

# Determining the role of factors leading to metabolic syndrome in coronary disease based on data mining approaches

شناسایی نقش عوامل بروز سندرم متابولیک در ابتلا به بیماری عروق  
کرونر - رویکرد داده کاوی

Mehregan Ghobakhloo, Mohammad Mehdi Sepehri, Mohammad Hossein Hasheminejad, and Elmira Homayounfar

**Abstract** — Growth of cardiovascular diseases and their effects on society and the high cost of imports, caused the medical community to pursue plans for further evaluation, prevention, early detection and effective treatment so valuable knowledge can be created by using data mining and knowledge discovery in cardiology centers that the discovered knowledge can improve the quality of service by medical center's managers and It can also be used by doctors to treat and predict cardiovascular disease by their disease history. In this research, to improve the diagnosis of cardiac diseases a classification - feature selection approach has been proposed that were evaluated by using a data set of patient records of Amir-Al-Momenin hospital in Tehran. Our proposed method is selected by using the different ways of classification on existing data of healthy persons and patients to reach the most efficient way and highest accuracy. At the end the learned classifier are evaluated by a set of test specimen. Finally evaluation shows that the classification tree method C4.5 classification accuracy has earned the highest standard of accuracy rate as 95.8%<sup>1</sup>.

**Keywords** — Diagnosis of heart disease, Data mining, Classification, Tree method.

<sup>1</sup> M. Ghobakhloo is with Faculty Industrial Engineering, Tarbiat Modares University, Tehran, Iran (email: ghobakhloo.m@gmail.com)  
M.M. Sepehri is with Faculty of Industrial Engineering, Tarbiat Modares University, Tehran, Iran (email: mehdi.sepehri@modares.ac.ir, Corresponding Author)  
M.H. Hasheminejad is with Faculty of Computer Engineering, Tarbiat Modares University, Tehran, Iran (email: smh.hasheminejad@modares.ac.ir)  
E. Homayoundar is with Faculty Industrial Engineering, Tarbiat Modares University, Tehran, Iran (email: elmira.homayounfar@gmail.com)

سندرم متابولیک که تماماً به عنوان عوامل خطر بیماریهای قلبی عروقی اثبات شده اند، افزایش خطر بیماریهای قلبی عروقی را دو چندان می نماید. لذا محققین به دنبال یافتن روش‌هایی ارزان و دقیق برای تشخیص و پایش این بیماری هستند. همکاری متخصصین در زمینه کامپیوتر و پزشکی راه حل جدیدی را در تحلیل این داده ها و به دست آوردن الگوهای مفید و کاربردی ارائه می دهد که همان داده کاوی می باشد [۳]. در طول ۲۰ سال اخیر، علاقمندی به ارزیابی و بهبود کیفیت زندگی بیماران مبتلا به بیماریهای مزمن افزایش چشمگیری یافته است و بهبود عملکرد روزانه و کیفیت زندگی بیماران مبتلا به بیماریهای مزمن به صورت یک هدف در آمده است [۴].

تاکنون تحقیقات متعددی بر روی مجموعه داده های بیماری قلبی مبتنی بر روش های داده کاوی انجام گرفته است [۵-۱۳]. متداولترین روش هایی که محققین برای تشخیص و پایش بیماری قلبی استفاده نموده اند شامل روش های دسته بندی و خوشه بندی است. ولی آنچه تاکنون کمتر مورد توجه قرار گرفته بررسی قضیه با انواع روش ها و رسیدن به یک اجماع جهت تشخیص بیماری قلبی در کمترین زمان و با صرف کمترین هزینه و کار بر روی مجموعه داده های خاص می باشد. بدین منظور در این تحقیق بهره گیری از انتخاب خصیصه و سپس دسته بندی پیشنهاد شده است که با استفاده از روش Cross-Validation با  $fold=10$  دسته بندی شده است. در این تحقیق روش های دسته بندی  $C4.5$ ، Naive Bayes، SVM و KNN به منظور یادگیری دسته بندها انتخاب شده است. نتایج ارزیابی بدست آمده حاکی از آن است که استفاده از روش  $C4.5$  در مجموعه داده بیماران دارای بالاترین دقت ارزیابی است.

