تشخیص خودکار بیماری کرونا (کوید-۱۹) با استفاده از تکنیک‌های داده‌کاوی: *یک گزارش کوتاه*

سید علی اکبر عرب‌زاده[۱]، وحید جمشیدی[۱*]، مسعود سعید[۱]، رستم یزدانی[۲]، مهدیه جمشیدی[۳]

۱- گروه مهندسی کامپیوتر، دانشکده مهندسی، دانشگاه شهید باهنر کرمان، کرمان، ایران.
۲- گروه داخلی، دانشکده پزشکی، دانشگاه علوم پزشکی کرمان، کرمان، ایران.
۳- گروه داخلی، دانشکده پزشکی، دانشگاه علوم پزشکی رفسنجان، رفسنجان، ایران.

[*] نویسنده مسئول: کرمان، بزرگراه امام خمینی، میدان پژوهش، دانشگاه شهید باهنر کرمان، گروه مهندسی کامپیوتر.
تلفن: ۰۳۴-۳۳۲۵۷۵۰۱
E-mail: vjamshidi@uk.ac.ir

دریافت: ۱۴۰۰/۰۶/۲۰ ویرایش: ۱۴۰۰/۰۶/۲۷ پذیرش: ۱۴۰۰/۰۹/۲۳ آنلاین: ۱۴۰۰/۱۰/۰۱

چکیده

زمینه و هدف: تشخیص زود هنگام بیماری همه‌گیر کرونا (کوید-۱۹) نقش مهمی در جلوگیری از مرگ‌ومیر دارد. مطالعه حاضر با هدف استخراج ویژگی‌های کلیدی بیماری و یافتن بهترین روش‌های داده‌کاوی که باعث بهبود صحت تشخیص بیماری کرونا می‌شود، صورت گرفته است.

روش بررسی: در این مطالعه، اطلاعات و علایم بالینی مربوط به بیماران مبتلا به کوید-۱۹، از کلینیک‌های عفونی بیمارستان افضلی پور کرمان و علی‌بن‌ابیطالب رفسنجان از اسفند ۱۳۹۸ تا بهمن ۱۳۹۹ به‌مدت ۱۲ ماه گردآوری شدند. مدل‌های پیش‌بینی با استفاده از ترکیب‌های مختلف از ویژگی‌های بیماری و هفت روش داده‌کاوی، ایجاد شده و مورد آزمایش قرار گرفتند.

یافته‌ها: ۹ ویژگی کلیدی و سه تکنیک برتر داده‌کاوی شناسایی شدند. نتایج آزمایش‌ها نشان می‌دهند که مدل پیش‌بینی ارایه شده با استفاده از ویژگی‌های کلیدی شناسایی شده و بهترین تکنیک داده‌کاوی به‌دست آمده، دقت ۸۳/۱۹٪ را برای تشخیص بیماری کرونا فراهم می‌کند.

نتیجه‌گیری: با توجه ویژگی‌های کلیدی و تکنیک‌های داده‌کاوی به‌دست آمده از این مطالعه می‌توان با تکمیل مجموعه داده و استفاده از این روش، با دقت بالا ابتلا به کوید- ۱۹ را در افراد با ویژگی‌های بالینی مختلف تشخیص داد.

کلمات کلیدی: داده‌کاوی، تشخیص، علایم بالینی، کروناویروس، بیماری کوید-۱۹، همه‌گیر.

مقدمه

در چندماه گذشته، بیماری کوید- ۱۹ سبک زندگی عادی افراد جامعه را به شدت دستخوش تغییر کرده است. با توجه به اینکه در دنیای امروز کشورها دارای ارتباطاتی قوی و پیوسته هستند، اثرات این بیماری فراتر از مرگ‌ومیر و بیماری افراد جامعه است. بررسی‌های اقتصادی نشان داده است که شیوع این بیماری، بخش‌های بزرگی از اقتصاد در حوزه گردشگری، بازار سرمایه، بازار ارز، کسب و کارهای خرد و عمومی، قیمت کالاها و تولید ناخالص را تحت تاثیر قرار داده است. با توجه به این مهم، هر تلاشی در راستای کاهش این آسیب‌ها اهمیت بسیاری دارد. در این راستا، تشخیص سریع‌تر و دقیق‌تر بیماری می‌تواند تا حد زیادی از شیوع این بیماری و آسیب‌های تصاعدی وارده جلوگیری کند.

