مقاله پژوهشی

مجله انفورماتیک سلامت و زیست پزشکی
مرکز تحقیقات انفورماتیک پزشکی
دوره چهارم شماره چهارم. ص۳۰۴-۲۹۱، ۱۳۹۶

ارایه روشی جهت تشخیص سندرم متابولیک بر مبنای الگوریتم داده کاوی KNN
مطالعه موردی: بیمارستان شهدای کارگر یزد

جواد لعل دشتی[۱]، محسن محمدی[۲]*، فرهنگ پدیدران مقدم[۳]

• دریافت مقاله: ۹۶/۶/۲۸ • پذیرش مقاله: ۹۶/۱۲/۳

مقدمه: سندروم متابولیک به معنای وجود گروهی از عوامل خطرساز برای بروز بیماری‌های قلبی- عروقی و دیابت در یک شخص است. وجود علائم و ویژگی‌های مختلف این بیماری، تشخیص را برای پزشکان دشوار می‌کند. داده‌کاوی امکان تحلیل داده‌های بالینی بیماران برای تصمیم‌گیری‌های پزشکی را فراهم می‌کند. هدف این مقاله، ارائه یک مدل برای افزایش دقت پیش‌بینی سندرم متابولیک است.

روش: در این مطالعه کاربردی-توصیفی، پرونده پزشکی ۱۴۹۹ بیمار مبتلا به سندرم متابولیک با تعداد ۱۵ ویژگی مورد بررسی قرار گرفت. اطلاعات بیماران از پایگاه داده استاندارد بیمارستان فوق تخصصی شهدای گارگر یزد جمع‌آوری شد. هر یک از بیماران حداقل به مدت یک سال تحت پیگیری بودند. در این مقاله برای پیش‌بینی و تشخیص سندرم متابولیک، از الگوریتم کلونی زنبورعسل برای بهینه سازی نتایج الگوریتم داده‌کاوی KNN استفاده شد و یک مدل جدید ارائه گردید.

نتایج: بر اساس تابع هدف برای پیش‌بینی عارضه افزایش چربی خون از روش پیشنهادی، الگوریتم‌های گرگ خاکستری، ازدحام ذرات و ژنتیک برای بهبود عملکرد الگوریتم KNN استفاده شد. تحلیل‌های صورت گرفته نشان می‌دهد که مدل پیشنهادی با دقت پیش‌بینی ۰/۹۲۱ از روش‌های فازی، ماشین بردار پشتیبان، درخت‌تصمیم و شبکه عصبی دقت بیشتری دارد.

نتیجه‌گیری: جستجو در پایگاه داده‌های پزشکی برای رسیدن به دانش و اطلاعات جهت پیش‌بینی، تشخیص و تصمیم‌گیری از کاربردهای داده‌کاوی در پزشکی است. می‌توان از الگوریتم‌های وراثتی برای بهینه‌سازی تکنیک‌های داده‌کاوی استفاده کرد. پیش‌بینی و تشخیص صحیح سندرم متابولیک با استفاده از هوش مصنوعی و یادگیری ماشین، شانس درمان موفق را بالا می‌برد.

کلید واژه‌ها: سندرم متابولیک، الگوریتم کلونی زنبور عسل، درخت تصمیم

•**ارجاع:** ارایه روشی جهت تشخیص سندرم متابولیک بر مبنای الگوریتم داده کاوی KNN، مطالعه موردی: بیمارستان شهدای کارگر یزد. مجله انفورماتیک سلامت و زیست پزشکی ۱۳۹۶؛ ۴(۴): -.

۱. کارشناسی ارشد مهندسی کامپیوتر، موسسه آموزش عالی غیرانتفاعی اشراق، بجنورد، ایران
۲. دکتری فناوری اطلاعات، استادیار گروه کامپیوتر مجتمع آموزش عالی فنی مهندسی اسفراین، اسفراین، ایران
۳. دکتری فناوری اطلاعات، استادیار گروه کامپیوتر، موسسه آموزش عالی غیرانتفاعی اشراق، بجنورد، ایران
* **نویسنده مسئول:** گروه کامپیوتر، مجتمع آموزش عالی فنی مهندسی اسفراین
• **شماره تماس:** ۰۵۸۳۷۲۶۶۷۱۱
• **Email:**Mohsen@esfarayen.ac.ir

مقدمه

سندروم متابولیک به معنای وجود گروهی از عوامل خطرساز برای بروز بیماری‌های قلبی- عروقی و دیابت در یک شخص است. افراد دچار سندروم متابولیک در معرض خطر بیشتری برای ابتلا به بیماری‌های کرونری قلب و سایر بیماری‌های مربوط به تشکیل پلاک‌های چربی در دیواره سرخرگ‌ها مانند سکته مغزی و بیماری عروق محیطی هستند. تشخیص زود هنگام سندرم متابولیک شانس درمان موفقیت‌آمیز بیمار را بالا می‌برد [۱]. بهره‌گیری از راه‌کار داده‌کاوی با رویکرد استخراج دانش از اطلاعات موجود در راستای اهدافی چون نحوه تشخیص بیماری، کاهش هزینه‌های درمانی و میزان خطای نحوه درمان مؤثر بوده و موجب بهبود عملکرد سازمان‌های بهداشتی می‌شود [۲]. مطالعات مختلف نشان داده‌اند که وجود سندرم متابولیک تأثیر به سزایی روی فراوانی مرگ‌ومیر ناشی از بیماری‌های قلبی-عروقی دارد. سندرم متابولیک به عنوان مجموعه‌ای از اختلالات متابولیکی شامل پرفشاری خون، چاقی، اختلال لیپیدها و افزایش مقاومت به انسولین از جمله عوامل تأثیرگذار در بروز مرگ‌ومیر و ناتوانی ناشی از بیماری های قلبی-عروقی تلقی می‌شود [۲].

سندروم متابولیک با متابولیسم بدن و نیز با شرایطی تحت عنوان مقاومت به انسولین در ارتباط است. به صورت طبیعی، دستگاه گوارش و کبد غذاهایی را که می‌خورید به قند (گلوکز) تجزیه می‌کنند. خون، این قند را به بافت‌های بدن منتقل می کند، جایی که سلول‌ها از آن به عنوان سوخت استفاده می‌کنند. گلوکز به کمک انسولین وارد سلول‌های بدن می‌شود. در افراد مبتلا به مقاومت انسولین، سلول‌ها به صورت طبیعی به انسولین پاسخ نمی‌دهند و گلوکز به آسانی نمی‌تواند وارد سلول ها شود. در این حالت، بدن بیشتر انسولین ترشح می‌کند. در نتیجه مقادیر انسولین در خون بسیار بالا خواهد بود. در نهایت زمانی که بدن قادر به ساخت انسولین کافی برای کنترل گلوکز خون نیست، فرد به دیابت مبتلا خواهد شد. حتی چنان چه سطوح قند خون به حدی زیاد نباشد که بتوان آن را دیابت در نظر گرفت، باز سطوح افزایش یافته گلوکز مضر هستند. پزشکان به این شرایط، مرحله پیش از دیابت می‌گویند. انسولین افزایش یافته، سطح تری گلیسرید و دیگر چربی‌های خون را نیز افزایش می‌دهد. این شرایط با چگونگی عملکرد کلیه‌ها نیز تداخل می‌کنند و موجب افزایش فشار خون می‌شوند. این اثرات، فرد را در معرض خطر ابتلا به بیماری قلبی، سکته، دیابت و دیگر بیماری‌ها قرار می‌دهد. برخی از افراد به صورت ژنتیکی مستعد مقاومت به انسولین هستند و این استعداد را از والدین خود به ارث می‌برند. از عوامل محیطی، چاق بودن و بی تحرک بودن از عوامل اصلی ابتلا به سندرم متابولیک هستند [۳].