در ادامه این تحقیق، ابتدا مجموعه داده مراجعین بخش قلب بیمارستان امیرالمومنین تهران در بخش ۳ معرفی می گردد، سپس روش پیشنهادی در بخش ۵ بیان می گردد. در بخش ۱۰، نتایج ارزیابی روش مذکور ارائه می شود و این نتایج در بخش ۱۱ با سایر تحقیقات انجام شده مقایسه شده است. در نهایت در بخش ۱۲ نتیجه گیری انجام گرفته و کارهای آتی بیان شده است.

### ۳. مجموعه داده های مورد استفاده

داده های مورد استفاده برای انجام اعمال داده کاوی، عموماً داده‌هایی هستند که برای اهداف دیگری گرد آمده اند. در زمینه داده کاوی در حوزه سلامت، به طور معمول این داده ها از پرونده های مراجعین مراکز درمانی استخراج شده اند. برای انجام این تحقیق از داده های مربوط به پرونده های مراجعین بخش قلب بیمارستان امیرالمومنین تهران استفاده گردید. سپس با در نظر گرفتن اطلاعات موجود در پایگاه داده به دست آمده از کلینیک کلیولند که در مخزن یادگیری ماشین<sup>۳</sup> UCI موجود است و این پایگاه داده دارای ۷۰ نوع ویژگی برای بیماران است که از آن میان ۱۳ ویژگی اساسی را برشمرده است، در این تحقیق نیز از بین این ۱۳ ویژگی ۵ مورد موثر در سندرم متابولیک مورد بررسی قرار گرفت.

### ۴. ویژگی های مورد استفاده

شیوع بسیار بالای سندرم متابولیک در کشورهای پیشرفته و در حال توسعه توجه ویژه بسیاری از محققین را به خود جلب نموده است. تحقیقات حاضر یکی از فاکتورهای ایجاد بیماری قلبی را سندرم متابولیک برشمرده اند [۱۷]. سندرم متابولیک مشتمل بر مجموعه ای از اختلالات متابولیکی همچون چاقی شکمی،

## ۱. چکیده

رشد چشمگیر بیماری‌های قلبی و عروقی و اثرات و عوارض آنها و هزینه های بالایی که بر جامعه وارد می کند، باعث شده که جامعه پزشکی به دنبال برنامه‌هایی جهت بررسی بیشتر، پیشگیری، شناسایی زود هنگام و درمان موثر آن باشد. از اینرو با استفاده از داده‌کاوی و کشف دانش در سیستم مراکز قلب و عروق می توان دانش ارزشمند را ایجاد کرد که این دانش کشف شده میتواند باعث بهبود کیفیت سرویس به وسیله مدیران مرکز شود و همچنین می تواند به وسیله پزشکان استفاده شود تا رفتار آینده بیماران قلبی و عروقی از روی سابقه داده شده پیش بینی کنند. در این تحقیق به منظور بهبود روشهای تشخیص بیماریهای قلبی یک رویکرد دسته بندی - انتخاب خصیصه پیشنهاد شده است که با استفاده از مجموعه داده مربوط به پرونده های مراجعین بخش قلب بیمارستان امیرالمومنین تهران مورد ارزیابی قرار گرفته است. روش پیشنهادی پس از بکارگیری روشهای مختلف دسته بندی بر روی داده های موجود شامل افراد سالم و بیمار و رسیدن به بهینه ترین روش با بالاترین دقت انتخاب می شود. در نهایت دسته بندی های یادگیری شده توسط مجموعه نمونه های آزمون مورد ارزیابی قرار می گیرند. ارزیابی روش های دسته بندی نشان می دهد که روش دسته بندی درختی  $C4.5$  بالاترین دقت را کسب نموده است که دارای معیار صحت به میزان ۹۵.۸٪ می باشد.