با توجه به حجم عظیم داده‌های در دسترس در مورد این بیماری، تکنیک‌های داده‌کاوی می‌توانند رویکرد موثری جهت استفاده و پردازش این حجم از اطلاعات داشته باشند. روش‌های متنوعی توسط پژوهشگران در حوزه داده‌کاوی به‌منظور دستیابی به تشخیص بیماری‌ها، معرفی شده‌اند. به‌عنوان مثال می‌توان به پژوهش Srinivas و همکاران اشاره کرد که از روش‌های داده‌کاوی برای شناسایی و استخراج اطلاعات مفید از مجموعه داده‌های بالینی استفاده می‌کند.[۱]

در این مقاله به‌صورت خلاصه تعدادی از کارهای انجام شده را شرح می‌دهیم.

Sepehri و همکاران، مقاله "ساخت مدلی برای تعیین نوع درمان سنگ حالب" را برای طراحی سیستمی پیشنهاد کردند که می‌تواند به‌منظور انتخاب نوع درمان سنگ کلیه براساس ویژگی‌های سنگ و بیمار، در تصمیم‌گیری کمک کند.[۲] این مقاله ضمن استفاده از روش درخت تصمیم، به دقت ۸۵٪ در تصمیم‌گیری نائل شده است. همچنین جهت پیش‌بینی ابتلای ساکنان آمریکا به دیابت براساس سبک زندگی، تحقیقاتی انجام شده است و در نهایت الگوهای پیش‌بینی استخراج شده‌اند.[۳] در مقاله "راهکاری جهت کشف روابط بین عوامل خطرزای قلبی" با استفاده از داده‌کاوی روابطی بین عوامل خطر آفرین در بیماری‌های قلب و عروق شناسایی شده است. در این مقاله نیز از درخت تصمیم استفاده شده است.[۴] در مقاله "تشخیص سکته مغزی به کمک تکنیک‌های داده‌کاوی" تحقیقی برای تشخیص دقیق و کنترل موثر سکته مغزی صورت گرفته است که در آن از تکنیک‌های درخت تصمیم و نزدیکترین همسایه برای دسته‌بندی داده‌ها استفاده شده است و به‌ترتیب دقت ۹۵/۴۲٪ و ۹۴/۱۸٪ به‌دست آمده است.[۵]

تحقیقاتی نیز در زمینه پیش‌بینی سرطان‌های مختلف با استفاده از تکنیک‌های داده‌کاوی صورت گرفته است. به‌عنوان مثال Li و همکاران، با داده‌کاوی و استفاده از پروفایل پروتئومی سرم (Serum proteomic profile)، به پیش‌بینی سرطان پرداخته‌اند.[۶] در این پژوهش به‌دلیل پیچیدگی ناشی از پروفایل پروتئومی و حجم زیاد داده‌های تحت آنالیز، آنها از تکنیک‌های داده‌کاوی جهت یافتن الگوهای پیچیده استفاده کردند و در نهایت توانستند با دقت قابل‌قبولی در تشخیص سرطان تخمدان موفق شوند.

شایان ذکر است در بیماری‌های مغزی نیز داده‌کاوی به کمک آمده است. Miller و همکاران مطالعاتی در این خصوص انجام داده‌اند. آنها با استفاده از داده‌کاوی، دو کلاس بیماران مبتلا به آلزایمر و نرمال را تعریف کردند و به مدد داده‌کاوی نه تنها به نتایج ارزشمندی در پیش‌بینی زودهنگام آلزایمر دست یافتند بلکه راهکارهایی جهت کمک به بیماران مبتلا نیز پیشنهاد کردند.[۷]

افزون‌بر تحقیقات فوق‌الذکر، فعالیت‌های مشابهی در علم پزشکی با استفاده از تکنیک‌های داده‌کاوی جهت یافتن الگوهای پیش‌بینی کارآمد انجام شده است که از مهمترین آنها می‌توان به سرطان پستان، بررسی فاکتورهای موثر در ژن‌ها، بیماری‌های قلبی و عروقی، سرطان پروستات و دیابت اشاره کرد.[۸-۱۲]

اما باید به این نکته توجه داشت که کارایی تکنیک‌های داده‌کاوی، تا حد زیادی به تکنیک مورد استفاده و ویژگی‌های انتخاب شده بستگی دارد. از طرف دیگر، مجموعه داده‌های پزشکی در حوزه بهداشت و درمان بسیار زیاد و گاهی متناقض می‌باشند که این امر باعث شده است که استفاده از روش‌های داده‌کاوی بدون آماده‌سازی مناسب داده‌ها، ثمربخش نباشد. زیرا همانگونه که در مراجع بیان شده است ازدیاد داده‌ها و تناقض در مجموعه داده‌های خام، اثرات منفی بر نتایج الگوریتم‌ها خواهد گذاشت.[۱۳-۱۶] بنابراین، آماده‌سازی و پیش‌پردازش داده‌ها، از ملزومات استفاده از پتانسیل حداکثری روش‌های داده‌کاوی می‌باشد. افزون‌براین، ویژگی‌های نامرتبط و یا زاید نیز می‌توانند عملکرد تکنیک‌های داده‌کاوی را تضعیف نمایند.[۱۷] بنابراین، جهت دستیابی به دقت بالا در شناسایی بیماری کرونا با استفاده از روش‌های داده‌کاوی، علاوه‌بر آماده‌سازی داده‌ها، استفاده از ویژگی‌های کلیدی و تکنیک‌های داده‌کاوی مناسب، از اهمیت بالایی برخوردار هستند.

