگی است که می تواند یک متغیر، مشاهده، مشخصه، و غیره باشد که می تواند برای هر شیء اندازه گیری شود. این جدول به نام سیستم اطلاعات است. به شیوه ای رسمی، یک سیستم اطلاعاتی یک جفت S=(A ،U) است که در آن U یک مجموعه متناهی غیر خالی از اشیاء است به نام جهان (Universe) و A یک مجموعه متناهی غیر خالی از صفات است به طوری که $\alpha : V \to U$ و به ازای هر $A \in \alpha$ و V_α مقدار α نامیده می شود.

در استفاده عملی از سیستم اطلاعات، یک نتیجه از طبقه بندی به عنوان تصمیم گیری داده و بیان شده توسط تنها ویژگی های خاص وجود دارد. این نوع از سیستم اطلاعات ، سیستم تصمیم گیری نامیده می شود. بعبارت دیگر، یک سیستم تصمیم گیری گاهی اوقات جدول تصمیم گیری نامیده می شود که یک سیستم اطلاعات با شکل $S = (D \cup A ،U)$ است که در آن $\{d\}$ و $d \notin A$ یک ویژگی تصمیم گیری و یا تصمیم گیری ساده است. عنصر A ویژگی های شرطی و یا شرط ساده نامیده می شود. جدول تصمیم گیری به صورت زیر تعریف می شود.

$$DS=(U,C \cup D) \qquad (۴-۳)$$

مرحله بعد فیلتر کردنن قانون است که مطالعات بسیاری برای روش فیلتر کردن قانون [۶۷-۶۴] ارائه شده است. در این تحقیقRST ترکیبی مبتنی بر قانون اندازه گیری اهمیت و پشتیبانی فیلتر کردن برای انتخاب قوانین با تبدیل قوانین به جداول تصمیم گیری ارائه شده است. روش فیلتر بر اساس پشتیبانی قانون برای انتخاب قوانین جهت کاهش تعدادی از قوانین قبل از اعمال قانون اندازه گیری اهمیت برای انتخاب قواعد مهم اعمال می شود.

اصلاحات با اعمال این روش به سیستم تصمیم گیری و تبدیل قوانین به جداول تصمیم گیری بر اساس داده های تست شده معرفی می شود که مجموعه داده کامل CAD به جای داده آموزشی که مجموعه داده ناقص CAD برای اندازه گیری اهمیت استفاده می شود. و R به صورت زیر تعریف می شود:

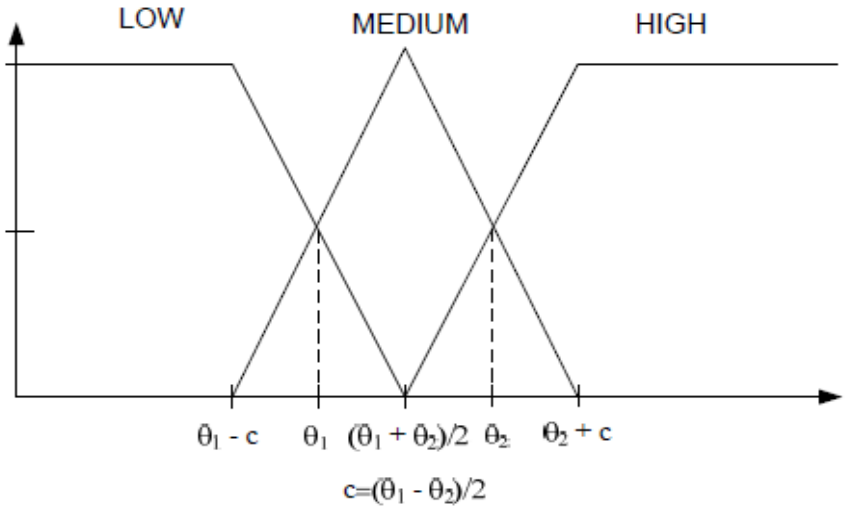
$$R=\{R_1,R_2,\ldots,R_j\} \quad (۵-۳)$$

مرحله بعد، تولید یک سیستم پشتیبانی تصمیم گیری فازی (FDSS) است.

قوانین RST تولید شده انتخابی ، قطعی (Crisp) می باشد و فازی سازی برای این قوانین قطعی صورت می گیردو توابع عضویت فازی به شکل ذوزنقه و مثلث می باشد.

تمام شروط عددی بر اساس مقدار گسسته "کاهش" فازی سازی می شود. به عنوان مثال، سن صفت عددی است با استفاده از استدلال منطقی به سه مقدار گسسته [θ2,*] ,[θ1, θ2) and (*,θ1) که به معنی "کمتر از θ۱" ،"مساوی یا بزرگتر از θ۱ "، و کمتر از θ۲ " و" مساوی یا بزرگتر از θ۲".

دو تابع عضویت ذوزنقه ای و تک مثلثی ویژگی سن LOW(کم)، MEDIUM(متوسط) و HIGH (بالا) می باشد، همانگونه که در شکل زیر نشان داده شده است.



شکل۳-۶: تابع عضویت سن [۶۲]

در اینجا روش فازی وزنی بر اساس پشتیبانی از قوانین RST انتخاب شده از داده های آموزشی ارائه شده است. اگر قانون قطعی n ام دارای پشتیبانی SP(n) باشد سپس وزن قاعده فازی به صورت زیر است.

$$w(n) = \frac{sp(n)}{Max(sp(1), \ldots, sp(i))} \quad (۶-۳)$$

RST پیشنهادی بر اساس روش انتخاب قانون قادر به انتخاب تنها ۲۷ قانون است که چند قانون از ۲۷ قانون در جدول ۳-۶ نشان داده شده است. این مجموعه قوانین با استفاده از مجموعه داده کامل CAD تست شده و در مقایسه با روش های دیگر که در جدول ۳-۷ نشان داده شده است و روش انتخاب پیشنهادی در این تحقیق دارای عملکرد بهتر در دقت و پوشش است.