**کلمات کلیدی - تشخیص بیماری قلبی، داده کاوی، دسته بندی، روش**

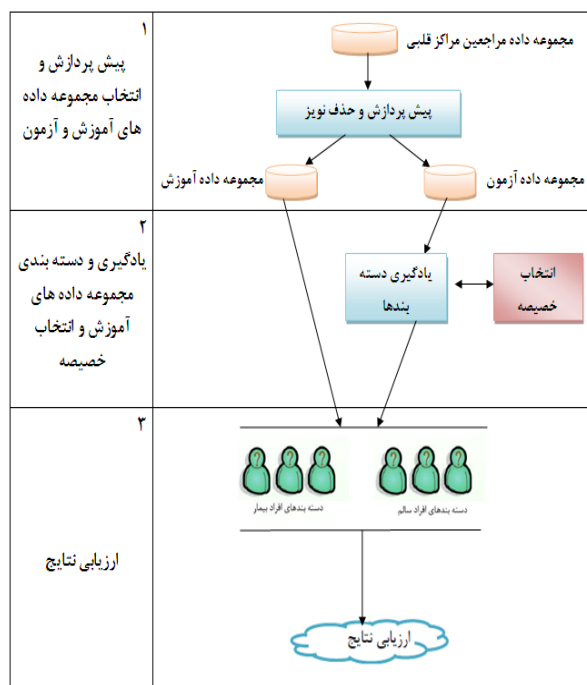
**درختی**

## ۲. مقدمه

در ابتدای قرن بیستم ۱۰٪ کل مرگ و میرها به علت بیماریهای قلبی عروقی بود [۱]. در انتهای همین قرن موارد مرگ و میر ناشی از بیماریهای قلبی به ۲۵٪ افزایش یافت و 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 شود. رشد چشمگیر این بیماریها و اثرات و عوارض آنها و هزینه های بالایی که بر جامعه وارد می کند باعث شده که جامعه پزشکی به دنبال برنامه هایی جهت بررسی بیشتر، پیشگیری، شناسایی زود هنگام و درمان موثر آن باشد. لبتوجه به اینکه بیماریهای قلبی و عروقی در جهان و در کشور ما نیز شایعترین عامل مرگ و میر انسانها هستند اطلاع رسانی درست و مناسب در مورد آنها لازم است. قلب عضوی در قفسه سینه است که کار آن به جریان درآوردن خون درون رگهای بدن است تا اکسیژن و مواد غذایی به همه یاخته ها و اعضاء بدن برسد، به همین دلیل سالم بودن و عملکرد درست آن در سلامت انسان نقش حیاتی دارد. بیماری قلبی به هر بی نظمی که بر توانایی قلب برای انجام وظایفش به صورت نرمال تاثیر گذارد اطلاق می گردد [۲]. یکی از شایع ترین دلایل بروز بیماری قلبی ابتلا به سندروم متابولیک است. سندرم متابولیک<sup>۲</sup> به مجموعه‌ای از اختلالات متابولیک اطلاق می‌شود که عبارت هستند از: فشار خون بالا، سطوح افزایش یافته ی انسولین در خون، چربی انباشته اضافی در اطراف شکم و سطوح افزایش یافته ی چربی خون که وقوع همزمان آنها در هر شخص بیشتر از خطر وقوع احتمالی هر یک به تنهایی است و در سندروم متابولیک در خیلی از موارد، اکثریت قریب به اتفاق این شرایط، همزمان با هم وجود دارند و ریسک ابتلا به بیماری قلبی، سخته و دیابت را افزایش می‌دهند. حضور توأم اجزای

<sup>3</sup> Machine Learning Repository

<sup>2</sup> Metabolic Syndrome



شکل 1- شمایی از روش پیشنهادی مبتنی بر انتخاب خصیصه - دسته بندی

## ۷. انتخاب خصیصه

در گام دوم روش پیشنهادی، انتخاب خصیصه از بین زیرمجموعه داده ها می باشد. تعداد زیرمجموعه های مذکور با در نظر گرفتن ۵ فاکتور موثر در بروز بیماری قلبی و با استفاده از فرمول  $2^n - 1$  محاسبه شده که در این گام از بین ۳۱ زیرمجموعه موجود انتخاب بهترین خصیصه صورت می گیرد. هدف از این گام بررسی تمامی زیرمجموعه های موجود پایگاه داده با فرض وجود مجموعه داده ای با دو برجسب بیمار و سالم به منظور یافتن بهینه ترین و کم هزینه ترین روش شناسایی بیماری است. در مرحله بعد به ازای هر زیرمجموعه داده، یک دسته بندی یادگیری می شود و شناسایی بیماری توسط بررسی و رای گیری بین این دسته بندها انجام می گردد.