در همین راستا، این تحقیق تلاش کرده است تا با شناسایی ویژگی‌های کلیدی و تکنیک‌های داده‌کاوی مناسب، به تشخیص بیماری کرونا با دقت بالا بپردازند.

روش بررسی

در این مطالعه، داده‌های بیماری کرونا از کلینیک عفونی بیمارستان افضلی‌پور کرمان از اسفند ۱۳۹۸ تا بهمن ۱۳۹۹ به‌مدت دوازده ماه گردآوری شده‌اند. مجموعه داده ای تنظیم شده، شامل ۱۰۲ رکورد می باشد.

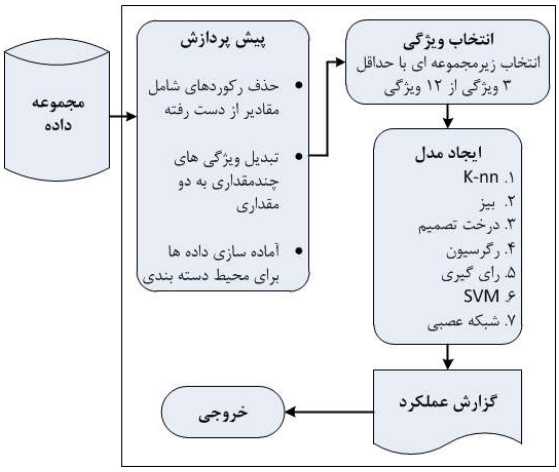
با بررسی پرونده‌های بیماران مبتلا به بیماری کرونا و آنالیز آنها، جدول ۱ استخراج شده است. همانگونه که در جدول ۱ نشان داده شده است ۱۳ ویژگی در نظر گرفته شده که ۱۲ ویژگی آن مربوط به تشخیص بیماری کرونا و یک ویژگی به‌عنوان خروجی (ویژگی پیش‌بینی شده) برای نشان دادن شدت بیماری کرونا در فرد می‌باشد. خروجی یا ویژگی NUM عددی از صفر تا چهار است که «صفر» برای

جدول ۱: شرح ویژگی‌های مجموعه داده‌ای بیمارستان افضلی‌پور

ویژگی	نوع	توضیحات
سن	عددی	سن بیمار به سال
سرفه خشک	اسمی	عدد یک برای بله و صفر برای خیر
گلودرد	اسمی	ارزش صفر: عدم گلودرد ارزش یک: گلودرد
تب	عددی	دمای بدن (به واحد درجه سانتی‌گراد)
بیماری زمینه‌ای	اسمی	عدد یک برای بله و صفر برای خیر
خستگی	اسمی	عدد یک برای بله و صفر برای خیر
ESR	اسمی	عدد یک به معنای وجود التهاب و صفر به معنای عدم وجود التهاب است
تنگی نفس	اسمی	عدد یک برای بله و صفر برای خیر
درد عضلانی	اسمی	عدد یک برای بله و صفر برای خیر
CBC	عددی	مقدار گلبول‌های سفید موجود در خون
CPK	اسمی	عدد یک به معنای نتیجه نرمال و صفر به معنای نتیجه غیرعادی است
LDH	عددی	مقدار موجود در نتایج آزمایش‌ها
NUM	اسمی	تشخیص بیماری کرونا: صفر به معنای عدم وجود و یک به معنای وجود

نشان دادن عدم وجود بیماری کرونا و مقادیر یک تا چهار برای نشان دادن شدت بیماری هستند (چهار بیشترین شدت بیماری است).