جدول۳-۶: قوانین RST انتخاب شده

Rule No.	Rules
1.	oldpeak ([0.3,*)) AND slope (2) AND thal (7) => num (1)
2.	fbs (0) AND thalach ([33,*)) AND slope ((1) AND ca ([*,1)) AND thal (3) => num (1)
3.	fbs (0) AND cal([1,*)) AND thal (7) =>(1)
4.	sex 1) AND fbs (0) AND thalach ([33,*)) AND excang (0) AND ca ([*,1)) AND thal (3) => num (1)
5.	sex (1) AND fbs (0) AND restecg (0)) AND oldpeak ([0.3,*)) AND thal (7) => num (1)

27.	Age ([53,*)) AND tresbps ([129,*)) AND restecg (0) AND excang (0) AND ca ([*,1)) => num (1)

جدول۳-۷: کارایی قانون انتخابی

Selection Methods	Accuracy	Coverage	Number of rules
Proposed Method	0.852	0.937	27
Support Based (Training Data)	0.847	0.799	29
Support Based (Testing Data)	0.844	0.868	27
Michalski μ=0.5	0.845	0.785	27
Torgo	0.845	0.785	27
Brazdil	0.845	0.785	27
Pearson	0.845	0.785	27
Cohen	0.863	0.65	29

برای تست عملکرد سیستم پشتیبانی تصمیم گیری فازی (FDSS) ، هر چهار مجموعه داده ها از – UCI CAD که در Cleveland ، Hungarian ، Long Beach وSwitzerland می باشد استفاده می شود. تمام مجموعه داده ها شامل ارزش های از دست رفته است. Cleveland تنها ارزش های از دست رفته در شش اشیاء را دارد . Switzerland دارای بیشترین تعداد از ارزش های از دست رفته[1] است . برای مقایسه، ANN

[1] Missing Value

پرسپترون چند لایه ، k نزدیکترین همسایه ، C4.5 وروش RIPPER نرم افزار WEKA برای تشخیص CAD در چهار مجموعه داده ها UCI - CAD و مجموعه داده ها بیمارستان تخصصی Ipoh اجرا شده است. نتایج را می توان در جداول زیر دید و دیده می شود که FDSS دارای عملکرد خوب در هر پنج مجموعه داده است. تنها روش K- NN عملکرد بهتری دررابطه با دقت روی مجموعه داده Hungarian وLong Beach دارد، به این خاطر که داده های آموزشی از این مجموعه داده می آیند. K-NN بر اساس شباهت با محاسبه فاصله اقلیدسی بر اساس داده های تست شده یا طبقه بندی شده و یا داده های آموزشی می باشد . مجموعه داده Hungarian و Long Beach بیشترین شباهت برای مجموعه آموزشی را دارند و برای Cleveland ، Switzerland و Ipoh، FDSS نتایج بهتری ارائه شده است.

جدول۳-۸: کارایی FDSS روی مجموعه داده Cleveland

Methods	Accuracy	Sensitivity	Specificity
FDSS	0.83	0.81	0.85
MLP-ANN	0.81	0.77	0.85
k-NN	0.81	0.84	0.79
C4.5	0.82	0.79	0.85
RIPPER	0.83	0.82	0.84

جدول۳-۹ : کارایی FDSS روی مجموعه داده Hungarian

Methods	Accuracy	Sensitivity	Specificity
FDSS	0.84	0.70	0.91
MLP-ANN	0.54	0.44	1.00
k-NN	0.92	0.87	0.95
C4.5	0.66	0.52	0.89
RIPPER	0.68	0.54	0.87

جدول۳-۱۰: کارایی FDSS روی مجموعه داده Long Beach

Methods	Accuracy	Sensitivity	Specificity
FDSS	0.75	0.83	0.49
MLP-ANN	0.76	0.75	1.00
k-NN	0.85	0.85	0.81
C4.5	0.77	0.78	0.60
RIPPER	0.78	0.81	0.59

جدول۳-۱۱ کارایی FDSS روی مجموعه داده Switzerland

Methods	Accuracy	Sensitivity	Specificity
FDSS	0.70	0.71	0.50
MLP-ANN	0.81	0.93	0.06
k-NN	0.62	0.96	0.10
C4.5	0.53	0.97	0.10
RIPPER	0.41	0.96	0.08

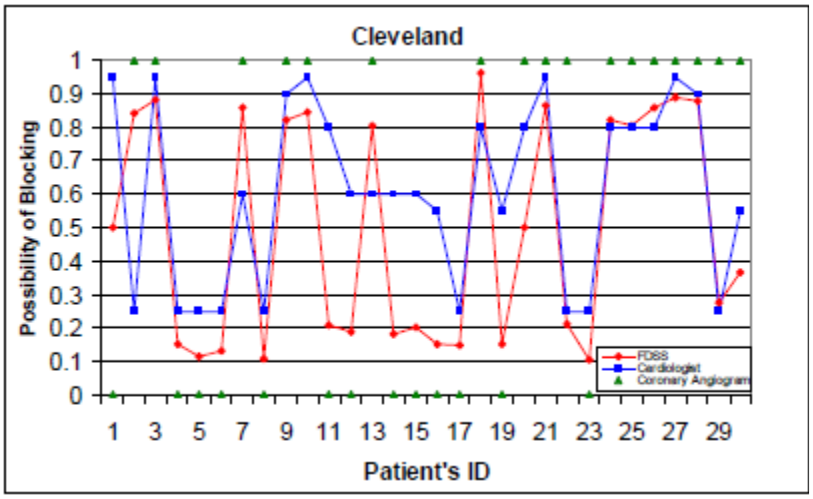
جدول۳-۱۲: کارایی FDSS روی مجموعه داده Ipoh

Methods	Accuracy	Sensitivity	Specificity
FDSS	0.82	1.00	0.56
MLP-ANN	0.77	0.92	0.56
k-NN	0.73	0.69	0.78
C4.5	0.55	0.85	0.11
RIPPER	0.68	0.69	0.67

جدول۳-۱۳: کارایی FDSS و ۳ متخصص قلب و عروق

Metrics	Diagnosis Results			
	FDSS	Cardiologist#1	Cardiologist#2	Cardiologist#3
Accuracy	0.87	0.67	0.67	0.73
Sensitivity	0.82	0.82	0.76	0.76
Specificity	0.92	0.46	0.54	0.69

FDSS قادر به دادن مقادیر درصدی تا حدودی واقعی از گرفتگی عروق کرونر است. اما متخصصان قلب یا آنژیوگرافی نمی توانند چنین نتیجه دهند. به عنوان مثال، نتایج گرفتگی برآورد شده توسط یک متخصص قلب و عروق و FDSS از ۳۰ نفر از مجموعه داده Cleveland در شکل ۳-۷ نشان داده شده است.



شکل۳-۷: نتیجه تشخیص FDSS، متخصص قلب و عروق و آنژیوگرافی عروق کرونر [۶۲]

در تحقیق بعدیGamberger و همکاران، سیستم یادگیری استقراء توسط مینیمم منطقی [1] (ILLM) را ارائه کردند. هدف استفاده از روش یادگیری ماشین برای پیدا کردن اطلاعات استخراج شده مهم و مفید از داده های پزشکی در تشخیص بیماری عروق کرونری است[۶۸].

تحقیق دیگری در رابطه با تئوری Rough Set (RST) برای مدل سازی قدرت موثر تست های قلبی توسط Komorowski و Ohrn ارائه شده است. این کار با اسکن یک گروه از بیماران با استفاده از روش Rough Set بررسی و شناسایی شد[۶۹].

[1] Inductive Learning by Logic Minimization

کار پژوهشی دیگر در زمینه تشخیص خودکار در CAD بر اساس القاء قانون و مدل سازی فازی توسط Tsipouras، و همکاران ارائه شده است. روش القاء قانون برای استخراج قوانین به طور غیر مستقیم با الگوریتم C4.5 انجام شد [۷۰] [۷۱].