نکته مهمی که باید در این بخش ذکر شود این است که انتخاب خصیصه به صورت مجزا بر روی دو گروه از نمونه ها یعنی گروه بیماران قلبی و گروه افراد سالم انجام می شود و مجموعه های بدست آمده به مرحله بعد جهت دسته بندی انتقال داده می شود.

## ۸. دسته بندی

پس از اینکه خصیصه های مختلف مجموعه داده آموزش استخراج گردید، برای هر خصیصه بدست آمده بر اساس یک روش دسته بندی، یک دسته بندی، یادگیری می شود. پس از یادگیری این دسته بندها، هر نمونه از مجموعه آزمون، به همه دسته بندها داده می شود و نظر آنها پرسیده می شود.

نکته ای که باید به آن توجه شود این است که تا کنون روش های دسته بندی متعددی ارائه شده است اما نمی توان چنین نتیجه گیری کرد که یکی از این روش ها، بهترین روش دسته بندی است [۱۵]. برای تشخیص بهترین روش دسته بندی در هر حوزه باید روش های متعدد ارزیابی شوند تا بهترین روش دسته بندی انتخاب گردد. به همین منظور در تحقیق حاضر از روش های دسته بندی KNN<sup>۴</sup>، SVM

دیس لیپیدمی، اختلال در هموستاز قند خون و پرفشاری خون می باشد [۱۸]. بنابر نظر ATP III بیماری های قلبی عروقی اولین پیامد سندرم متابولیک هستند [۱۹]. در افراد مبتلا به این سندرم، خطر مرگ و میر از بیماریهای قلبی عروقی و مرگ و میر از تمام علل بیشتر از افراد فاقد این سندرم است [۲۰]. پنج ویژگی موثر در سندرم متابولیک به اختصار عبارتند از:

۱. بالا بودن وزن یا همان چاقی شکمی
۲. فشار خون Systolic در حال استراحت: فشار خون در زمان ورود به بیمارستان بر حسب میلیمتر جیوه
۳. فشار خون Diastolic در حال استراحت: فشار خون در زمان ورود به بیمارستان بر حسب میلیمتر جیوه
۴. کلسترول سرم خون یا همان دیس لیپیدمی: بر حسب میلی گرم بر دسی لیتر (mg/dl)
۵. قند خون ناشتا یا همان اختلال در هموستاز قند خون

## ۵. روش پیشنهادی مبتنی بر انتخاب خصیصه - دسته بندی

روش پیشنهادی شامل سه گام: ۱. پیش پردازش، ۲. یادگیری دسته بندی و ۳. ارزیابی است. در گام پیش پردازش دو فعالیت اصلی انجام می شود: (۱) پیش پردازش و حذف نویز، (۲) تعیین مجموعه های داده های آموزش و آزمون. پس از حذف نویز و انجام پیش پردازش، مجموعه داده آزمون و آموزش تعیین می گردد که در اینجا از روش 10 Folds Cross validation برای تعیین این دو مجموعه داده، استفاده شد که در آن ۹۰٪ داده ها به عنوان مجموعه آموزش و ۱۰٪ به عنوان مجموعه آزمون بصورت تصادفی انتخاب می گردد. سپس در گام دوم روش پیشنهادی دسته بندهایی با استفاده از مجموعه آموزش یادگیری می شود و در نهایت در گام سوم روش پیشنهادی ارزیابی دسته بندها انجام می گیرد. در شکل ۱ شمایی از روش مذکور نمایش داده شده است و در بخش هی بعد هر یک از گام ها به تفصیل آورده شده است. واحد های اندازه گیری