Karegowda و همکاران [۴] ترکیبی از الگوریتم ژنتیک و شبکه‌های انتشار خطا را به‌منظور انجام پیش‌بینی برای داده‌های پزشکی ارائه داده‌اند که در آن الگوریتم ژنتیک می تواند برای بهبود عملکرد الگوریتم شبکه‌های انتشار خطا استفاده شود. این مطالعه، کاربرد الگوریتم ژنتیک را برای مقداردهی اولیه و بهینه‌سازی وزن‌های شبکه‌های انتشار خطا نشان می‌دهد. اولین مرحله الگوریتم ژنتیک، نمایش کروموزوم‌ها است. برای شبکه‌های انتشار خطا با یک لایه پنهان و m گره شامل n گره ورودی و p گره خروجی، تعداد وزن‌ها برابر است با .(n+p)*m هر کروموزوم از (n+p)*m ژن ساخته می‌شود. در این مطالعه، تابع برازندگی هر کروموزوم در الگوریتم ژنتیک با استفاده از روش بهینه‌سازی مینیمم محاسبه می‌شود. تابع برازندگی به‌صورت Fitness (Ci)=1/E است که در آن E خطای محاسبه شده به‌عنوان خطای مربع میانگین ریشه در لایه خروجی می‌باشد. عملگر برش مورد استفاده، برش یک نقطه‌ای، دو نقطه‌ای و چند نقطه‌ای می‌باشد. پس از انجام عملگر جهش و تولید داده‌ها، جمعیت جدید به‌عنوان ورودی به شبکه‌های انتشار خطا داده می‌شود تا برازندگی هر کروموزوم را محاسبه کند و به‌دنبال آن فرایندهای انتخاب، تولید مجدد، برش و جهش برای جمعیت جدید انجام می‌شود. این روند تا رسیدن به کروموزوم‌های هم‌گرا ادامه می‌یابد. این کروموزوم‌های هم‌گرای به‌دست آمده، وزن‌های ارتباطی بهینه برای شبکه‌های انتشار خطا هستند. داده مورد استفاده برای این مدل PID است. در این مطالعه، مشخصه‌های قابل توجهی با استفاده از دو روش شناسایی شدند. روش درخت تصمیم و روش GA-CFS به‌عنوان ورودی مدل ترکیبی برای تشخیص مرض قند استفاده شدند. نتایج حاصل حاکی از آن است که رویکرد الگوریتم ژنتیک بهینه‌سازی‌شده شبکه‌های انتشار خطا نسبت به حالت بدون بهینه‌سازی الگوریتم ژنتیک، عملکرد بهتری را نشان داده است.

Shen و همکاران [۵] یک طرح جدید برای تنظیم پارامترهای ماشین‌های بردار پشتیبان پیشنهاد داده‌اند که در آن از الگوریتم بهینه‌سازی مگس میوه استفاده می‌شود. این طرح ماشین بردار پشتیبان مبتنی بر بهینه‌سازی مگس میوه نام دارد که با موفقیت در تشخیص پزشکی مورد استفاده قرار گرفته

است. رویکرد جدید این مطالعه در ارائه روش مبتنی بر بهینه‌سازی مگس میوه نهفته است که هدف آن بالابردن قابلیت تعمیم طبقه‌بندی‌کننده ماشین بردار پشتیبان از طریق بررسی تکنیک هوش ازدحامی جدید برای تنظیم پارامتر بهینه به‌منظور طبقه‌بندی داده‌های پزشکی است. در روش پیشنهادی، تکنیک بهینه‌سازی مگس میوه به‌طور مؤثر و کارآمد به پارامترهای موجود در ماشین بردار پشتیبان می‌پردازد. مدل ارائه شده عمدتاً شامل دو روش است: بهینه‌سازی پارامتر درونی و ارزیابی عملکرد طبقه‌بندی بیرونی. در طول رویه بهینه‌سازی پارامتر درونی، پارامترهای ماشین بردار پشتیبان به‌صورت پویا از طریق تجزیه و تحلیل اعتبارسنجی متقاطع ۵ لایه‌ای توسط بهینه‌سازی مگس میوه تنظیم می‌شوند. سپس پارامترهای بهینه به‌دست آمده به مدل پیش‌بینی ماشین بردار پشتیبان تغذیه می‌شوند تا با استفاده از تجزیه و تحلیل اعتبارسنجی متقاطع ده لایه‌ای، این کار طبقه‌بندی را برای تشخیص در حلقه بیرونی اجرا کند. گذشته از آن، کارایی و اثر بخشی ماشین بردار پشتیبان مبتنی بر بهینه‌سازی مگس میوه در مقایسه با چهار مجموعه داده پزشکی شناخته شده و معروف شامل: مجموعه داده‌های سرطان پستان ویسکانسین، مجموعه داده‌های بیماری PID، مجموعه داده‌های بیماری پارکینسون و مجموعه داده‌های بیماری تیروئید به لحاظ دقت طبقه‌بندی، میزان حساسیت، ویژگی و مدت زمان پردازش به‌دقت مورد ارزیابی قرار گرفته است. از چهار همتای رقابتی از جمله ماشین بردار پشتیبان مبتنی بر الگوریتم بهینه‌سازی ازدحام ذرات، ماشین بردار پشتیبان مبتنی بر الگوریتم ژنتیک، ماشین بردار پشتیبان مبتنی بر بهینه‌سازی تغذیه باکتریایی و ماشین بردار پشتیبان مبتنی بر تکنیک جستجوی شبکه برای اهداف مقایسه‌ای استفاده شد. نتایج تجربی این مطالعه نشان داد که روش ماشین بردار پشتیبان مبتنی بر بهینه‌سازی مگس میوه پیشنهاد شده می‌تواند پارامترهای الگوی بسیار مناسب‌تری را به‌دست بیاورد و به طور قابل توجهی زمان محاسبه را کاهش دهد که در نهایت به‌دقت بالا در طبقه‌بندی منجر می‌شود. در این روش پیشنهادی امید بر آن است که بتوان از آن به‌عنوان یک راه‌حل جایگزین و یک ابزار مفید بالینی در تصمیم‌گیری های پزشکی استفاده نمود.

روش پیشنهادی برای کمک به تشخیص و پیش‌بینی بیماری های داخلی از طریق تکنیک‌های داده‌کاوی، دقیق‌تر از روش‌های موجود، به عنوان سیستم پشتیبان تصمیم‌گیری در سیستم‌های پزشکی کاربرد خواهد داشت. در این پژوهش سندرم متابولیک مورد بررسی قرار می‌گیرد. هدف تحقیق ارائه الگوریتمی جهت بهینه‌سازی تشخیص سندرم متابولیک بر مبنای الگوریتم K نزدیکترین همسایه (KNN) است به طوری که دقت تشخیص و پیش‌بینی سندرم متابولیک را افزایش دهد.