۳-۳-۳ ماشین بردار پشتیبان (SVM)

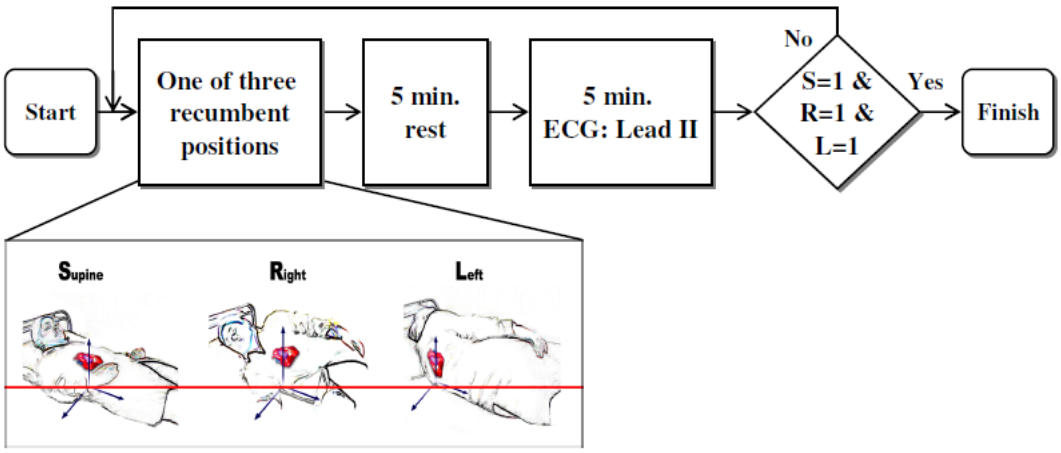
ازآنجایی که روشهایی مانند درخت تصمیم گیری را نمی توان به راحتی در مسائل مختلف به کار برد. این روش از جمله روش های نسبتاً جدیدی است که در سال های اخیر کارایی خوبی نسبت به روش های قدیمی تر برای طبقه بندی از جمله شبکه های عصبی پرسپترون نشان داده است. مبنای کاری دسته بندی کنندۀ SVM دسته بندی خطی داده ها است و در تقسیم خطی داده ها سعی می شود خطی انتخاب گردد که حاشیه اطمینان بیشتری داشته باشد. ماشین بردار پشتیبان حداکثر حاشیه الگوریتم طبقه بندی ریشه در تئوری یادگیری آماری است. این روش برای طبقه بندی داده های هر دو خطی و غیر خطی است. در واقع از یک نگاشت غیر خطی در ابعاد جدید برای تبدیل داده های آموزشی اصلی به یک بعد بالاتر استفاده می کند. با استفاده از نگاشت غیر خطی مناسب داده های دو کلاس توسط یک ابرصفحه جداشده اند و خطاهای طبقه بندی به حداقل می رسد.

دراولین تحقیق I. Babaog Lu [۷۲]، در سال ۲۰۱۰ ،از داده های تست ورزش استفاده کرد و یک سیستم دسته بندی با استفاده از ماشین بردار پشتیبان طراحی شد .در این تحقیق، به جای استفاده ازتمام تمام ویژگی ها، از بهینه سازی ازدحام ذرات باینری و ژنتیک الگوریتم، برای انتخاب ویژگی استفاده شد .با توجه به کاهش بعد مجموعه داده، کارایی روش، بهتر از زمانی بود که از ماشین بردار پشتیبان به تنهایی استفاده شد .نتایج به دست آمده حاکی از این بود که استفاده از بهینه سازی ازدحام ذرات باینری، برای انتخاب ویژگی موفق تر از الگوریتم ژنتیک می باشد .دقت روش، زمانیکه از بهینه سازی ازدحام ذرات باینری برای انتخاب ویژگی استفاده شد، برابر ۸۱,۴۶ ٪ و در موردی که از ژنتیک الگوریتم به عنوان انتخاب کننده ی ویژگی استفاده شد، ۷۹,۱۷ ٪ گزارش شد در حالیکه دقت برای زمانی که از بردار ماشین پشتیبان به تنهایی استفاده شد، ۷۶,۶۷ ٪ گزارش شده است.

در تحقیق دیگرHeon Gyu Lee و همکاران[۷۳]، تشخیص بیماری عروق کرونر با استفاده از ویژگی های خطی و غیر خطی HRV یعنی دامنه زمانی ودامنه فرکانس را مورد بررسی قرار دادند.در این مطالعه روش

جدیدی برای توسعه ویژگی های چند پارامتری از جمله ویژگی های خطی و غیر خطی HRV [1] (تغییرات ضربان قلب) برای تشخیص بیماری های قلبی عروقی ارائه شد. این مطالعه، ویژگی های خطی و غیر خطی از HRV را برای سه حالت خوابیدن، یعنی خوابیده به پشت(طاق باز)، موقعیت جانبی چپ و راست با استفاده از طبقه بندی کننده SVM تحلیل می کند که دارای دقت قابل قبولی می باشد.

ابتدا فرآیند استخراج ویژگی های خطی و غیر خطی با تجزیه و تحلیل HRV از سیگنال های زیستی [2]اولیه (خام) صورت گرفت. سیگنال زیستی ECG با تپش در طول ۵ دقیقه برای هر ۳ حالت خوابیده به ترتیب خوابیده به پشت، موقعیت های جانبی راست و چپ ظبط شد.



شکل۳-۸: اندازه گیری سیگنال زیستی ECG برای هر موقعیت [۷۳]

در این تپش، سیگنال آنالوگ اندازه گیری شده به یک سیگنال دیجیتال با فرکانس نمونه برداری ۵۰۰ هرتز تبدیل شد.

همانطور که گفته شد ویژگی های خطی HRV دامنه زمانی و دامنه فرکانس می باشد. در دامنه فرکانس فواصل بین ضربان قلب با استفاده از سری فوریه سریع (FFT) محاسبه می شود و در دامنه زمانی فواصل بین ضربان لحظه ای قلب اندازه گیری شد.

[1] Heart Rate Variability
[2] Biosignal

در ویژگی های غیر خطی ضربان قلب را با استفاده از روش غیر خطی مانند آنتروپی[1] یا بی نظمی تقریبی (ApEn) تجزیه و تحلیل می کنیم.