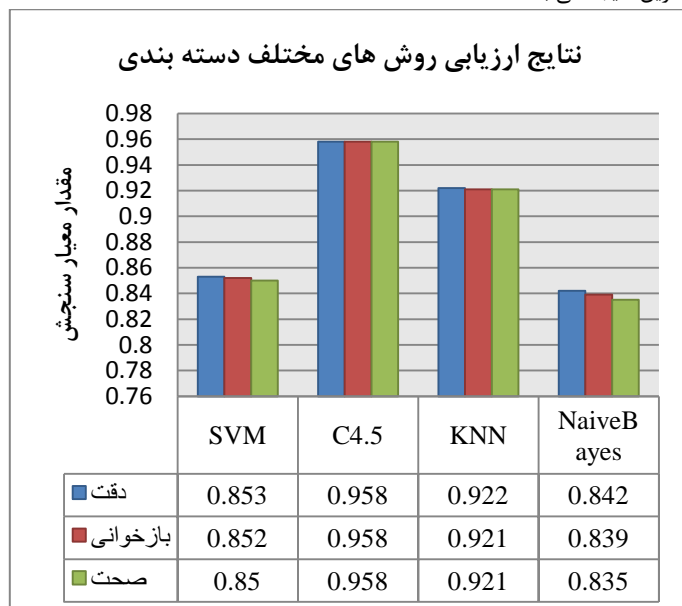
## ۶. پیش پردازش

در روش پیشنهادی منظور از پیش پردازش نمونه های بیماری، جایگزینی داده های از دست رفته می باشد. برای انجام چنین پیش پردازشی، در این تحقیق از رگرسیون خطی ابزار Weka [۱۴] استفاده می گردد. وجود نویز و داده های پرت<sup>۴</sup> همواره تاثیر بدی بر روی نتایج دسته بندی ایجاد می کند به همین سبب در استفاده از روش های داده کاوی ابتدا تلاش می شود که این نمونه ها از مجموعه داده پاکسازی شود.

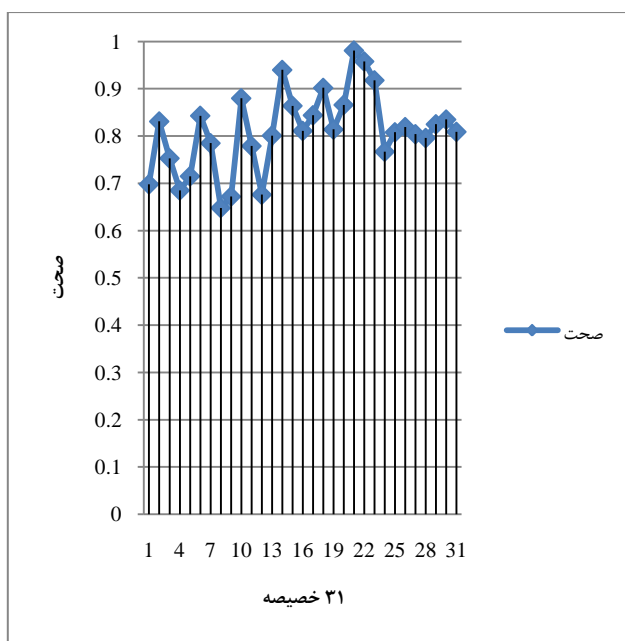
<sup>۵</sup> K-Nearest Neighbor

<sup>۴</sup> Outlier

معیارهای دقت و بازخوانی به همراه یکدیگر افزایش می یابد و یکی قربانی دیگری نمی شود. معمولاً بیشتر روش های دسته بندی یا دارای دقت خوبی هستند یا بازخوانی خوبی دارند و برای افزایش یک معیار، معیار دیگر را قربانی می کنند اما این روش به همراه انتخاب خصیصه مذکور از این ایراد مبرا است [۱۶]. نتایج انتخاب خصیصه های مختلف در شکل ۳ آمده که از این میان خصایص شامل اضافه وزن، فشار خون و قند خون که خصیصه شماره ۲۲ می باشد با صحت ۰.۹۵۸ دارای بهترین نتیجه می باشد.



شکل ۲- شمایی از نتایج ارزیابی روش های مختلف دسته بندی



شکل ۳- شمایی از نتایج صحت 31 زیرمجموعه از خصایص

### ۱.۱. مقایسه با کارهای مشابه

در بررسی های انجام شده، تحقیقات بسیاری در زمینه افزایش دقت پیش بینی بیماری قلبی انجام شده است. همانطور که در بخش ۱۰ ذکر شد، این تحقیق، با

، C4.5 و بیز ساده استفاده شده است و یکی از نوآوری های این تحقیق تشخیص بهترین روش دسته بندی برای مجموعه داده بیماری های قلبی است.