روش

این مطالعه از نوع کاربردی-توصیفی است. برای تهیه مجموعه داده‌های مربوط به سندرم متابولیک، در ابتدا پرسشنامه استاندارد بر اساس ویژگی‌های جدید طراحی شد و در بیمارستان شهدای کارگر یزد تکمیل و جمع‌آوری گردید. داده‌ها مربوط به سال‌های ۱۳۹۳ تا ۱۳۹۴ است. این مجموعه داده شامل ۱۴۹۹ نمونه بوده که تعداد ۵۲ نمونه فاقد اطلاعات کامل می‌باشد در این مقاله بعضی از اطلاعات موجود در پرونده مانند نام، نام‌خانوادگی، شماره پرونده بیمار و نشانی حذف شدند. در مرحله بعدی، پرونده بیمارانی را که فقط یک بار مراجعه داشته‌اند، کنار گذاشته شد، زیرا اطلاعات کامل از آزمایش‌ها و عوارض آن‌ها در دسترس نبود. مواردی که ارزش صفر برای ویژگی‌های فشار خون، قند خون ناشتا، قند خون ۲ ساعت بعد از غذا، تری‌گلیسیرید داشتند، به دلیل اهمیتی که این ویژگی‌ها در نتیجه نهایی دارند، حذف شدند. در تحقیق Harper و همکاران [۷] ثابت شد که حذف عاقلانه، یک روش کارآمد به‌جای جایگزین کردن ویژگی‌های پر اهمیت با تکنیک‌هایی مانند میانگین، انتساب تصادفی، انتساب رگرسیون و مدل‌های بیزی است. نمونه‌هایی که دارای چند ویژگی مفقوده بودند نیز حذف گردیدند و سایر مقادیر مفقوده با استفاده از مقادیر شایع و پر احتمال مقداردهی شدند. برای هر بیمار تعداد ۱۵ ویژگی ثبت شد. بازه مقادیر اولیه ویژگی‌های بالینی بیمار در جدول ۱ نمایش داده شد. قالب مناسب داده‌ها به عنوان ورودی داده کاوی در نتایج و خروجی تأثیرگذار است.

اگر مقادیر ویژگی‌های مجموعه داده در دامنه متفاوتی قرار داشته باشند، احتمال بروز خطا در یافته‌ها افزایش می‌یابد. به قرار دادن داده‌های یک جامعه آماری در دامنه مشابه، نرمال سازی گفته می‌شود [۷]. در مدل پیشنهادی نحوه نرمال‌سازی به روش Max/Min و در بازه [۱-۰] است [۸].

جدول ۱: نحوه نرمال سازی مجموعه داده‌ها

	ویژگی	بازه مقادیر اولیه	نحوه نرمال سازی
۱	سن	[۶۵ و ۱۸]	Max/Min
۲	تحصیلات	[۱۹۵ و۱۵۰]	Max/Min
۳	BMI	۱- دچار کمبود وزن شدید ۲- کمبود وزن ۳- عادی ۴- اضافه وزن ۵- چاقی کلاس ۱ ۶- چاقی کلاس ۲ ۷- چاقی کلاس ۳	پس از مرتب سازی از روش Max Min نرمال می‌شود.
۴	تری‌گلیسرید	۱- نرمال ۲- نزدیک به حد نرمال ۳-سطح خطرناک ۴- بسیار پر خطر	Max Min
۵	فشار خون بالا	۱- افت فشار خون ۲- نرمال ۳-در معرض فشار خون بالا ۴- فشار خون بالا	پس از مرتب سازی از روش Max Min نرمال می‌شود.
۶	سابقه سندرم در خانواده	۱- ندارد ۲- دارد	Max Min
۷	قند ناشتا	۱- طبیعی ۲- پیش دیابتی ۳- دیابتی	پس از مرتب سازی از روشMax Min نرمال می‌شود.
۸	قند دو ساعت بعد از غذا	۱- طبیعی ۲- پیش دیابتی ۳- دیابتی	پس از مرتب سازی از روش Max Min نرمال می‌شود.
۹	LDL	۱- مقدار مطلوب ۲- مقادیر نزدیک به حد مطلوب ۳- سطح در مرز شروع خطر ۴- سطح خطرناک ۵- سطح بسیار پر خطر	پس از مرتب سازی از روش Max Min نرمال می‌شود.
۱۰	HDL	۱- بالا ۲- پایین	Max Min
۱۱	جنسیت	۱- مرد ۲- زن	Max Min
۱۲	میزان فعالیت بدنی	۱-بیشتر از سه روز در هفته ۲-کمتر از سه روز در هفته ۳-بیشتر از سی دقیقه در روز ۴-روزانه	Max Min
۱۳	مصرف سیگار	۱- ندارد ۲- دارد	Max Min
۱۴	مصرف دارو	۱- دارو مصرف نمی کند ۲- دارو مصرف می کند	Max Min
۱۵	برچسب	۱- طبیعی ۲- دارای اختلالات سندرم متابولیک	نرمال‌سازی نمی‌شوند

در حل مسئله مقادیر گمشده هر نمونه، در صورتی که نمونه مذکور، دارای مقادیر گمشده زیادی باشد، آن نمونه از دیتاست حذف شده و در غیر این صورت مقادیر گمشده با بیشترین فراوانی مشخصه مذکور تکمیل می‌شوند. در یک دیتاست، مواردی پیش می‌آید که مقادیر گمشده در حوزه‌های یک رکورد وجود دارد. داده‌های موجود در یک دیتاست، زمانی که وارد الگوریتم می‌شوند باید کامل و بدون مقادیر گمشده یا داده‌های صدمه دیده باشند. همچنین مواردی که مقادیر احتمالاً غلط به ویژگی‌های یک رکورد تخصیص یافته باشد، باید تصحیح و در صورت عدم تصحیح از دیتاست حذف گردد. این مشکل بیشتر ناشی از عدم یکپارچگی قسمت‌های جمع‌آوری کننده داده‌ها، منابع داده و شکل‌های جمع‌آوری داده‌ها است؛ بنابراین وجود یک دیتاست کامل و خالی از مقادیر Miss Value در کشف روابط پنهانی بین داده‌های موجود در دیتاست، نقشی بسیار مهم و حیاتی دارد. با توجه به اینکه نتایج حاصل از تحلیل داده‌های ناقص می‌تواند به سوگرائی منجر شود؛ لذا اهمیت دارد که تحلیل این نوع داده ها در مسیری مناسب و صحیح قرار داده شود.

هر یک از ویژگی‌ها و یافته‌ها در تشخیص و پیش‌بینی سندرم متابولیک از اهمیت خاصی برخوردار هستند. به بیان دیگر همه ویژگی‌ها دارای ارزش یکسان نیستند. به عنوان مثال در تشخیص بیماری دو ویژگی BMI و وجود قند خون دارای

اهمیت متفاوتی هستند. این که هر یک از ویژگی‌ها دارای چه ارزشی است و چقدر در تشخیص بیماری نقش دارد، مسئله مهمی است. در این مقاله روشی ارائه می‌شود که ارزش و نقش هر یک از ویژگی‌ها به طور دقیق مشخص شده و بیماری تشخیص داده می‌شود.

در روش پیشنهادی برای مشخص کردن ارزش و نقش هر یک از ویژگی‌ها در تشخیص بیماری از الگوریتم‌های فراابتکاری برای بهبود عملکرد الگوریتم KNN استفاده می شود. در جدول ۲ نحوه وزن‌دار کردن ویژگی‌ها برای بهبود KNN نشان داده شد. در هر مرحله از اجرا، ویژگی‌های وزن دار به KNN داده می شود و بر اساس دقت به دست آمده، با استفاده از الگوریتم‌های تکاملی، مقدار بهینه به دست می آید.

جدول ۲: وزن‌دار کردن ویژگی‌ها

ویژگی	مقدار اولیه	مقدار نرمال	وزن	مقدار وزن‌دار
سن	۵۵	۰/۷۱	۰/۳	۰/۲۱
قد	۱۵۶	۰/۵۵	۰/۸	۰/۴۴
BMI	۳۴/۹	۰/۹	۰/۷	۰/۶۳

طبقه‌بند نزدیکترین همسایه (K Nearest KNN(Neighbor Classifier

طبقه‌بند k–نزدیکترین همسایه یک طبقه‌بند یادگیری با ناظر می‌باشد که نمونه‌ها را بر اساس شباهت و فاصله از نمونه‌های آموزشی طبقه‌بندی می‌کند. در اغلب موارد برای دسته‌بندی به‌کار می‌رود، هر چند که می‌توان از آن برای تخمین و پیش‌بینی نیز استفاده نمود. برای یک داده آزمایشی الگوریتم به‌دنبال K نمونه از نزدیک‌ترین نمونه‌ها می‌گردد (K نمونه مشابه).