جدول۳-۱۴: ویژگی های خطی و غیر خطی HRV

	Feature	Description
Linear	*nLF*	Normalized low frequency power
	nHF	Normalized high frequency power
	LF/HF	The ratio of low- and high-frequency power
	RRm	The mean of RR intervals
	SDRR	Standard deviation of all RR intervals
	SDSD	Standard deviation of differences between adjacent RR intervals
Nonlinear	*SD1*	Standard deviation of the distance of RR(i) from the line $y = x$ in the Poincare plot
	SD2	Standard deviation of the distance of RR(i) from the line $y = -x + 2RR_m$ in the Poincare plot
	SD2/SD1	The ratio of SD2 and SD1
	SD1SD2	SD1 x SD2
	ApEn	Approximate Entropy

سپس مدل طبقه بندی SVM ساخته شد، دراین مدل هر شی به یک نقطه در یک فضای ابعاد بالا نگاشت می شود، هر بعد مربوط به ویژگی هاست.مختصات نقاط، فرکانسی از ویژگی ها در ابعاد مربوطه می باشد.یادگیرنده SVM در گام آموزش، یک Hyper-Plane ماکزیمم ، هر کلاس را جدا می کند. در مرحله آزمایش، یک شی جدید را به یک نقطه در همان فضای ابعاد بالا بوسیله یک یادگیرنده Hyper-Plane موجود در گام آموزش تقسیم می کند و از شبکه تابع مبنای شعاعی [2] (RBF) استفاده کردند.

با توجه به جدول زیر SVM دارای عملکرد مناسبی در رابطه با تشخیص بیماری می باشد.

جدول۳-۱۵: عملکرد SVM

Classifier	TP	FP	Precision	Recall	Class
SVM	0.909	0.096	0.909	0.909	CAD
	0.904	0.091	0.904	0.904	Normal

[1] Entropy
[2] Radial Basis Function

۳-۳-۳-۱ الگوریتم SMO

بهینه سازی مینیمم متوالی (SMO) یک الگوریتم برای حل موثر مشکل بهینه سازی است که در طول آموزش از ماشین های بردار پشتیبان [1](SVM ها) ناشی می شود. که توسط جان پلات در سال ۱۹۹۸ در تحقیقات مایکروسافت معرفی شد. SMO به طور گسترده ای برای آموزش SVM استفاده می شود[۵۱].

انتشار الگوریتم SMO در سال ۱۹۹۸ هیجان زیادی را در طبقه بندی کننده SVM، به عنوان روشی که قبلا برای آموزش SVM در دسترس بوده و بسیار پیچیده تر بود و نیاز به حل کننده های QP (طرف سوم) پر هزینه داشت، ایجاد کرد.

در تحقیقی دکتر علیزاده ثانی و همکاران[۵۱]، با استفاده از مجموعه داده ای به نام ZAlizadeh ثانی با ۳۰۳ نفر و ۵۴ ویژگی، طبقه بندی کننده SVM با الگوریتم SMO را برای تشخیص بیماری عروق کرونر ایجاد کردند. در این مطالعه چندین ویژگی جدید مانند EF، منطقه RWMA، موج Q و موج وارون T به منظور افزایش دقت تشخیص در نظر گرفته شد و نتایج حاصل از روش آنژیوگرافی استاندارد به عنوان مقایسه پایه ، برای ارزیابی قابلیت پیش بینی الگوریتم طبقه بندی استفاده شد.

در انتخاب ویژگی، "وزن توسط SVM" در RapidMiner در نظر گرفته شد.از ضرایب بردار نرمال از SVM خطی به عنوان وزن ویژگی استفاده کردند. از میان ویژگی های بسیار، ۳۴ از آنها که وزن بالاتر از ۰,۶ داشتند، انتخاب شدند و الگوریتم بر روی آنها اعمال شد.

در بخش بعدی، یک الگوریتم برای ایجاد سه ویژگی جدید به نام شناسندهLAD، شناسنده LCX ، شناسنده RCA ارائه شده است. این ویژگی ها برای تشخیص اینکه آیا سه سرخرگ اصلی کرونر نزولی قدامی چپ (LAD) ، بطن چپ [2] (LCX) ، و یا عروق کرونر راست (RCA) مسدود شده است یا نه استفاده می شود.مقادیر بالا از هر یک از این ویژگی های ایجاد شده، احتمال زیاد داشتن CAD را نشان می دهد.

به عنوان نتیجه SMO با در نظر گرفتن انتخاب ویژگی و ایجاد ویژگی در تشخیص بیماری به دقت بالای ۹۴,۸٪ رسید.

[1] Support Vector Machines
[2] Left Circumflex

KNN ۴-۳-۳

در طبقه بندی k-NN [1] یا طبقه بندی بر اساس نزدیک ترین همسایه، خروجی، عضو کلاس است. یک شی با رای اکثریت همسایگان خود طبقه بندی شده، (k یک عدد صحیح مثبت و به طور معمول کوچک است.) اگر kبرابر با یک باشد، شی به کلاس نزدیک ترین همسایه واحد اختصاص داده خواهد شد.

در[۴۹]، تبدیل سریع فوریه و سیستم تشخیص ایمنی مصنوعی و نهایتا روش نزدیکترین همسایه مجاور، استفاده شد. در فاز آموزش، با استفاده از اطلاعات پیش پردازش شده از سیگنالهای داپلر، به سیستم ایمنی مصنوعی، آموزش داده می شود و سپس دسته بندی به روش نزدیکترین همسایه مجاور صورت می گیرد . دقت روش برابر ۹۹,۲۹٪، اختصاصی بودن برابر ۹۹,۸۳٪ و حساسیت برابر ۱۰۰٪ می باشد.

در پژوهش آقای مکرینی و همکارانشان[۵۵]، استفاده از یک شبکه عصبی LVQ4 یا چندی سازی برداری یادگیر و یک طبقه بندی کننده KNN پیشنهاد گردید. در این تحقیق در الگوریتم نزدیکترین همسایه مجاور استفاده شده درLVQ معیار فاصله اقلیدسی با یک معیار بر اساس وزن دهی جایگزین می گردد که این وزن دهی با توجه به اهمیت هر متغیر در تصمیم گیری می باشد. در این تحقیق از مجموعه داده مربوط به بیماری گرفتگی عروق کرونری که از جمله مجموعه دادههای موجود در UCI می باشد، جهت اعتبارسنجی روش استفاده شده و به دقت ۸۰ درصد رسیدند.

Statical Classifier ۵-۳-۳

یکی از انواع طبقه بندی کننده ها،طبقه بندی کننده های خطی است که خود دارای دو نوع متداول به نام های طبقه بندی کننده Bayesian و رگرسیون منطقی می باشد که در ادامه به توضیح تحقیق های انجام شده در این زمینه ها می پردازیم.

۱-۵-۳-۳ طبقه بندی کننده Logisitc Reggression

رگرسیون منطقی یکی از انواع مدل های طبقه بندی آماری می باشد که رابطه بین یک متغیر طبقه بندی شده وابسته و یک یا چند متغیر مستقل دیگر را اندازه گیری می کند.

در تحقیقی Detrano و همکاران، یک مدل تابع تبعیض [1] جدید برای تخمین احتمال بیماری عروق کرونر آنژیوگرافی ساختند [۷۴]. این تابع بر اساس رگرسیون منطقی [2] است که به راحتی قابل تفسیری نیست.

[1] K-Nearest Neighborhood

تحقیق دیگری، در نتایج تجربی رابرت Detrano دقت طبقه بندی صحیح حدود ۷۷٪ با رگرسیون منطقی مشتق شده از تابع گسسته را نشان داد [۷۵].