### ۹. معیارهای ارزیابی

برای بررسی میزان اثربخشی روش C4.5، از معیارهای بازخوانی، دقت و صحت استفاده شده است. این معیارها با استفاده از نتایج اعمال این روش بر روی مجموعه آزمون، محاسبه می شوند. در رابطه های ۱ تا ۳ این سه معیار معرفی شده است. در این روابط  $TP^y$  به معنای تعداد نمونه ای است که سالم بوده اند و به درستی تشخیص داده شده اند.  $FP^A$  نشانگر تعداد نمونه هایی است که بیمار بوده اند و به اشتباه سالم تشخیص داده شده اند.  $FN^A$  نشانگر تعداد نمونه هایی است که سالم بوده اند ولی به اشتباه بیمار تشخیص داده شده اند.  $TN^A$  حاوی تعداد نمونه هایی است که بیمار بوده اند و به درستی بیمار تشخیص داده شده اند.

$$\text{دقت} = \frac{TP}{TP+FP} \quad (۱)$$

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

$$\text{صحت} = \frac{TP+TN}{TP+TN+FN+FP} \quad (۳)$$

### ۱۰. ارزیابی روش پیشنهادی

در این بخش، روش پیشنهادی با استفاده از مجموعه داده مراجعین بخش قلب بیمارستان امیرالمومنین تهران که در بخش ۳ معرفی شده، مورد ارزیابی قرار می گیرد. در گام اول این روش بر روی مجموعه داده های جمع آوری شده، فیلترهای معرفی شده در بخش ۶ اعمال شده و تعداد نمونه های این مجموعه داده از ۵۶۰ به ۵۳۳ نمونه تقلیل داده شده است. این روش در گام نخست ۲۷ نمونه از این مجموعه داده را به عنوان نویز یا داده پرت در نظر گرفته و آنها را حذف می کند. در مجموعه داده بدست آمده ۳۲۲ نمونه برای افراد سالم و ۲۰۱ داده دیگر برای افراد مبتلا به بیماری قلبی اعلام شده است. پس از حذف نویز و داده پرت، در این روش مجموعه های آموزش و آزمون تعیین می گردد (مراجعه به شکل ۱). مجموعه داده بیماری مورد استفاده برای آموزش شامل ۴۸۰ نمونه شامل افراد سالم و بیمار می باشد و مابقی به عنوان مجموعه داده آزمون (۵۳ نمونه) در نظر گرفته می شود. نسبت مورد نظر ۹۰٪ داده آموزش به ۱۰٪ داده آزمون است.

در انتها با ارزیابی روشهای دسته بندی اعمال شده روش دسته بندی C4.5 بهترین نتایج را بر طبق معیارهای دقت، بازخوانی، صحت و نرخ خطا بدست آورد. نتایج این ارزیابی در شکل ۲ قابل مشاهده است. همچنین مشاهده می شود این روش توانسته معیار صحت را نسبت به سایر روش های دسته بندی به میزان قابل توجهی بهبود بخشد.

از سوی دیگر با در نظر گرفتن انتخاب خصیصه و اندازه گیری سه فاکتور اضافه وزن، فشار خون و قند خون با کاهش هزینه و زمان، امر تشخیص بیماری را تسریع نمود که این روش باعث بهبود ۶.۲٪ نرخ تشخیص شد، و نکته قابل توجه دیگر در خصوص خصیصه انتخاب شده این است که در اینجا بر خلاف سایر روشها، مقادیر

<sup>6</sup> Support Vector Machine

<sup>7</sup> True Positive

<sup>8</sup> False Positive

<sup>9</sup> False Negative

<sup>10</sup> True Negative

جدول 1  
نتایج تحقیقات انجام شده مشابه

بیشترین میزان معیار صحت بدست آمده	الگوریتم مورد استفاده	نویسندگان
٪۸۲.۸	شبکه عصبی پرسپترون توسعه یافته	Godara و Panday [۵]، ۲۰۱۲
٪۸۱	بیز ساده	Subbalakshmi و همکاران [۱۳]، ۲۰۱۱
٪۸۶	SMO	Nahar و همکاران [۶]، ۲۰۱۲
٪۸۴.۱۲	SVM	Godara و Kumari [۷]، ۲۰۱۱
٪۸۳.۷۰	ODANB	Srinivas و همکاران [۸]، ۲۰۱۰
٪۶۳.۱۵	خوشه بندی فازی	El-Rashidy و همکاران [۹]، ۲۰۱۰
٪۶۳.۳۳	تصمیم گیری درختی	Rajalaxmi و Abdullah [۱۰]، ۲۰۱۲
٪۸۱.۵۱	دسته بندی انجمنی وزن دار	Soni و همکاران [۱۱]، ۲۰۱۱
٪۸۹.۵۶	شبکه های عصبی	Boonjing و Khemphila [۱۲]، ۲۰۱۱
٪۹۵.۸	انتخاب خصیصه - تصمیم گیری درختی	روش دسته بندی - انتخاب خصیصه