نزدیکی دو نمونه، با به‌دست آوردن تشابه و یا فاصله میان این دو نمونه محاسبه می‌شود. هر نمونه می‌تواند از انواع داده‌ها تشکیل شده باشد که باید تشابه میان آن‌ها بررسی شود. پس از یافتن این K داده مشابه با نمونه آزمایشی، رأی اکثریت تعیین‌کننده برچسب کلاس داده آزمایشی می‌باشد. چنانچه مقدار ۱ برای K تنظیم شود، در این‌صورت کلاس نزدیک‌ترین داده به نمونه آزمایشی، به‌عنوان کلاس تخمینی ارائه می‌شود؛ اما به‌دلیل وجود داده‌های نویز و خارج از محدوده، مقدار ۱، عدد مناسبی برای K نیست. می‌توان مقدار مناسب را به‌صورت تجربی به‌دست آورد. برای مثال با ۱ شروع و برای مجموعه‌ای آزمایشی نرخ خطا را محاسبه کرد. با افزایش مقدار K این کار را تکرار می‌کنیم. مقداری از K که باعث حداقل نرخ خطا می‌شود، انتخاب مناسبی است.

روش k–نزدیکترین همسایه بار محاسباتی زیادی دارد، زیرا زمان محاسباتی به صورت نمایی از تمام نقاط افزایش می‌یابد، ولی دقت بالایی دارد [۹].

از سال ۱۹۷۰، روش انشعاب و تحدید برروی مسئله KNN اعمال شده است. در فضای اقلیدسی، این کار با نام شاخص‌های مکانی یا روش دسترسی مکانی شناخته می‌شود. روش‌های متعددی برای پارتیشن‌بندی فضا به منظور حل مسئله KNN ارائه و توسعه داده شده‌اند. در این بین، ساده ترین روش، استفاده از ساختمان داده k-d tree است که به صورت تکراری فضای جستجو را در یکی از ابعاد، به دو نیم‌بخش تقسیم می‌کند به گونه‌ای که هر نیم‌بخش –تقریباً– نیمی از نقاط بخشِ بزرگ‌تر را در برخواهد گرفت. پرس و جو با پیمایش درخت از ریشه تا برگ صورت می‌پذیرد به گونه‌ای که در هر مرحله، مقدار بردار گره در بُعد مشخصی، با همان بُعد از نقطه پرس و جو، مقایسه می‌شود؛ اگر مقدار نقطه پرس وجو بیشتر بود به زیر درخت سمت راست و در غیر این‌صورت به زیردرخت چپ حرکت می‌کنیم. همچنین بسته به اینکه فاصله این دو نقطه چقدر است، شاخه‌های کناری نیز ممکن است نیاز به بررسی داشته باشند. در مورد نقاط راندم توزیع شده تحلیل پیچیدگی زمانی مربوط به بدترین حالت صورت می‌گیرد. روش دیگر، استفاده از ساختمان داده درخت مستطیلی است که برای استفاده KNN در زمینه‌های پویا

به کار گرفته می‌شود. این ساختمان داده دارای الگوریتم‌های مؤثری برای درج و حذف از درخت می‌باشد.

ایجاد مدل تشخیص سندرم متابولیک توسط الگوریتم‌های تکاملی

الگوریتم‌های تکاملی زیر مجموعه‌ای از محاسبات تکاملی است و در شاخه هوش مصنوعی قرار می‌گیرد و شامل الگوریتم‌هایی جهت جستجو است که در آن‌ها عمل جستجو از چندین نقطه در فضای جواب آغاز می‌شود. الگوریتم‌های تکاملی به طور اساسی با دیگر روش‌های بهینه‌سازی و جستجوی مرسوم قدیمی تفاوت دارند. برخی از این تفاوت‌ها عبارت‌اند از:

۱- الگوریتم‌های فرگشت‌پذیر تنها یک تک نقطه را جستجو نمی‌کنند، بلکه جمعیتی از نقاط را به صورت موازی بررسی می‌نمایند.

۲- الگوریتم‌های فرگشت‌پذیر نیاز به اطلاعاتی ضمنی و دیگر دانش‌های مکمل ندارند؛ تنها تابع هدف و شایستگی مربوطه در جهت‌های جستجو تأثیر گذارند.

۳- الگوریتم‌های فرگشت‌پذیر از قوانین در حال تغییر احتمالی بهره می‌برند و نه موارد مشخص و معین.

۴- استفاده از الگوریتم‌های فرگشت‌پذیر به طور کلی خیلی سر راست است، زیرا هیچ‌گونه محدودیت‌هایی برای تعریف تابع هدف وجود ندارد.

۵- الگوریتم‌های فرگشت‌پذیر تعداد زیادی از پاسخ‌های قابل قبول را به دست می‌آورند و انتخاب پایانی بر عهده کاربر است؛ الگوریتم‌های تکاملی برای شناسایی این پاسخ‌های چندگانه به طور همزمان ذاتاً کارآمدند.

روند کلی الگوریتم‌های تکاملی برای انتخاب ویژگی‌های وزن دار در جهت تشخیص سندرم متابولیک به صورت زیر است:

۱-ایجاد جمعیت اولیه

جمعیت اولیه به تعداد n

اختصاص وزن به هر یک ویژگی‌ها به طور تصادفی در بازه [0 1] برای هر یک از اعضای جمعیت

۲-یافتن مقدار تابع هزینه (شکل ۱)

- Objective Function = Minimum (Error Rate)
- Error Rate = 1- Accuracy

۳-بقای بهترین‌ها و تولید نسل جدید و حذف ضعیف‌ترها با توجه به تابع هزینه.

۴-رفتن به مرحله ۲ و انجام مجدد کلیه روند تا زمانی که شرط توقف اعمال شود:

- رسیدن به تعداد ۱۰۰۰ تکرار
- رسیدن به هزینه حداقل

ایجاد مدل تشخیص سندرم متابولیک مبتنی بر الگوریتم GBC

الگوریتم GBC(Genetic Bee Colony) رویکرد جدیدی است که توان افزایش دقت و کاهش خطا در پیش بینی را با محاسبات حداقل و مرتبه زمانی کمتر از کارهای مشابه دارد [۱۰]. روال بهینه‌یابی در الگوریتم GBC براساس یک روند تصادفی- هدایت شده استوار می‌باشد. این روش، بر مبنای نظریه تکامل تدریجی و ایده‌های الگوریتم زنبورعسل پایه‌گذاری شده است. الگوریتم زنبور عسل هر نقطه را در فضای پارامتری متشکل از پاسخ‌های ممکن به عنوان منبع غذا تحت بررسی قرار می‌دهد. زنبورهای دیده‌بان – کارگزاران شبیه‌سازی شده – به صورت تصادفی فضای پاسخ‌ها را ساده می‌کنند و به وسیله تابع شایستگی کیفیت موقعیت‌های بازدید شده را گزارش می‌دهند. جواب‌های ساده شده رتبه‌بندی می‌شوند و دیگر زنبورها نیروهای تازه‌ای هستند که فضای پاسخ‌ها را در پیرامون خود برای یافتن بالاترین رتبه محل‌ها جستجو می‌کنند که گلزار نامیده می‌شود. الگوریتم به صورت گزینشی دیگر گلزارها را برای یافتن نقطه بیشینه تابع شایستگی جستجو می‌کند.

در نهایت، به منظور دستیابی به یک تعادل بین بهره‌برداری و اکتشاف در الگوریتم کلونی رنبور عسل (ABC) و بهبود توانایی‌های جستجوی محلی، از اپراتورهای جهش الگوریتم ژنتیک در طول فرآیند جایگزینی راه‌حل‌ها استفاده می‌شود. الگوریتم GBC با دادن جمعیت تصادفی اولیه از فضای جستجو یعنی ویژگی‌های استخراج شده شروع می‌شود:

$$X_{mi} = X_i^{min} + rand * (X_i^{max} - X_i^{min}) \quad (۱)$$

که در رابطه بالا X_{mi} یک بردار راه حل (ویژگی‌های انتخاب شده) برای مسئله بهینه‌سازی و $i=1,\ldots,n$ و $m=1,\ldots,SN$ می‌باشد. SN بیانگر تعداد جمعیت اولیه و هر X_i یک بردار n بعدی است.