در این تحقیق، برای انجام آزمایش از (Rapid Miner Studio, RapidMiner, USA) که نرم‌افزاری توانمند در مباحث داده‌کاوی می‌باشد استفاده شده است. این نرم‌افزار می‌تواند تمامی مراحل مورد نیاز، از آماده‌سازی اطلاعات اولیه گرفته تا بصری کردن نتایج، ارزیابی، سنجش اعتبار و بهینه‌سازی خروجی را در یک محیط یکپارچه و واحد انجام دهد. شکل ۱، گردش کار آزمایش را نشان می‌دهد. برای انجام آزمایش، مجموعه داده‌ای بیمارستان افضلی‌پور به Rapid Miner داده می‌شود. فرآیند داده‌کاوی از مرحله پیش‌پردازش آغاز می‌شود. انتخاب ترکیب‌های مختلف از ویژگی‌ها، ایجاد مدل و استخراج نتایج انجام می‌شود و نتایج به‌دست آمده مورد ارزیابی قرار می‌گیرند. در صورتی که نتایج ارزیابی مناسب باشد منجر به پیاده‌سازی مدل برای شناسایی بیماری می‌شود. در فرآیند داده‌کاوی تشخیص کرونا، زیر مجموعه‌ای حاوی حداقل ۳ ویژگی از ۱۲ ویژگی انتخاب می‌شود سپس ایجاد مدل برای هر ترکیب از آنها،



شکل ۱: گردش کار پیشنهاد شده برای آزمایش‌ها

انجام می‌شود. عملکرد هر مدل ساخته شده، براساس ویژگی‌ها و تکنیک داده‌کاوی در طول هر تکرار، ثبت می‌شود و نتیجه پس از اتمام کل فرآیند نشان داده می‌شود.

در شکل ۱، گردش کار پیشنهاد شده برای آزمایش‌ها نشان داده شده است مراحل آن به شرح زیر است:

پیش‌پردازش داده‌ها: داده‌ها پس از جمع‌آوری، پیش‌پردازش می‌شوند. در مجموعه داده‌ای افضلی‌پور، مقدار ویژگی (NUM) برای داشتن بیماری کرونا از رکوردهای چند مقداری (صفر برای عدم وجود و ۱،۲،۳،٤ برای وجود) به رکوردهای دو مقداری (صفر برای عدم وجود و یک برای وجود بیماری کرونا) تبدیل می‌شود. کار پیش‌پردازش داده‌ها با تبدیل کلیه مقادیر تشخیص (۲،۳،٤) به یک انجام می‌شود. مجموعه داده‌ای حاصل فقط شامل مقادیر صفر و یک است که صفر معرف فقدان و یک معرف ابتلا بیمار به کرونا است.

انتخاب ویژگی: از میان ۱۲ ویژگی مورد استفاده در تشخیص بیماری کرونا، فقط «سن» به اطلاعات شخصی هر بیمار اشاره دارد. ۱۱ ویژگی باقیمانده شامل خصوصیات بالینی گردآوری شده از معاینات پزشکی می‌باشند. در انجام آزمایش، ترکیبی از ویژگی‌ها برای استفاده در هفت روش داده‌کاوی (k-NN، درخت تصمیم (Decision tree)، بیز (Bayes)، رگرسیون (Regression)، رای‌گیری (Voting)،

SVM و شبکه عصبی (Neural network) مورد استفاده قرار می‌گیرد. ابتدا تمام ترکیبات احتمالی ۳ ویژگی از ۱۲ ویژگی انتخاب شده و با استفاده از هفت روش داده‌کاوی، هر ترکیب مورد آزمایش قرار می‌گیرد. در مرحله بعد، آزمایش برای انتخاب همه ترکیبات ممکن ٤ ویژگی از ۱۲ ویژگی انجام می شود. تعداد کل ترکیب‌های قابل دستیابی از مجموعه‌ای با ۱۲ ویژگی، به استثنای مجموعه تهی، ۱-۲n است. در این تحقیق، یک زیر مجموعه واحد از ویژگی‌ها نمی‌تواند کمتر از ۳ ویژگی داشته باشد. بنابراین، تمام زیر مجموعه‌های ترکیبی که با داشتن ۱ یا ۲ ویژگی حاصل می‌شوند، حذف می‌شوند. معادله‌ای که برای محاسبه تعداد کل ترکیب‌ها حاصل می‌شود عبارت است از:

$$\begin{aligned}\text{تعداد کل ترکیب‌ها} &= 2^n - \left(\frac{n!}{1!(n-1)!}\right) - \left(\frac{n!}{2!(n-2)!}\right) - 1 \\ &= 2^n - n - \frac{n(n-1)}{2} - 2 \\ &= 2^n - \left(\frac{2n-n^2-n}{2} + 1\right) \\ &= 2^n - \left(\frac{n^2-n}{2} + 1\right)\end{aligned} \quad (۱)$$

n تعداد کل ویژگی‌هایی است که برای تولید زیر مجموعه‌های ترکیبی به‌کار می‌رود که در این آزمایش ۱۲ است. بنابراین، در مجموع ۸۱۰۰ ترکیب از ویژگی‌ها در این آزمایش انتخاب و آزمایش شده‌اند.