به عنوان یکی از کارهای آتی در این زمینه، توصیه می شود که برای افزایش دقت از ترکیب این روش با روش خوشه بندی نیز استفاده شود.

### تقدیر و تشکر

با تشکر از اساتید و پزشکانی که در هرچه بهتر شدن این مستند یاری ام نموده اند.

### REFERENCES

- Mohammadifard N, Sarafzadgan N, Sadri GH, Malekafzali H, Shahrokhi SH, Tolooie H, Poormoghaddas M, Rafiei M, Tavasoli AA, Kelishadi A, Rabiei K, Bashardoust NA, Asgari M, Naderi GH, Changiz T, Yousefi AR. Isfahan healthy heart program: The community-based interventions for prevention and control of cardiovascular disease. *Journal of Research in Medical Sciences*. 2002; 7(1):1-8.
- Polat K, Güneş S. A hybrid approach to medical decision support systems: Combining feature selection, fuzzy weighted pre-processing and AIRS. *Computer Methods and Programs in Biomedicine*. 2007; 88:164-174.
- Koh HC, Tan G. Data mining applications in healthcare. *Journal of Healthcare Information Management*. 2011;19:65.
- Crakowski MS. Health-related quality of life outcomes in clinical research. *Am J of Epidemiology*. 1999;283:215.
- Panday P, Godara N. Decision support system for cardiovascular heart disease diagnosis using improved multilayer perceptron. *International Journal of Computer Applications*. 2012;45:12-20.
- Nahar J, Imam T, Tickle KS, Chen YPP. Computational intelligence for heart disease diagnosis: A medical knowledge driven approach. *Expert Systems with Applications*. 2012.

استفاده از مجموعه داده های خاص برگرفته از اطلاعات مراجعین بخش قلب بیمارستان امیرالمومنین تهران مورد ارزیابی قرار گرفته است. در این بخش نتایج سایر تحقیقات انجام شده بر روی داده های مشابه به منظور مقایسه با روش این تحقیق و داده های موجود آورده شده است.

تا کنون از سه رویکرد شبکه های عصبی، دسته بندی و خوشه بندی در تشخیص بیماری قلبی استفاده شده است. بیشترین رویکردی که در تشخیص بیماری قلبی مورد استفاده قرار گرفته است به کارگیری روش های دسته بندی می باشد. در [۱۳]، از روش دسته بندی بیز ساده به این منظور استفاده شده است. همچنین Srinivas و همکاران [۸]، الگوریتم دسته بندی بیز را توسعه داده و الگوریتمی به نام ODANB را به منظور تشخیص بیماری قلبی ارائه داده اند. در [۶] و [۷] به ترتیب از دو روش دسته بندی SVM و SMO جهت تشخیص بیماری استفاده شده است.

در تحقیق دیگری [۱۰] که در زمینه پیش بینی بیماری قلبی بر روی یک سری داده انجام شده است از تصمیم گیری درختی استفاده شده است. استفاده از دسته بند انجمنی وزن دار نیز در [۱۱] مورد توجه قرار گرفته است.

Khemphila و همکارش [۱۲] یک شبکه عصبی برای تشخیص بیماری قلبی طراحی کرده اند و اخیراً در [۵] از شبکه عصبی پرسپترون توسعه یافته برای پیش بینی بیماری قلبی استفاده شده است. El-Rashidy و همکاران [۹] برای تشخیص بیماری قلبی از روش خوشه بندی فازی استفاده نموده اند. مقایسه نتایج این تحقیقات در جدول ۱ آورده شده است.