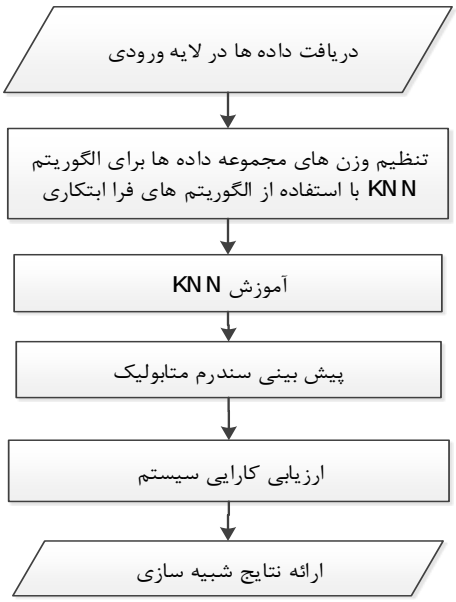
در الگوریتم GBC برای دستیابی به بهره‌برداری و اکتشاف متعادل از مزایای الگوریتم ABC و GA استفاده شده است.

در GBC عملگرهای GA در فرآیند بهره‌برداری روش ABC در مرحله زنبورعسل ناظر برای بهبود به اشتراک گذاری اطلاعات بین زنبورعسل کارگر و زنبور عسل ناظر برای یافتن راه‌حل بهینه استفاده شده است. شکل ۲ روند نمای الگوریتم پیشنهادی را نشان می‌دهد.

ALGORITHM 1

1) Read the training data from a file
2) Read the testing data from a file
3) Normalize the attribute values in the range of 0 to 1.
4) Let $x_1,x_2....x_m$ denote the m instances from data set $\{f_1, f_2, \ldots, f_n\}$, n = Number of features
5) Assign weight w_i to each instance x_i in the training set
6) Predict the class value from weighted dataset
7) Calculate the error rate as
*Error Rate = 1- (# of correctly classified examples / All) * 100*
8) Fitness Function=Minimize (Error Rate)

شکل ۱: تابع هزینه در الگوریتم پیشنهادی



شکل ۲: روندنمای رهیافت پیشنهادی

نتایج

در جدول ۳ مقایسه رهیافت پیشنهادی با الگوریتم‌های فازی، شبکه عصبی، درخت تصمیم برای پیش‌بینی سندرم متابولیک نمایش داده شد. ارزیابی مقادیر جدول بر حسب معیار Sensitivity بیان کننده برتری کلاسه‌بندی روش پیشنهادی با ویژگی‌های وزن‌دار نسبت به سایر روش‌ها در حالت ویژگی‌های با وزن یکسان است.

جدول ۳: مقایسه نتایج با معیار Sensitivity

	Decision Tree	SVM	MLP-NN	Fuzzy	KNN & GBC
وجود سندرم	۰/۹۵	۰/۹۸	۰/۹۵	۰/۹۵	۰/۹۸
عدم وجود سندرم	۰/۷۲	۰/۵۵	۰/۵۱	۰/۷۰	۰/۸۰

ارزیابی پیش‌بینی سندرم متابولیک بر اساس معیارهای Specificity در جدول ۴ مشخص کننده برتری کلاسه بندی روش پیشنهادی نسبت به روش‌های فازی، شبکه عصبی، درخت تصمیم است. مقایسه نتایج جدول بهبود کارایی کلاسه‌بندی با ویژگی‌های وزن‌دار نسبت به مدل سازی ویژگی‌ها با وزن یکسان را نشان می‌دهد.

جدول ۴: مقایسه نتایج با معیار Specificity

	Decision Tree	SVM	MLP-NN	Fuzzy	KNN & GBC
وجود سندرم	۰/۷۲	۰/۵۵	۰/۵۱	۰/۷۰	۰/۸۰
عدم وجود سندرم	۰/۹۵	۰/۹۸	۰/۹۵	۰/۹۵	۰/۹۸

معیار ارزیابی Precision برای مقایسه کارایی کلاسه بندی با ویژگی‌های وزن‌دار در مقابل مدل‌سازی ویژگی‌ها با وزن یکسان در جهت پیش‌بینی وجود یا عدم وجود عارضه سندرم متابولیک در جدول ۵ نشان داده شد. رهیافت پیشنهادی در مقایسه با روش‌های فازی، شبکه عصبی، درخت تصمیم نتایج بهتری در تشخیص بیماری را دارد.

جدول ۵: مقایسه نتایج با معیار Precision

	Decision Tree	SVM	MLP-NN	Fuzzy	KNN & GBC
وجود سندرم	۰/۹۱	۰/۸۸	۰/۸۶	۰/۹۱	**۰/۹٤**
عدم وجود سندرم	۰/۸۱	۰/۹۱	۰/۷۶	۰/۸۱	**۰/۹۱**

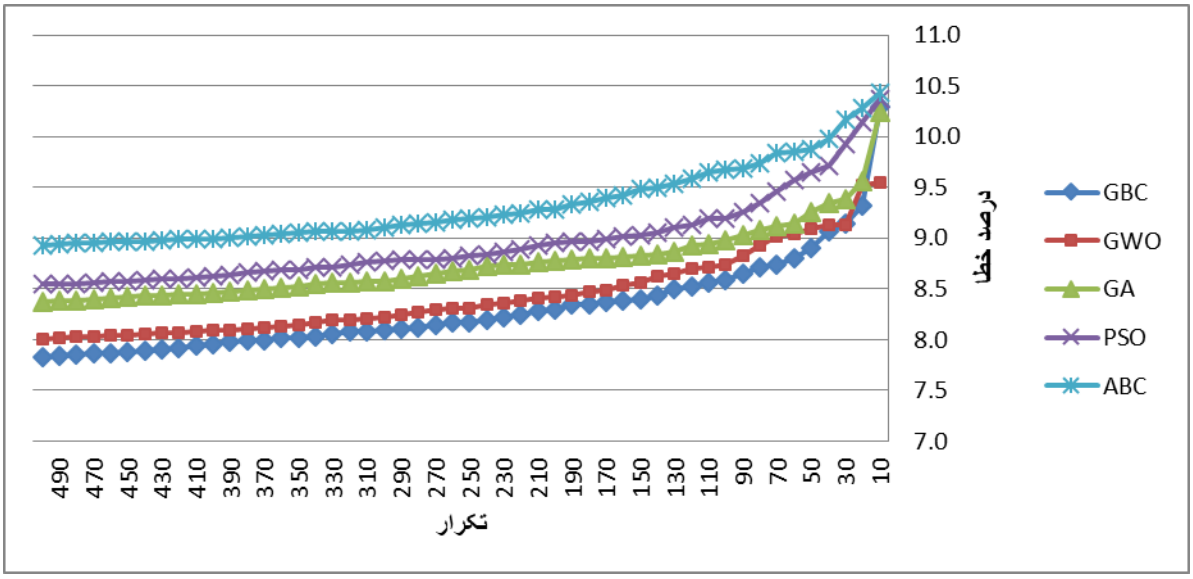
بر حسب معیار F-measure رهیافت پیشنهادی برای پیش‌بینی وجود یا عدم وجود عارضه سندرم متابولیک نسبت به سایر روش‌های مورد مقایسه کارایی بهتری دارد (جدول ۶). مقایسه نتایج روش پیشنهادی بر اساس این معیار نشان دهنده برتری روش پیشنهادی نسبت به سایر روش‌های مورد مقایسه می‌باشد. این ارزیابی نشان دهنده کارایی کلاسه‌بندی با ویژگی‌های وزن‌دار نسبت به مدل‌سازی ویژگی‌ها با وزن یکسان است.