در تحقیق بعدی، عمران کورت، TURE Mevlut، A. Turhan Kurum کارایی رگرسیون منطقی، طبقه بندی و درخت رگرسیون و شبکه های عصبی را برای پیش بینی بیماری عروق کرونر مقایسه کردند و به دقت ۷۰٪ رسیدند [۷۶].

۳-۳-۵-۲ طبقه بندی کننده Naïve Bayes

بیزین ساده بر پایه احتمالات است. در برنامه های کاربردی مانند طبقه بندی متن و تشخیص پزشکی، این روش دارای بازده بالا است. در این روش فرض بر یک فرض ساده است که مقادیر ویژگی ها مستقل هستند. که از فرمول بیز استفاده می کند[۷۷].

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)} \qquad (۷-۳)$$

در مطالعه ای دکتر علیزاده ثانی و همکاران[مقاله شماره ۶]، با بررسی اثرات مجموعه ای از ویژگی ها، از جمله داده های آزمایشگاهی و اطلاعات اکو که برخی از آنها در مطالعات قبلی در نظر گرفته نشده بودند و با استفاده از طبقه بندی کننده بیزین ساده و مقایسه آن با روش های استفاده شده دیگر، بیماری عروق کرونری را تشخیص دادند.

مجموعه داده شامل اطلاعات جمع آوری شده از ۳۰۳ بازدید کننده در مرکز پزشکی و تحقیقات قلب و عروق شهید رجایی تهران است که یکی از بزرگترین بیمارستان قلب در آسیا است. ویژگی های اضافه شده جدید از جمله محدوده ای با RWMA و جز جهشی یا تخلیه ای [۳] (EF)، هموگلوبین، WBC، پروتئین دفعی در ادرار، HDL، LDL، VHD که در CAD تاثیر به سزایی دارد در نظر گرفته شد.در نهایت این روش دارای دقت نسبتا بالایی، حدود ۷۵٪ گزارش شد و با روشهای دیگر مورد مقایسه قرار گرفت که نتایج در جدول زیر قابل مشاهده است.

[1] Discriminant function model
[2] Logistic regression
[3] Ejection Fraction

جدول۳-۱۶: مقایسه دقت الگوریتم ها با ویژگی های انتخاب شده

Algorithm Used	Accuracy	Sensitivity	Specificity
Naïve Bayes	74.89%±9.34%	72.22%	81.61%
C4.5	78.23%±4.09%	87.50%	55.17%
AdaBoost	76.86%±5.88%	78.70%	72.41%
SMO	82.16%±5.45%	90.74%	60.92%

در تحقیق دیگر، Srinivas و همکاران از روش بیزی و C4.5 بر روی مجموعه داده های مشابه استفاده کردند و به دقت ۸۲٪ رسیدند[۷۸].

در تحقیق دیگری، Rajkumar و Reena از الگوریتم های ساده بیز و درخت تصمیم در مجموعه داده UCI استفاده کردند و به دقت ۵۲,۳۳٪ رسیدند[۵۱]. همچنین Lavesson و Halling الگوریتم های Naïve Bayes , Bagging , Ada boost روی مجموعه داده اعمال کردند و به دقت ۷۱٪ با استفاده از Naïve Bayes رسیدند[۵۱].

۳-۴ نتیجه گیری

تکنیک های داده کاوی، نوید بزرگی را برای کشف الگوهای مخفی در داده ها خصوصا در داده های پزشکی ارائه و عرضه کردند که می تواند به متخصصین بالینی در تصمیم گیری در رابطه با تشخیص بیماری عروق کرونری کمک کند. در حال حاضر، آنژیوگرافی برای تعیین میزان و محل تنگی عروق قلب استفاده می شود که گران قیمت و دارای عوارض جانبی متعدد است به همین دلیل بسیاری از محققان انگیزه استفاده از داده کاوی برای تشخیص CAD را پیدا کردند.

از مطالعه ی فوق مشاهده می شود که دقت برای تحلیل تشخیص تکنیک های گوناگون طبقه بندی داده کاوی اعمال شده تا حد زیادی قابل قبول است و می تواند به متخصصان پزشکی در تصمیم گیری برای تشخیص بهترکمک کند. از این تکنیک های طبقه بندی در رابطه با تشخیص بیماری عروق کرونر استفاده کردیم و الگوریتم ها و تکنیک های مختلفی را مورد بررسی قرار دادیم که درخت تصمیم و شبکه های عصبی مصنوعی به طور موفقیت آمیزی در تشخیص این بیماری بکار گرفته شد.

فصل چهارم

نتیجه گیری و فعالیت های آتی

۴-۱ نتیجه گیری

بیماری گرفتگی عروق کرونری شایعترین بیماری قلبی بوده که سالانه مرگ و میر زیادی را در دنیا به همراه دارد.در حال حاضر استاندارد طلایی برای تشخیص گرفتگی شریانهای کرونری، آنژیوگرافی عروق کرونری می باشد که روشی گران و تهاجمی بوده و ممکن است مشکلاتی را برای بیمار به وجود آورد. هدف از این گزارش تشخیص بیماری گرفتگی عروق کرونری با استفاده از معیارهای غیرتهاجمی و تکنیک های طبقه بندی در داده کاوی می باشد.

در این گزارش تحقیقاتی در فصل اول به بیان مقدمات گزارش پرداختیم , اهداف و فرضیات مسئله را مورد بررسی قرار دادیم که هدف از این مطالعه علاوه بر تشخیص بیماری عروق کرونر، سهولت در تعیین نوع درمان می باشد.

در فصل دوم به معرفی و مرور اجمالی داده کاوی، روشهای آن، تعاریف دسته بندی شامل روش ها، مزایا و معایب پرداخته شد.

در فصل سوم به مرور مقالاتی پرداختیم که به پیش بینی و تشخیص بیماری عروق کرونر پرداختند و این روش ها را مقایسه کردیم. در واقع همه محققین در مطالعه های خود از مجموعه داده های خاصی استفاده کردند که در این مجموعه داده ها یا از داده های استاندارد استفاده کردند و یا داده ها را جمع آوری نمودند و در نهایت استخراج ویژگی را انجام دادند و تکنیک های طبقه بندی همانند درخت تصمیم , شبکه های عصبی، SVM ، قانون Bayes و غیره بررسی کردند و دقت حاصل از این روش های به دست آمده را مقایسه نمودند.البته محققین از متدهایی مانند رگرسیون ساده، رگرسیون چند جمله ای و رگرسیون خطی در تشخیص این بیماری استفاده نکردند که به عنوان کار آتی می توان روی این متدها کار کرد.

در فصل آخر هم به نتیجه‌گیری و پیشنهاد فعالیت‌های آتی پرداخته شد.