ایجاد مدل: پس از انتخاب ویژگی‌ها، مدل‌ها با استفاده از ۷ روش معروف داده‌کاوی ساخته می‌شوند: k-NN، درخت تصمیم، بیز، رگرسیون، SVM، شبکه عصبی و رأی‌گیری. سپس برای سنجش اعتبار کارایی مدل‌ها، از روش سنجش اعتبار Folds cross validation–10 استفاده می‌شود. در این روش، کل مجموعه داده به ۱۰ زیر مجموعه هم‌اندازه تقسیم می‌شود. سپس سنجش اعتبار ۱۰ بار تکرار می‌شود. در هر بار تکرار، یکی از ۱۰ مجموعه به‌عنوان مجموعه تست و مابقی به‌عنوان مجموعه آموزش انتخاب می‌شوند. سرانجام، میانگین نتایج حاصل از ۱۰ تکرار ارایه می‌شود. در افراز مجموعه داده به زیرمجموعه‌ها، از نمونه‌گیری طبقه‌ای استفاده می‌شود تا در هر زیر مجموعه تعداد نمونه یکسانی از مجموعه داده اصلی وجود داشته باشد.

کارایی مدل‌ها، با استفاده از سه معیار مطرح در این حوزه اندازه‌گیری شده است: دقت، صحت و معیار F. منظور از «دقت» درصد موارد صحیح تشخیص داده شده در بین کلیه موارد است. «معیار F» میانگین وزنی دقت و فراخوانی است. «صحت» درصد تشخیص‌های صحیح برای نمونه مثبت است.

برای شناسایی ویژگی‌های کلیدی، این سه معیار کارایی مورد استفاده قرار گرفتند و برای شناسایی تکنیک داده‌کاوی برای ایجاد بهترین مدل، معیارهای دقت و صحت مورد استفاده قرار گرفتند. برای شناسایی ویژگی‌های کلیدی، سه معیار ذکر شده، درک بهتری از رفتار ترکیبات متفاوت ویژگی‌ها ارایه می‌دهند. از طرف دیگر، تجزیه و تحلیل تکنیک‌های داده‌کاوی، بر روی بهترین مدل‌های اجرا تمرکز دارد که می‌تواند دقت بالایی در پیش‌بینی و تشخیص بیماری کرونا فراهم کند زیرا دقت و صحت بهترین معیارهای ارزیابی کارایی هستند. برای هر روش داده‌کاوی، معیارها به‌صورت جداگانه اندازه‌گیری شدند و تمام نتایج برای تجزیه و تحلیل ثبت شدند.

یافته‌ها

جدول ۲، ۳، و٤ کارایی ۷ تکنیک داده‌کاوی را در سه معیار مختلف نشان می‌دهند. براساس تجزیه و تحلیل نشان داده شده در جداول، می‌توانیم مشاهده کنیم که بالاترین میزان دقت (۸۲/٤۳٪) توسط SVM با ۹ ویژگی انجام شده است. از طرف دیگر، بالاترین میزان صحت (۱۰۰/۰۰٪) با استفاده از ۳ ویژگی (یعنی سن، تب، گلودرد) با درخت تصمیم حاصل شده است، در حالی‌که بیشترین میزان F (۹۸/۳۵٪) با بیز با ۹ ویژگی انجام شده است. براساس جدول ۲ سه تکنیک برتر SVM، بیز و شبکه عصبی هستند که دقتی بیش از ۷۵٪ دارند. همچنین نتایج نشان می‌دهند که دو روش درخت تصمیم و k-NN کمترین میزان دقت را نسبت به سایر تکنیک‌ها دارند. درحالی‌که روش درخت تصمیم بالاترین میزان صحت را دارا می‌باشد.

با استفاده از هر روش داده‌کاوی، میانگین دقت، میانگین صحت، و میانگین F بر روی ۸۱۰۰ ترکیب مختلف از ویژگی‌ها نیز محاسبه شده است. نتایج نشان دادند که بالاترین میانگین دقت در روش‌های داده‌کاوی مربوط به بیز، شبکه عصبی و SVM به‌ترتیب با داشتن میانگین ۷۱/۲۰٪ و ۷۰/۱۲٪ و ٦۹/۱۵٪ می‌باشد. بالاترین میانگین صحت مربوط به شبکه عصبی، بیز و SVM به‌ترتیب با داشتن میانگین ۷۱/۳۲٪ و ۷۰/۷٦٪ و ٦۹/۱۵٪ می‌باشد. درحالی‌که بالاترین میانگین F