جدول ۶: مقایسه نتایج با معیار F-Measure

	Decision Tree	SVM	MLP-NN	Fuzzy	KNN & GBC
وجود سندرم	۰/۹۳	۰۰/۹۳	۰/۹۰	۰/۹۳	۰/۹۶
عدم وجود سندرم	۰/۷۶	۰/۶۹	۰/۶۱	۰/۷۵	۰/۸۵

در شکل ۳ محور عمودی نشان‌دهنده درصد خطای پیش بینی عارضه و محور افقی نشان‌دهنده تعداد تکرارهای اجرای الگوریتم می‌باشد. در تکرارهای اول چون جمعیت اولیه به صورت تصادفی است کاهش خطا ملموس می‌باشد؛ اما در تکرارهای بعدی آهنگ کاهش خطا کمتر می‌شود و در نهایت روش پیشنهادی در انتهای شبیه‌سازی به کاهش خطای بهتری دست یافت. در شکل ۳ مقایسه مقادیر خطا برای پیش‌بینی سندرم متابولیک با استفاده بهینه‌سازی وزن ویژگی ها از طریق پردازش تکاملی در ۵۰۰ تکرار نمایش داده شده است. نتایج شکل ۴ نشان دهنده برتری روش GBC نسبت

به الگوریتم‌های گرگ خاکستری، PSO، GA و ABC است.



شکل ۳: درصد خطا برای پیش‌بینی سندرم متابولیک در مدل پیشنهادی

مدل پیشنهادی با سه روش فازی [۱۲]، درخت‌تصمیم [۱۳] ، ماشین بردار پشتیبان و شبکه عصبی پرسپترون چند لایه (MLP) [۱۴] مقایسه شد. ارتباط بین کلاس‌های واقعی و کلاس‌های پیش‌بینی شده با استفاده از ماتریس Confusion قابل محاسبه است. در شکل ۴، پارامترهای موردنیاز ماتریس Confusion ذکرشده است.

برای مقایسه مدل پیشنهادی با سایر روش‌ها از معیارهای Accuracy، Sensitivity، Specificity، Precision و F-Measure با توجه به شکل ۴ طبق روابط زیر استفاده می‌شود [۱۵]:

$$Accuracy = (TP + TN) / All \quad (۲)$$

$$Sensitivity = TP / (TP + FN) \quad (۳)$$

$$Specificity = TN / (FP + TN) \quad (۴)$$

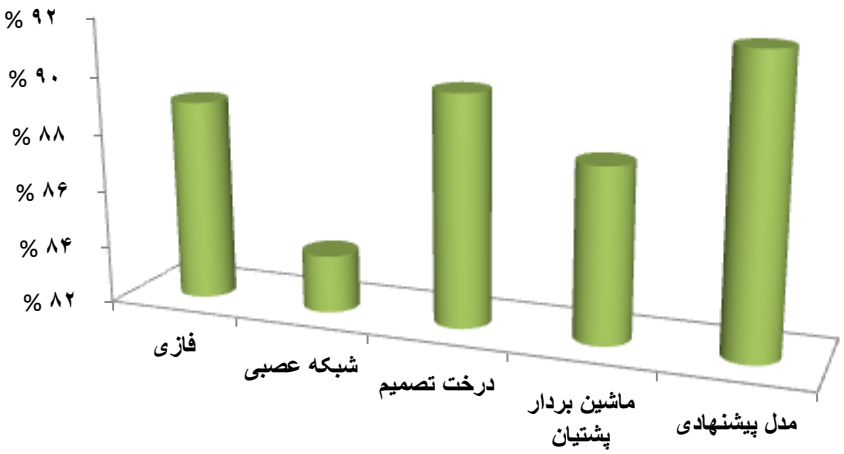
$$Precision = TP / (TP + FP) \quad (۵)$$

$$Recall = TP / (TP + FN) \quad (۶)$$

$$F\ Measure = \frac{2 * Precision * Recall}{Precision + Recall} \quad (۷)$$

TP: تعداد رکوردهایی که به درستی، مثبت تشخیص داده می‌شوند.
TN: تعداد رکوردهایی که به درستی، منفی تشخیص داده می‌شوند.
FP: تعداد رکوردهایی که به غلط، مثبت تشخیص داده می‌شوند.
FN: تعداد رکوردهایی که به غلط، منفی تشخیص داده می‌شوند.

شکل ۴ نشان دهنده نمودار نتایج تشخیص روش‌های مختلف با معیار Accuracy است. همان‌گونه که مشاهده می‌شود مدل پیشنهادی دقت بیشتری نسبت به سایر روش‌ها دارد. همچنین به ترتیب در جدول ۳ تا ۶ مقایسه نتایج پیش بینی بیماری با معیارهای Sensitivity، Specificity، Precision و F-Measure مورد مقایسه قرار گرفت. نتایج مقایسه نشان دهنده برتری عملکرد مدل پیشنهادی می‌باشد. مقادیر جدول نشان دهنده عملکرد بهتر روش پیشنهادی است.



	فازی	شبکه عصبی	درخت تصمیم	ماشین بردار پشتیان	مدل پیشنهادی
دقت	% ۸۹	% ۸۴	% ۹۰	% ۸۸	% ۹۲

شکل ۴: نمودار نتایج با معیار Accuracy

بحث

در این مطالعه، مجموعه‌ای از الگوهای کمتر شناخته شده و مؤثر در بروز سندرم متابولیک بر اساس یک پایگاه داده بومی مورد پردازش قرار گرفت. طراحی و تکمیل پرسشنامه‌ها، ثبت علایم بالینی و نتایج آزمایشگاهی در مجموعه داده‌های مورد استفاده، توسط نویسندگان این مقاله در بیمارستان فوق تخصصی شهدای کارگر یزد جمع‌آوری شد. پژوهش‌های مربوط به پیش‌بینی سندرم متابولیک محدودیت‌هایی از قبیل تعداد کم بیماران برای ایجاد مدل، داده‌های از دست رفته و متغیرهای ناقص را دارا می‌باشند. در این پژوهش تعداد قابل قبولی از بیماران با متغیرهای مناسب با حداقل داده‌های ازدست رفته به کار گرفته شد. مطالعه حاضر با به‌کارگیری ویژگی‌های بالینی و آزمایشگاهی بیماران با استفاده الگوریتم GBC و داده‌کاوی، سندرم متابولیک را پیش‌بینی می‌کند. مدل پیشنهادی این مطالعه شامل نرمال‌سازی ویژگی‌ها، انتخاب ویژگی‌ها، وزن دهی به ویژگی‌ها و معرفی روشی برای پیش‌بینی سندرم متابولیک با استفاده از ویژگی‌های وزن‌دار است.

ویژگی‌های سابقه فامیلی بالاترین کارایی در دسته‌بندی مدل پیشنهادی را دارا بودند. در این مطالعه علاوه بر شناسایی مهم ترین ویژگی‌ها، با استفاده از انتخاب ویژگی‌ها بر اساس الگوریتم ژنتیک به عملکردی بهتر بر حسب شاخص‌های دقت، حساسیت و ویژگی دست یافتیم.