۴-۲ فعالیت های آتی

در ساخت مدل طبقه بندی برای پیش بینی احتمال ابتلا به بیماری عروق کرونر، داده ها توسط افراد متخصص از منابع مختلفی جمع آوری می شود و متخصصین در زمان جمع آوری این داده ها برخی از فاکتورهایی که مهم نبودند را حذف کرده باشند. در فصل سوم بعضی از مجموعه داده های مربوط به گرفتگی شریانهای کرونری شامل تعداد بسیار زیادی داده گم شده بود. این داده ها به خصوص برای ویژگی هایی که تا

حد زیادی در تشخیص مؤثرند می تواند دقت دسته بندی را بسیار تحت تأثیر قرار دهد .در پژوهش های آتی سعی داریم با روش هوشمند مسآله داده های گم شده را حل نماییم.

در واقع می توان ساخت درخت تصمیم را با مقادیر گم شده (Missing Value) انجام داد و با معیار های جداسازی مختلف مانند آنتروپی، ویژگی های مختلف را استخراج کرد و دوباره طبقه بندی انجام داد و با یک معیاری مانند Naive Bayes این مدل جدید را با مدل قبلی مقایسه کرد و مدلی که دارای دقت بالاتری است را به عنوان مدل بهتر در نظر گرفت.

از طرفی در این گزارش محققین از متدهایی مانند رگرسیون ساده، رگرسیون چند جمله ای و رگرسیون خطی (موجود در درختواره) در تشخیص این بیماری استفاده نکردند که به عنوان کار آتی می توان روی این متدها کار کرد و با استفاده از این روش ها سیستمی برای تشخیص بیماری عروق کرونری ایجاد نمود.

مراجع

[١] R.Das, I. Turkoglu, A. Sengur "Effective diagnosis of heart disease through neural networks ensembles," Expert Systems with Applications, pp. 7675–7680, 2009.

[٢] American heart association (AHA). http://www.americanheart.org (last accessed: January 2011).

[٣] S. B. Kotsiantis, Supervised Machine Learning: A Review of ClassificationTechniques, Information, Vol.31, pp.249-268, 200

[٤] Ian H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Technniques, Vol.1, Morgan Kaufmann publications, PP. 363-483, 2005

[٥] Bradley, P. S. Fayyad, U. M. Mangasarian, O.L., Mathematical Programming for Data mining: Formulations and Challenges. INFORMS Journal on Computing 11, PP.217 – 238,1999

[٦] Larose, Daniel T., Discovering knowledge in data: an introduction to data mining, John Wiley & Sons, Inc., 2005

[٧] Hussein A. Abbass, Ruhul A. Sarker, Charles S. Newton, Data Mining: A Heuristic Approach, Idea Group Publishing, 2002.

[٨] R. S. Michalski, I. Bratko, and M. Kubat, "Machine Learning and Data Mining: Methods and Applications," Wiley, New York, 1998.

[٩] D. Martens, M. D. Backer, R. Haesen, J. Vanthienen, M. Snoeck, and B. Baesens, "Classification with Ant Colony Optimization," IEEE Trans on Evolutionary Computation, vol. 11, pp. 651-656, 2007.

[١٠] K. Fukunaga, "Introduction to Statistical Pattern Recognition," New York: Academic, 1972.

[١١] H. Ishibuchi, T. Murata, and I. B. Turksen, "Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern

classification problems", Fuzzy Sets and Systems, vol. 89, no. 2, pp. 135-149, July, 1997.

[١٢] M. Zolghadri Jahromi, M. Taheri, A proposed method for learning rule weights in fuzzy rule-based classification systems, Fuzzy Sets and Systems, Vol. 159, PP.449 – 459, 2008

[١٣] H. Ishibuchi, T. Yamamoto, Rule weight specification in fuzzy rule-based classification systems, IEEE Trans. Fuzzy Systems ,Vol.13 , PP.428 –435, 2005.

[١٤] E .G. Mansoori, M.J. Zolghadri, S.D. Katebi, A weighting function for improving fuzzy classification systems performance, Fuzzy Sets and Systems, Vol. 158, PP. 583–591, 2007.

[١٥] M.J. Zolghadri, E.G. Mansoori, Weighting fuzzy classification rules using receiver operating characteristics (ROC) analysis, Inform. Sci, Vol. 177, PP. 2296–2307, 2007

[١٦] H. Ishibuchi, T. Nakashima, T. Morisawa, Voting in fuzzy rule-based systems for pattern classification problems, Fuzzy Sets and Systems, Vol. 103, PP. 223–238, 1999

[١٧] Minas A. Karaolis, *Member, IEEE*, Joseph A. Moutiris, Demetra Hadjipanayi, and Constantinos S. Pattichis, *Senior Member*, *IEEE*,Assessment of the Risk Factors of Coronary Heart Events Based on Data Mining With Decision Trees

[١٨] R. Chaves, J. Ramirez, J.M. Górriz, M. Lopez, D. Salas-Gonzalez, I. Alvarez, F. Segovia, SVM-based computer-aided diagnosis of the Alzheimer's disease using t-test NMSE feature selection with feature correlation weighting, Neuroscience Letters, Vol. 461, PP.293-297,2009.

[١٩] J. Ramirez, J.M. Gorriz, D. Salas-Gonzalez, A. Romero, M. Lopez, I. Alvarez, M. Gomez-Rio, Computer-aided diagnosis of Alzheimer's type dementia combining support vector machines and discriminant set of features, Information Sciences, 2009.

[٢٠] R. Lin, An intelligent model for liver disease diagnosis, Artificial Intelligence in Medicine, Vol. 47, PP. 53—62, 2009.

[٢١] F. Temurtas, A comparative study on thyroid disease diagnosis using neural networks, Expert Systems with Applications, Vol. 36, PP. 944–949, 2009.

[٢٢] H. Temurtas, N. Yumusak, , F. Temurtas, A comparative study on diabetes disease diagnosis using neural networks, Expert Systems with Applications, Vol. 36, PP. 8610–8615, 2009.

[٢٣] I. Turkoglua, A. Arslan, E. Ilkay, "An expert system for diagnosis of the heart valve diseases", Expert Systems with Applications, pp. 229–236, 2002.

[٢٤] A. Sengur, "An expert system based on principal component analysis, artificial immune system and fuzzy k-NN for diagnosis of valvular heart diseases, Computers in Biology and Medicine, pp. 329 – 338,2008.

[٢٥] A. Sengur, "An expert system based on linear discriminant analysis and adaptive neuro-fuzzy inference system to diagnosis heart valve diseases", Expert Systems with Applications, pp. 214–222, 2008.

[٢٦] E. Avci, "A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier", Expert Systems with Applications, pp. 10618–10626, 2009.

[٢٧] R. Das, I. Turkoglu, A. Sengur," Diagnosis of valvular heart disease through neural networks ensembles", computer methods and programs in biomedicine, pp. 185–191, 2009.

[٢٨] D.-Y. Tsai, M. Tomita, "A computer-aided system for discrimination of dilated cardiomyopathy using echocardiographic images." IEICE Trans. Fundamentals, pp. 1649- 1654, 1995.