مربوط به SVM، رگرسیون و بیز بترتیب با داشتن میانگین ۸۰/۹۸٪ و ۸۰/۲۵٪ و ۸۰/۱۷٪ می‌باشد. حال باید با تجزیه و تحلیل این نتایج، ویژگی‌های کلیدی و تکنیک‌های داده‌کاوی که تأثیر بسزایی در ایجاد بهترین مدل دارند، برای تشخیص بیماری کرونا شناسایی و انتخاب شوند:

– انتخاب ویژگی‌های کلیدی: به‌منظور شناسایی ویژگی‌های کلیدی، تجزیه و تحلیلی انجام شد تا مشخص شود چند بار یک ویژگی در مدلی که بالاترین میزان دقت، صحت و F را دارد ظاهر می‌شود. از ۸۱۰۰ ترکیب ویژگی‌ها، ترکیب‌هایی که منجر به بالاترین کارایی یک تکنیک خاص شده‌اند، مشخص شدند. جدول ۵ تجزیه و تحلیل ویژگی‌ها را نشان می‌دهد. در این جدول، زمانیکه یک ویژگی در ترکیبی با بالاترین کارایی ظاهر می‌شود، شمرده می‌شود تا مشخص شود هر ویژگی چندبار تاثیرگذار بوده است. به عبارت دیگر ردیف اول جدول ۵ نشان می‌دهد که چند بار هر یک از این ویژگی‌ها در بین ترکیب‌هایی که به بالاترین دقت در بین ۷ تکنیک منجر شده‌اند، وجود دارد. به‌همین ترتیب، ردیف دوم و سوم رخداد ویژگی‌هایی را نشان می‌دهد که بالاترین صحت و میزان F را به خود اختصاص داده‌اند. در آخر، جمع‌بندی از همه رخدادهای هر ویژگی محاسبه شده است.

از بین ۱۱ ویژگی، تب، ویژگی است که بیشترین تعداد رخداد را داشته و ۲۰ بار در همه ترکیب‌ها ظاهر شده است. این موضوع نشان می‌دهد که این ویژگی مهمترین ویژگی است که با دقت، صحت و میزان F بالا در تشخیص‌ها موثر است. در این تحقیق، ویژگی‌هایی که حداقل ۱۰ بار ظاهر شده‌اند و بیشترین کارایی را داشته‌اند، به‌عنوان ویژگی‌های کلیدی در تشخیص بیماری کرونا شناخته شده‌اند. براساس آنالیز جدول ۵، ۹ ویژگی به‌عنوان ویژگی‌های مهم در تشخیص بیماری کرونا شناخته شده‌اند: «سن»، «تب»، «گلو درد»، «تنگی نفس»، «سرفه خشک»، «درد عضلانی»، «بیماری زمینه‌ای»، «خستگی» و «ESR».

– انتخاب روش‌های برتر داده‌کاوی: برای تکمیل مدل پیشنهادی خود، باید ببینیم براساس ویژگی‌های کلیدی به‌دست آمده، کدامیک از تکنیک‌های داده‌کاوی بالاترین کارایی را ارایه می‌کنند.

سه تکنیک برتر داده‌کاوی که بالاترین میانگین دقت و صحت را دارند عبارتند از: شبکه عصبی، بیز و SVM. به‌منظور نهایی کردن انتخاب سه تکنیک برتر، نتایج با نتایج به‌دست‌آمده در جداول ۲ و ۳ نیز مورد بررسی قرار گرفتند و تایید شدند. برای سنجش اعتبار ویژگی‌های کلیدی و تکنیک‌های داده‌کاوی به‌دست آمده، از مجموعه

جدول ۲: بالاترین میزان دقت حاصل در هر روش داده‌کاوی

تکنیک	ترکیب	دقت
SVM	سن، تب، گلودرد، تنگی نفس، سرفه خشک، درد عضلانی، بیماری زمینه‌ای، خستگی، ESR	٪۸۲/۴۳
شبکه عصبی	سن، تب، گلودرد، تنگی نفس، سرفه خشک، درد عضلانی، بیماری زمینه‌ای، خستگی، ESR،CBC، LDH	٪۷۲/۸۳
بیز	سن، تب، گلو درد، تنگی نفس، سرفه خشک، CBC، LDH	٪۷۷/۱۷
رگرسیون	سن، تب، بیماری زمینه‌ای، تنگی نفس، سرفه خشک، خستگی، ESR، CBC، LDH	٪۶۳/۰۴
رأی‌گیری	سن، تب، گلو درد، تنگی نفس، سرفه خشک، درد عضلانی، بیماری زمینه‌ای، خستگی، LDH	٪۶۳/۰۴
درخت تصمیم	سن، CPK، بیماری زمینه‌ای، تنگی نفس، سرفه خشک، خستگی، ESR، CBC، LDH	٪۵۱/۰۹
k-NN	سن، CPK، بیماری زمینه‌ای، تنگی نفس، سرفه خشک، خستگی، ESR، CBC، LDH	٪۴۴/۵۷