Uzer و همکاران [۱۶] از ماشین‌های بردار پشتیبان برای طبقه‌بندی بهره گرفته‌اند. در پایگاه داده‌های مورد استفاده موجود، برخی از ویژگی‌های نه چندان متمایز و زائد وجود دارند. این ویژگی‌ها عوامل تأثیرگذار عمده‌ای در موفقیت ابزار طبقه‌بندی‌کننده و زمان پردازش سیستم محسوب می‌شوند. در سیستم ایجاد شده در این مطالعه تلاش شده است تا با حذف این ویژگی‌های زائد، میزان سرعت و موفقیت سیستم افزایش یابد؛ لذا هدف این مطالعه بررسیِ تأثیری است که حذف ویژگی‌های غیر ضروری و منسوخ در مجموعه داده‌ها بر موفقیت طبقه‌بندی با استفاده از طبقه‌بندی‌کننده ماشین بردار پشتیبان می‌گذارد. الگوریتم انتخاب ویژگی بر مبنای الگوریتم کلونی زنبور عسل که در این مطالعه ابداع شد، اولین نمونه از الگوریتم‌های کلونی زنبور عسل مورد استفاده در زمینه انتخاب ویژگی به‌شمار می‌رود. در این روش، یک فرایند جستجو اجرا می‌شود تا بهترین زیرمجموعه از ویژگی‌ها را بیابد. این روش برگرفته از روشی است که به‌طور معمول در تشخیص بیماری‌های کبد استفاده می‌شد. در این مقاله از مجموعه داده‌های حاصل از پایگاه داده (University of UCI(California, Irvine برای تشخیص بیماری‌های، هپاتیت، اختلالات کبدی و بیماری استفاده شد. همچنین، محققان برای این مجموعه داده‌ها از روش اعتبارسنجی متقاطع ۱۰ لایه‌ای برای به‌دست آوردن دقت در این طبقه‌بندی استفاده کردند. نتایج این بررسی نشان داد که کارایی و عملکرد این روش در مقایسه با سایر نتایج به‌دست آمده بسیار موفقیت‌آمیز بوده و به‌نظر می‌رسد روش امیدوارکننده‌ای برای برنامه‌های کاربردی تشخیص الگو باشد.

Shen و همکاران [۵]، یک طرح جدید برای تنظیم پارامترهای ماشین‌های بردار پشتیبان پیشنهاد داده‌اند که در آن از الگوریتم بهینه‌سازی مگس میوه استفاده می‌شود. این طرح ماشین بردار پشتیبان مبتنی بر بهینه‌سازی مگس میوه نام دارد که با موفقیت در تشخیص پزشکی مورد استفاده قرار گرفته است. رویکرد جدید این مطالعه در ارائه روش مبتنی بر بهینه‌سازی مگس میوه نهفته است که هدف آن بالابردن قابلیت تعمیم طبقه‌بندی‌کننده ماشین بردار پشتیبان از طریق بررسی تکنیک هوش ازدحامی جدید برای تنظیم پارامتر بهینه به‌منظور طبقه‌بندی داده‌های پزشکی است. در روش پیشنهادی، تکنیک بهینه‌سازی مگس میوه به‌طور مؤثر و کارآمد به پارامترهای موجود در ماشین بردار پشتیبان می‌پردازد. در طول رویه بهینه‌سازی پارامترها، پارامترهای ماشین بردار پشتیبان به‌صورت پویا از طریق تجزیه و تحلیل اعتبارسنجی متقاطع پنج لایه‌ای توسط بهینه‌سازی مگس میوه تنظیم می‌شوند. سپس پارامترهای بهینه به‌دست آمده به مدل پیش‌بینی ماشین بردار پشتیبان تغذیه می‌شوند تا با استفاده از تجزیه و تحلیل اعتبارسنجی متقاطع ده لایه‌ای، این کار طبقه‌بندی را برای تشخیص در حلقه بیرونی اجرا کند. گذشته از آن، کارایی و اثر بخشی ماشین بردار پشتیبان مبتنی بر بهینه‌سازی مگس میوه در مقایسه با چهار مجموعه داده پزشکی شناخته شده و معروف شامل مجموعه داده‌های ویسکانسین مجموعه داده‌های بیماری PID، مجموعه داده‌های بیماری پارکینسون و مجموعه داده های بیماری تیروئید به لحاظ دقت طبقه‌بندی، میزان حساسیت، ویژگی و مدت زمان پردازش به‌دقت مورد ارزیابی قرار گرفته است. از چهار همتای رقابتی از جمله ماشین بردار پشتیبان مبتنی بر الگوریتم بهینه‌سازی ازدحام ذرات، ماشین بردار پشتیبان مبتنی بر الگوریتم ژنتیک، ماشین بردار پشتیبان مبتنی بر بهینه‌سازی تغذیه باکتریایی و ماشین بردار پشتیبان مبتنی بر تکنیک جستجوی شبکه برای اهداف مقایسه‌ای

استفاده شد. نتایج تجربی این مطالعه نشان داد که روش ماشین بردار پشتیبان مبتنی بر بهینه‌سازی مگس میوه پیشنهاد شده می‌تواند پارامترهای الگوی بسیار مناسب‌تری را به‌دست بیاورد و به طور قابل توجهی زمان محاسبه را کاهش دهد که در نهایت به‌دقت بالا در طبقه‌بندی منجر می‌شود. در این روش پیشنهادی امید بر آن است که بتوان از آن به‌عنوان یک راه‌حل جایگزین و یک ابزار مفید بالینی در تصمیم‌گیری‌های پزشکی استفاده نمود.

Hayashi و همکاران [۱۷] از K-Means برای حذف داده‌های نوسان‌دار و از الگوریتم ژنتیک برای کشف مجموعه بهینه‌ای از ویژگی‌ها با کمک ماشین بردار پشتیبان به‌عنوان طبقه‌بندی‌کننده برای دسته‌بندی داده‌ها استفاده کرده‌اند. مجموعه داده مورد استفاده در این مطالعه، PID بوده که از ۷۶۸ نمونه حاضر در آن، ۲۶۸ مورد در کلاس "تست مثبت بیماری" و ۵۰۰ نفر در کلاس "تست منفی بیماری" قرار گرفتند. روند کار این الگوریتم به‌طور خلاصه به این‌صورت است: (۱) تمیز کردن داده‌ها با جایگزین کردن مقادیر از دست رفته با میانگین انجام می‌شود. (۲) مجموعه داده تمیز شده به‌منظور حذف داده‌های دورافتاده، متناقض و نویزدار، با استفاده از K-Means خوشه‌بندی شده و داده‌های کاهش‌یافته برای انتخاب ویژگی‌های مطلوب با الگوریتم ژنتیک مورد استفاده قرار می‌گیرند. (۳) مجموعه داده کاهش‌یافته با استفاده از طبقه‌بندی ماشین بردار پشتیبان به‌منظور رسیدن به‌دقت بهتری نسبت به روش‌های موجود در متون طبقه‌بندی می‌شوند. به‌منظور افزایش قابلیت اطمینان عملکرد طبقه‌بندی‌کننده، روش اعتبارسنجی متقاطع ۱۰ لایه‌ای به‌کار برده شده است. برای نشان دادن عملکرد الگوریتم‌های یادگیری ماشین نظارت شده، از ماتریس درهم‌ریختگی استفاده شده است.

در روش پیشنهادی با استفاده از کاهش تعداد متغیرها و وزن‌دار کردن ویژگی‌ها با به‌کارگیری الگوریتم GBC برای افزایش دقت با هدف طراحی و ارزیابی یک مدل پزشک‌یار در تشخیص سندرم متابولیک انجام شد. مدل پزشک‌یار طراحی شده در این پژوهش در تشخیص سندرم متابولیک موفق بوده است و دسته‌بندی را با دقت قابل قبولی انجام می‌گردد.

آزمایش‌ها و شبیه‌سازی نشان داد سیستم پزشک‌یار معرفی شده در این پژوهش بر روی مجموعه داده بیماران بومی مبتلا به سندرم متابولیک بیمارستان شهدای کارگر یزد به دقت ۹۲/۱٪ رسیده است که بالاتر از تحقیقات مشابه بر روی مجموعه داده‌های متفاوت بوده است.