[٢٩] S. Watanabe, D.Y.Tsai, K.Kojima, et. ul, " De- termination of weighting coefficients of neural networks trained by multiple-fitness-based genetic algorithm." Med. Imag. Tech., pp. 557-558,1997.

[٣٠] D.-Y. Tsai, S. Watanabe, K. Kojima, ef.af, "A method for optimization of fuzzy reasoning using genetic algorithms and its application to recognition of medical images" Med. Imag. Inform. pp. 53-60, 1998.

[٣١] D-Y Tsai, "Comparison of Four Computer-Aided Diagnosis Schemes for Automated Discrimination of Myocardial Heart Disease", IEEE, 2000.

[٣٢] H. Yan, J. Zheng, Y.Jiang, Ch. Pengl, Q. Li , " Development of a decision support system for heart disease diagnosis using multilayer perceptron." IEEE,2003

[٣٣] H. Yan, Y. Jiang, J.Zhenge, Ch. Pengc, Q. Li , "A multilayer perceptron-based medical decision support system for heart disease diagnosis", Expert Systems with Applications, pp. 272–281, 2006.

[٣٤] Chau, M.; Shin, D., "A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms". Proceedings of IEEE International Conference on Dependable, Autonomic and Secure Computing 2009, pp. 183-187.

[٣٥] J. R. Quinlan, "Simplifying decision trees," *Int. J. Man-Mach. Stud.*, vol. 27, pp. 221–234, 1987.

[٣٦] C. Ordonez, "Comparing association rules and decision trees for disease prediction," in *Proc. Int. Conf. Inf. Knowl. Manage.,Workshop Healthcare Inf.Knowl. Manage*. Arlington, VA, 2006, pp. 17–24.

[٣٧] C. Ordonez, E. Omiecinski, L. de Braal, C. A. Santana, N. Ezquerra, J. A. Taboada, D. Cooke, E. Krawczvnska, and E. V. Garcia, "Mining constrained association rules to predict heart disease," in *Proc. IEEE Int. Conf. Data Mining (ICDM 2001)*, pp. 431–440.

[٣٨] C. L. Tsien, H. S. F. Fraser, W. J. Long, and R. L. Kennedy, “Using classification trees and logistic regression methods to diagnosemyocardial infraction,” in *Proc. 9th World Congr. Med. Inf.*, vol. 52, pp. 493–497, 1998.

[٣٩] K. Polat, S. Sahan, H. Kodaz, and S. Guenes, “A hybrid approach to medical
decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS,” *Comput.Methods Programs Biomed.*, vol. 88, no. 2, pp. 164–174, 2007.

[٤٠] S. A. Pavlopoulos, A. Ch. Stasis, and E. N. Loukis, “A decision treebased method for the differential diagnosis of aortic stenosis from mitral regurgitation using heart sounds,” *Biomed. Eng. OnLine*, vol. 3, p. 21, 2004.

[٤١] M. Shouman, T. Turner, Using decision tree for diagnosing heart disease patients, in: Proceedings of the 9-th

[٤٢] M. Shouman, T. Turner, Using Decision Tree for Diagnosing Heart Disease Patients, Proceedings of the 9-th Australasian Data Mining Conference (AusDM'11), Ballarat, Australia, 2011.

[٤٣] F. Temurtas, A comparative study on thyroid disease diagnosis using neural networks, Expert Systems with Applications, Vol. 36, PP. 944–949, 2009.

[٤٤] V. Cario, V. Hans, D. Richard. Journal of Analytical Atomic Spectrometry, 8: 781, 1993.

[٤٥] W.You, Y. Wang, B.Wo, S. Lv, A.Zhan, W. Sun, "Recognition of Coronary Heart Disease Patients by RBF Neural Network Basing on Contents of Microelements in Human Blood ", Second International Symposium on Computational Intelligence and Design, IEEE, 2009.

[٤٦] Chen Tian-hua , Xing Su-xia , Guo Pei-yuan , Yu Zhen, "Coronary Heart Disease Based on Neural Network and Heart Sound Signals", IEEE, 2009.

[٤٧] D.Itchhaporia, R.Almassy, L. Kaufman, P.Snow, W. Oetgen, "Artificial neural networks can predict significant coronary disease", Elsevier, 1995.

[٤٨] S. Kara , F. Dirgenali , "A system to diagnose atherosclerosis via wavelet transforms, principal component analysis and artificial neural networks", Expert Systems with Applications, pp. 632–640, 2007.

[٤٩] F. Latifoglu, H. Kodaz, S. Kara, S. Güne¸"Medical application of Artificial Immune Recognition System (AIRS): Diagnosis of atherosclerosis fromcarotid arteryDoppler signals", Computers in Biology and Medicine, pp. 1092 – 1099, 2007.

[٥٠]Oleg Yu. Atkov (MD, PhD)a, Svetlana G. Gorokhova (MD, PhD)b,∗, Alexandr G. Sboev (PhD)c, Eduard V. Generozov (PhD)d,Elena V. Muraseyeva (MD, PhD)e, Svetlana Y. Moroshkinad, Nadezhda N. Cherniyc , Coronary heart disease diagnosis by artificial neural networks including genetic polymorphisms and
clinical parameters, Received 18 November 2011; accepted 21 November 2011 Available online 2 January 2012.

[٥١] I.Babaoglu ,O. K. Baykan , N. Aygul, K. Ozdemir,M. Bayrak, "Assessment of exercise stress testing with artificial neural network in determining coronary artery disease and predicting lesion localization", Expert Systems with Applications, pp. 2562–2566, 2009.

[٥٢] A. Khemphila, V. Boonjing, Heart disease Classification using Neural Network and Feature Selection, 21st International Conference on Systems Engineering, University of Nevada, Las Vegas, USA, 2011.

[٥٣] Brennan, M., Palaniswami, M., Kamen, P.: Do existing measures of Poincaré plot geometry reflect nonlinear features of heart rate variability? IEEE Trans. Biomed. Eng. 48(11), 1342–1347 (2001) 11. Moraru, L., Tong, S.

[٥٤] E.Leitgeb, H.Koller M.D.," Toward improving exercise ECG for detecting ischemic heart disease with recurrent and feed forward neural nets.", IEEE, 1994.

[٥٥] C. E. Pedreira, L. Macrini, E. S. Costa, "Input and Data Selection Applied to Heart Disease Diagnosis", Proceedings of International Joint Conference on Neural Networks, IEEE, 2005.

[٥٦] OJ. Newman, S. Hettich, C.L. Blake, CJ. Merz,. UCI Repository of machine learning databases.

http://www.ics.uci.edu!~mleam!MLRepositol\.html.Irvine,CA: University of California, Department of Information and Computer Science. 1998

[٥٧] Markos G. Tsipouras, *Student Member, IEEE*, Themis P. Exarchos, *Student Member, IEEE*, Dimitrios I. Fotiadis, *Senior Member, IEEE*, Anna P. Kotsia, Konstantinos V. Vakalis, Katerina K. Naka, and Lampros K. Michalis "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling" IEEE TRANSACTIONS VOL. 12, NO. 4, JULY 2008 447

[٥٨] K. Polat , S. Sahan, S. Gunes ," Automatic detection of heart disease using an artificial immune recognition system (AIRS) with fuzzy resource allocation mechanism and k-nn (nearest neighbour) based weighting" , Expert Systems with Applications, pp. 625–631, 2007.