همانطور که در جدول ۱ مشاهده می شود، از بین رویکردهای مختلف رویکرد شبکه عصبی تا کنون بیشترین مقدار معیار صحت را داشته است و رویکرد خوشه بندی کمترین مقدار معیار صحت را بدست آورده است. روش انتخاب خصیصه - تصمیم گیری درختی C4.5 با ترکیب دو رویکرد انتخاب خصیصه و دسته بندی به روش تصمیم گیری درختی توانسته نسبت به سایر رویکردها، بخصوص رویکرد شبکه عصبی کارایی بیشتری را بدست آورد و معیار صحت را ۶.۲٪ بهبود بخشد.

### ۱۲. نتیجه گیری

تشخیص بیماری قلبی با استفاده از روش های یادگیری ماشین یکی از چالش های حوزه سلامت است. در این تحقیق کار بروی مجموعه داده ای خاص با بهره گیری از رویکرد انتخاب خصیصه دسته بندی ارائه شده است. گام های این روش عبارتند از: (۱) پیش پردازش (۲) انتخاب خصیصه و دسته بندی و (۳) ارزیابی. در روش مذکور ابتدا مجموعه آموزش به دو گروه افراد بیمار و سالم تقسیم می شود، سپس ۳۱ زیرمجموعه از کل مجموعه داده ها استخراج گردید که در مرحله سوم به ازای تمامی زیرمجموعه ها یک دسته بند یادگیری می شود. در این روش از چهار روش دسته بندی، SMO، C4.5 و بیز ساده استفاده شده است و نتایج ارزیابی های انجام شده در بخش ۴ این تحقیق، نشان می دهد که از بین روش های دسته بندی مورد استفاده، روش دسته بندی C4.5 دارای کارایی بیشتری می باشد که با انتخاب سه خصیصه فشار و قند و اضافه وزن جهت تشخیص بیماری و بکارگیری روش C4.5 ارزیابی و با سایر تحقیقات انجام شده مقایسه شده است. نتایج این ارزیابی نشان می دهد، که معیار صحت به میزان ۶.۲٪ بهبود یافته و به ۹۵.۸٪ رسیده است.

7. Kumari M, Godara S. Comparative study of data mining classification methods in cardiovascular disease prediction. *International Journal of Computer Science and Technology*. 2011;2:304-308.
8. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering*. 2010;2:250-255.
9. El-Rashidy MA, Taha TE, Ayad NMA, Sroor HS. An intelligent model for automated heart disease diagnosis. 2010.
10. Abdullah AS, Rajalaxmi R. A data mining model for predicting the coronary heart disease using random forest classifier. *International Conference in Recent Trends in Computational Methods, Communication and Controls*. 2012.
11. Soni J, Ansari U, Sharma D, Soni S. Intelligent and effective heart disease prediction system using weighted associative classifiers. *International Journal on Computer Science and Engineering*. 2011;3:2385-2392.
12. Khemphila A, Boonjing V. Heart disease classification using neural network and feature selection. *Systems Engineering (ICSEng)*. 2011; 406-409.
13. Subbalakshmi G, Ramesh K, Rao MC. Decision support in heart disease prediction system using naive bayes. *Indian Journal of Computer Science and Engineering (IJCSE)*. 2011;2:170-176.
14. [www.cs.waikato.ac.nz/ml/weka/index\\_downloading.html](http://www.cs.waikato.ac.nz/ml/weka/index_downloading.html) .
15. Sebastiani F. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*. 2002;34:1-47.
16. Zangooei MH, Jalili S. PSSP with dynamic weighted kernel fusion based on SVM-PHGS. *Knowledge-Based Systems*. 2011;22:424-442.
17. Haffner S, Taegtmeier H. Epidemic obesity and the metabolic syndrome. *Circulation*. 2003;108:1541-5.
18. Volek JS, Fernandez ML, Feinman RD, Phinney SD. Dietary carbohydrate restriction induces a unique metabolic state positively affecting atherogenic dyslipidemia, fatty acid partitioning, and metabolic syndrome. *Prog Lipid Res*. 2008;47:307-18.
19. Grundy SM, Brewer HB, Cleeman JI, Smith SC, Lenfant C. Definition of metabolic syndrome. *Arterioscler Thromb Vasc Biol*. 2004;24:13-18.
20. Ford ES. Risks for all-cause mortality, cardiovascular disease, and diabetes associated with the metabolic syndrome: a summary of the evidence. *Diabetes Care* 2005;28:1769-78.