جدول ۳: بالاترین میزان صحت حاصل در هر روش داده‌کاوی

تکنیک	ترکیب	صحت
درخت تصمیم	سن، تب، گلودرد	٪۱۰۰/۰۰
SVM	سن، تب، بیماری زمینه‌ای، سرفه خشک، خستگی	٪۹۱/۸۰
بیز	سن، تب، گلودرد، سرفه خشک، LDH	٪۷۵/۳۲
شبکه عصبی	سن، تب، ESR	٪۷۲/۱۵
رأی‌گیری	سن، تب، گلودرد، تنگی نفس، سرفه خشک، درد عضلانی، LDH، خستگی، ESR	٪۶۳/۰۴
رگرسیون	سن، تب، خستگی، درد عضلانی، CBC، خستگی، LDH، ESR، CPK	٪۶۳/۰۴
k-NN	سن، تب، گلودرد	٪۵۶/۳۶

جدول ٤: بالاترین میزان F در هر روش داده‌کاوی

تکنیک	ترکیب	مقیاس F
بیز	تب، درد عضلانی، سرفه خشک، بیماری زمینه‌ای، خستگی، ESR	٪۹۸/۳۵
SVM	سن، تب، گلودرد، تنگی نفس، سرفه خشک، درد عضلانی، بیماری زمینه‌ای، خستگی، ESR	٪۸٦/۱۵
رگرسیون	سن، تب، درد عضلانی، سرفه خشک، بیماری زمینه‌ای، خستگی، ESR، LDH	٪۸۱/۷٤
شبکه عصبی	سن، تب، گلودرد، تنگی نفس، سرفه خشک، درد عضلانی، خستگی، بیماری زمینه‌ای، ESR، CBC، LDH، PKR	٪۷۵/۳۸
رأی‌گیری	درد عضلانی، بیماری زمینه‌ای، خستگی، ESR، CBC، LDH، PKR	٪٦۳/۰٤
k-NN	گلودرد، خستگی، ESR، CBC، LDH، PKR	٪٦۲/۳٤
درخت تصمیم	گلودرد، خستگی، ESR، CBC، LDH، CPK	٪۵۳/۸۵

داده‌های بیماری کرونای کلینیک عفونی بیمارستان علی‌بن‌ابیطالب رفسنجان نیز استفاده شده است (در مجموع ۹۷ رکورد). گردش کار آزمایش (شکل ۱)، بر روی مجموعه داده‌های کلینیک علی‌بن‌ابیطالب اعمال گردید. نتایج نشان می‌دهند که ویژگی‌های کلیدی شناسایی شده، دقت را در سه روش برتر داده‌کاوی، بهبود بخشیده‌اند. این موضوع، نتایج به‌دست آمده از بیمارستان افضلی‌پور را در مورد ویژگی‌های کلیدی جهت تشخیص بیماری کرونا تایید می‌کند. از میان ۹ ویژگی، ۸ مورد از آنها ویژگی‌های بالینی گردآوری شده از معاینات پزشکی می‌باشند. فقط یک ویژگی «سن»، ویژگی آماری محسوب می‌شود. این نشان می‌دهد در تکنیک‌های داده‌کاوی، ویژگی‌های بالینی، بیشتر از ویژگی آماری در تشخیص بیماری کرونا موثر هستند.

بحث

صنعت بالینی دارای داده‌های عظیمی از بیماران است که پردازش نشده‌اند. یافتن راهی برای پردازش این داده‌های خام و تبدیل آن به یک گوهر اطلاعاتی می‌تواند زندگی افراد زیادی را نجات بخشد. استفاده از تکنیک‌های داده‌کاوی، یک روش کارآمد برای تجزیه و تحلیل این داده‌های خام با حجم زیاد می‌باشد که می‌تواند با شناخت درست از گذشته، آینده را پیش‌بینی کند و با پیش‌بینی‌های دقیق، بینش جدیدی را در مورد تشخیص و پیشگیری بیماری‌ها ارایه کند. هر چند تصمیم‌هایی که توسط انسان گرفته می‌شوند بیشتر مطلوب هستند اما وقتی با حجم زیادی از داده‌ها روبه‌رو باشیم تصمیمات انسانی عملکرد پایین‌تری را ارایه می‌دهند. همچنین کارایی و دقت تصمیمات زمانی که انسان‌ها با حجم و فشار کاری مواجه هستند کاهش می‌یابد.