در روش پیشنهادی برای مشخص کردن ارزش و نقش هر یک از ویژگی‌ها در تشخیص بیماری به طور تصادفی برای هر ویژگی، وزنی بین بازه [۱ ۰] اختصاص داده می‌شود که نشان دهنده درجه اهمیت ویژگی است. مقادیر وزن‌دار ویژگی‌ها به عنوان ورودی به الگوریتم KNN داده شده و توسط الگوریتم های تکاملی بهینه می‌گردد. بر اساس تابع هدف برای پیش بینی عارضه افزایش چربی خون از روش پیشنهادی، الگوریتم های گرگ خاکستری، ازدحام ذرات و ژنتیک برای بهبود عملکرد الگوریتم KNN استفاده شد.

نتیجه‌گیری

جستجو در پایگاه داده‌های پزشکی برای رسیدن به دانش و اطلاعات جهت پیش‌بینی، تشخیص و تصمیم‌گیری از کاربردهای داده‌کاوی در پزشکی است. می‌توان از الگوریتم‌های وراثتی مانند الگوریتم ژنتیک برای بهینه‌سازی تکنیک‌های داده کاوی استفاده کرد. پیش‌بینی و تشخیص صحیح سندرم متابولیک با استفاده از هوش مصنوعی و یادگیری ماشین، شانس درمان موفق را بالا می‌برد. در این مقاله برای پیش‌بینی و تشخیص سندرم متابولیک، از الگوریتم GBC برای بهینه سازی نتایج الگوریتم KNN استفاده شد و یک مدل جدید ارائه گردید. نتایج شبیه‌سازی نشان می‌دهد که مدل پیشنهادی با دقت پیش‌بینی ۰/۹۲۱ از روش‌های فازی، ماشین بردار پشتیبان، درخت‌تصمیم و شبکه عصبی دقت بیشتری دارد.

تشکر و قدردانی

از همکاران و نویسندگان این مقاله همچنین مسئولین محترم بیمارستان شهدای کارگریزد به دلیل پشتیبانی از این پژوهش تشکر و قدردانی می‌نماید.

References

1. Worachartcheewan A, Shoombuatong W, Pidetcha P, Nopnithipat W, Prachayasittikul V, Nantasenamat C. Predicting metabolic syndrome using the random forest method. The Scientific World Journal 2015;2015: 1-10.

2. Worachartcheewan A, Nantasenamat C, Isarankura-Na-Ayudhya C, Pidetcha P, Prachayasittikul V. Identification of metabolic syndrome using decision tree analysis. Diabetes Research and Clinical Practice. 2010;90(1):e15-e8.

3. Edrisi M, Gharipour M, Faroughi A, Javeri F, Shahgholi B, Gharipour A, et al. Decision support in prediction of metabolic syndrome with data mining methods. Pars Journal of Medical Sciences (Jahrom Medical Journal) 2011; 9(2): 48 - 58. Persian

4. Karegowda AG, Manjunath AS, Jayaram MA. Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima Indians diabetes. International Journal on Soft Computing. 2011 May;2(2):15-23.

5. Shen L, Chen H, Yu Z, Kang W, Zhang B, Li H, et al. Evolving support vector machines using fruit fly optimization for medical data classification. Knowledge-Based Systems 2016;96:61-75.

6. Jayalakshmi T, Santhakumaran A Novel classification method for diagnosis of diabetes mellitus using artificial neural networks. International Conference on Data Storage and Data Engineering; 2010 Feb 9-10; Bangalore, India: IEEE; 2010. p. 159-63.

7. Harper PR. A review and comparison of classification algorithms for medical decision making. Health Policy 2005;71(3):315-31.

8. AlJarullah A. Decision tree discovery for the diagnosis of type II diabetes. International Conference on Innovations in Information Technology; 2011 Apr 25-27; Abu Dhabi, United Arab Emirates: IEEE; 2011. p. 303-7.

9. Fang X. Are you becoming a diabetic? a data mining approach. Sixth International Conference on Fuzzy Systems and Knowledge Discovery; 2009 Aug 14-16; Tianjin, China: IEEE; 2009. p. 18-22.

10. Alshamlan HM, Badr GH, Alohali YA. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational Biology and Chemistry 2015;56:49-60.

11. Huang MJ, Chen MY, Lee SC. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. Expert Systems with Applications 2007;32(3):856-67.

13. Prilutsky D, Rogachev B, Marks RS, Lobel L, Last M. Classification of infectious diseases based on chemiluminescent signatures of phagocytes in whole blood. Artif Intell Med 2011;52(3):153-63.

13. Sheikhpour R, Sarram MA, Sheikhpour R. Particle swarm optimization for bandwidth determination and feature selection of kernel density estimation based classifiers in diagnosis of breast cancer. Applied Soft Computing 2016;40:113-31.

14. Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. *Age* 58, no. 13, pp. 10-110, 2006.

15. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent MC. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. Proc AMIA Symp 2000: 156–60.

16. Uzer MS, Yilmaz N, Inan O. Feature selection method based on artificial bee colony algorithm and support vector machines for medical datasets classification. The Scientific World Journal 2013;2013: 1-10.

17. Hayashi Y, Yukita S. Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. Informatics in Medicine Unlocked. 2016 Jan 1;2:92-104.

Original Article

Journal of Health and Biomedical Informatics
Medical Informatics Research Center
2018; 4(4):291-304

A Method for the Diagnosis of Metabolic Syndrome based on KNN Data Mining Algorithm: A case study in Shohada-ye Kargar Hospital in Yazd, Iran

Javad Lal dashti[1], Mohsen Mohammadi[2*], Farhang Padidarn Moghadam[3]

• **Received:** 19 Sep, 2017 • **Accepted:** 22 Feb, 2018

Introduction: Metabolic syndrome is a group of risk factors for developing cardiovascular diseases and diabetes in an individual. The presence of various signs and symptoms makes the diagnosis of this disease difficult. Data mining can provide clinical data analysis of patients for medical decision-makings. The purpose of this study was to provide a model for increasing the predictive accuracy of metabolic syndrome.

Method: In this applied-descriptive study, the medical records of 1499 patients with metabolic syndrome with 15 characteristics were investigated. Patients' information is collected from the standard database of Yazd Shohada-ye kargar Hospital. Each patient was followed for at least one year. In this paper, GBC algorithm was used to optimize the results of KNN data mining algorithm to predict and diagnose metabolic syndrome, and a new model was presented.

Results: Based on the objective function to predict the increase of blood lipids in the proposed method, gray wolf algorithms, particle swarm and genetics were used to improve the performance of the KNN algorithm. The analyses show that the proposed model with the precision accuracy of 0.921 has a greater accuracy compared to fuzzy methods, backup vector machine, tree decomposition and neural network.

Conclusion: Search in medical databases for the purpose of obtaining knowledge and information to predict, diagnose, and decision making are some applications of data mining in medicine. Hereditary algorithms can be used to optimize data mining techniques. The prediction and proper diagnosis of metabolic syndrome by using artificial intelligence and machine learning increases the chance of successful treatment.

Keywords: Metabolic syndrome, Bee colony algorithm, Decision tree

• **Citation:**. A Method for the Diagnosis of Metabolic Syndrome based on KNN Data Mining Algorithm:A case study in Shohada-ye Kargar Hospital in Yazd, Iran. Journal of Health and Biomedical Informatics 2018; 4(4): 291-304

1. MS.c. in Computer Engineering Eshragh Institute of Higher Education, Bojnourd, Iran
2. Ph.D. in Information Technology, Associate Professor, Computer Dept., Esfarayen University of Technology, Esfarayen, Iran
3. Ph.D. in Information Technology, Associate Professor, Computer Dept., Eshragh Institute of Higher Education, Bojnourd, Iran

*__Correspondence__: Computer Dep., Esfarayen University of Technology
• **Tel:** 05837266711 • **Email:** Mohsen@esfarayen.ac.ir