[٥٩] K. N. Anooj, Clinical decision support system: risk level prediction of heart disease using weighted fuzzy rules and decision tree rules, Central European Journal of Computer Science, Vol.1, PP. 482-498, 2011.

[٦٠]Vahid Khatibi, Gholam Ali Montazer, "A fuzzy-evidential hybrid inference engine for coronary heart disease risk assessment", *Information Technology Department, School of Engineering, Tarbiat Modares University, P.O. Box: 14115-179, Tehran, Iran,*2010.

[٦١] K. Polat, S. Gunes, A hybrid approach to medical decision support systems: combining feature selection, fuzzy weighted pre-processing and AIRS, Computer Methods and Programs in Biomedicine 88 (2007) 164–174.

[٦٢]Noor Akhmad Setiawan1, P.A. Venkatachalam2 and Ahmad Fadzil M.Hani3 "Diagnosis of Coronary Artery Disease Using Artificial Intelligence Based

Decision Support System", Proceedings of the International Conference on Man-Machine Systems (ICoMMS) 11 – 13 October 2009, Batu Ferringhi, Penang, MALAYSIA.

[٦٣] D. J. Newman, S. Hettich, C. L. Blake, and C. J. Merz, "UCI Repository of machine learning databases," University California Irvine, Department of Information and Computer Science, 1998.

[٦٤] M. Maddouri and J. Gammoudi, "On Semantic Properties of Interestingness Measures for Extracting Rules from Data," in *Adaptive and Natural Computing Algorithms*, 2007, pp. 148.

[٦٥] J. Li, "Rough set based rule evaluations and their applications." Waterloo, Ontario: University of Waterloo, 2007, pp. 191.

[٦٦] J. Li, P. Pattaraintakorn, and N. Cercone, "Rule Evaluations, Attributes, and Rough Sets: Extension and a Case Study," in *Transactions on Rough Sets VI*, 2007, pp. 152.

[٦٧] J. Li and N. Cercone, "Introducing a Rule Importance Measure," in *Transactions on Rough Sets V*, 2006, pp. 167.

[٦٨] M. G. Tsipouras, T. P. Exarchos, D. I. Fotiadis, A. P. Kotsia, K. V. Vakalis, K. K. Naka, and L. K. Michalis, "Automated Diagnosis of Coronary Artery Disease Based on Data Mining and Fuzzy Modeling," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 12, pp. 447, 2008.

[٦٩] Moraru, L., Tong, S., Malhotra, A., Geocadin, R., Thakor, N., Bezerianos, A.: Investigation of the effects of ischemic preconditioning on the HRV response to transient global ischemia using linear and nonlinear methods. Med. Eng. & Physics 27, 465–473 (2005)

[٧٠] Tulppo, M.P., Makikallio, T.H., Takala, T.E.S., Seppanen, T.: Quantitative beat-to-beat analysis of heart rate dynamics during exercise. Am J. Physiol. 271, 244–252 (1996)

[٧١] Tulppo, M.P., Husghson, R.L., Makilkallio, T.H., Airaksinen, K.E.J., Huikuri, H.V.: Effectsof exercise and passive head-up tilt on fractal and complexity properties of heart rate dynamics. Am J. Phyisol. Heart Circ. Physiol. 280, H1081–H1087 (2001)

[٧٢] I. Babaog Lu, O.Fındık, M. Bayrak "Effects of principle component analysis on assessment of coronary artery diseases using support vector machine", Expert Systems with Applications, pp. 2182–2185, 2010.

[٧٣] Heon Gyu Lee1, Ki Yong Noh2, and Keun Ho Ryu1, Mining Biosignal Data: Coronary Artery Disease Diagnosis Using Linear and Nonlinear Features of HRV

[٧٤] Tompkins, W.J.: Bimedical digital signal processing, p. 07458. Prentice Hall PTR, Upper Saddle River, New Jersey (1995)

[٧٥] Detrano, R.; Steinbrunn, W.; Pfisterer, M., "International application of a new probability algorithm for the diagnosis of coronary artery disease". American Journal of Cardiology, Vol. 64, No. 3, 1987, pp. 304-310.

[٧٦] Kurt, I.; Ture, M.; Turhan, A., "Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease". Journal of Expert Systems with Application, Vol. 3, 2008,
pp. 366-374.

[٧٧]Roohallah Alizadehsani1, Jafar Habibi1, Zahra Alizadeh Sani2,+, Hoda Mashayekhi1, Reihane Boghrati1, Asma Ghandeharioun1, Behdad Bahadorian2 " Diagnosis of Coronary Artery Disease Using Data mining based on Lab Data and Echo Features"

[٧٨] K. Srinivas, G. Raghavendra Rao, A. Govardhan, *"Analysis of Coronary Heart Disease and Prediction of Heart Attack in Coal Mining Regions Using Data Mining Techniques"*, The 5th International Conference on Computer Science & Education, China,pp.1344-1349, 2010.

واژگان انگلیسی به فارسی

ACCURACY	دقت
AMERICAN HEART ASSOCIATION (AHA)	انجمن قلب آمریکا
ARTIFICIAL NERUAL NETWORK	شبکه های عصبی مصنوعی
CLASSIFCATION	طبقه بندی
CRISP MODEL	مدل قطعی
CORONARY ARTERY DISEASE (CAD)	بیماری عروق کرونری
DECISION TREE	درخت تصمیم
DATA MINING	داده کاوی
DECISION SUPPORT SYSETM	سیستم پشتیبان تصمیم گیری
FEATURE EXTRACTION	استخراج ویژگی
FEATURE SELECTION	انتخاب ویژگی
FASTING BLOOD SUGER	قند خون ناشتا
FUZZY RULES SET	مجموعه قوانین فازی
GAIN RATIO	نسبت بهره
GINI INDEX	شاخص جینی
KNOWLEDGE DISCOVERY	کشف دانش
LEARNING VECTOR QUANTIZATION	چندی سازی برداری یادگیر
NEAREST NEIGHBORS	نزدیکترین همسایه
REGRESSION	رگرسیون
REDUCED ERROR PRUNING	هرس خطای کاهش یافته
RESTING BLOOD PRESSURE	فشارخون در حالت استراحت

SUPPORT VECTOR MACHINE	ماشین بردار پشتیبان
SENSITIVITY	حساسیت
SPECIFICITY	اختصاصی بودن
TRAINING DATA SET	مجموعه داده آموزشی
TEST DATA SET	مجموعه داده تست
VOTING	رای گیری