امروزه بیماری کرونا به یکی از دلایل اصلی مرگ‌ومیر در جهان تبدیل شده است. تشخیص این بیماری در اسرع وقت برای جلوگیری از مرگ‌ومیر بسیار مهم است. در این تحقیق برای دستیابی به دقت بالا در تشخیص بیماری کرونا، در ابتدا یک گردش کار کامل و موثر برای استفاده از روش‌های داده‌کاوی پیشنهاد گردید که شامل مـراحل

جدول ۵: مقایسه ویژگی‌ها براساس تعداد دفعات رخداد در روش‌های داده‌کاوی با بالاترین میزان کارایی

ویژگی / رخداد	تب	درد عضلانی	تنگی نفس	بیماری زمینه‌ای	CBC	ESR	خستگی	LDH	گلودرد	سن	CPK
رخداد با بالاترین میزان دقت	۷	٤	٦	٤	۳	٤	۵	۱	۷	۷	۲
رخداد با بالاترین میزان F	۷	٤	٦	٤	۳	٤	٦	۱	۷	۷	۲
رخدا با بالاترین میزان صحت	٦	٤	۲	٤	۲	۲	۳	۲	٤	۵	۰
تعداد کل رخدادها	۲۰	۱۲	۱٤	۱۲	۸	۱۰	۱٤	٤	۱۸	۱۹	٤

پیش‌پردازش داده‌ها، انتخاب ویژگی‌ها، ایجاد مدل، ثبت عملکرد مدل و نشان دادن نتیجه می‌باشد.

برای شناسایی ویژگی‌های کلیدی، از سه معیار دقت، صحت و معیار F استفاده شد و برای شناسایی روش‌های برتر داده‌کاوی، معیارهای دقت و صحت در نظر گرفته شد. برای هر روش داده‌کاوی، معیارها به‌صورت جداگانه اندازه‌گیری شدند و تمام نتایج برای تجزیه و تحلیل ثبت شدند. سرانجام ویژگی‌های کلیدی و بهترین روش‌های داده‌کاوی که باعث تشخیص بیماری کرونا با دقت بالا می‌شوند، استخراج شدند.

آزمایش اول برروی مجموعه داده‌ای کلینیک عفونی بیمارستان افضلی پور کرمان با استفاده از گردش کار پیشنهاد شده انجام شد. بالاترین میزان دقت (۸۲/۴۳٪) توسط SVM با ۹ ویژگی به‌دست آمد. از طرف دیگر، بالاترین میزان صحت (۱۰۰/۰۰٪) با استفاده از ۳ ویژگی (یعنی سن، تب، گلودرد) با درخت تصمیم حاصل شد، در حالی که بیشترین میزان F (۹۸/۳۵٪) با بیز با ۹ ویژگی به‌دست آمد. همچنین نتایج نشان دادند که دو روش درخت تصمیم و k-NN کمترین میزان دقت را نسبت به سایر تکنیک‌ها دارند. درحالی‌که روش درخت تصمیم بالاترین میزان صحت را دارا می‌باشد.

همچنین با اعمال تکنیک‌های داده‌کاوی بر روی ۸۱۰۰ ترکیب مختلف از ویژگی‌ها، مشاهده گردید بالاترین میانگین دقت در روش‌های داده‌کاوی مربوط به بیز، شبکه عصبی و SVM به‌ترتیب با داشتن میانگین ۷۱/۲۰٪ و ۷۰/۱۲٪ و ۶۹/۱۵٪ می‌باشد. بالاترین میانگین صحت مربوط به شبکه عصبی، بیز و SVM به‌ترتیب با داشتن میانگین ۷۱/۳۲٪ و ۷۰/۷۶٪ و ۶۹/۱۵٪ می‌باشد. درحالی‌که بالاترین میانگین F مربوط به SVM، رگرسیون، و بیز به‌ترتیب با داشتن میانگین ۸۰/۹۸٪ و ۸۰/۲۵٪ و ۸۰/۱۷٪ می‌باشد. بنابراین با در نظر گرفتن دو معیار دقت و صحت، سه تکنیک برتر داده‌کاوی که به بالاترین میزان دقت و صحت رسیده‌اند عبارتند از: شبکه عصبی، بیز و SVM.

به‌منظور شناسایی ویژگی‌های کلیدی، تجزیه و تحلیلی انجام شد تا مشخص شود چند بار یک ویژگی در مدلی که بالاترین میزان دقت، صحت و F را دارد ظاهر می‌شود. براساس آنالیز انجام شده، ۹ ویژگی به‌عنوان ویژگی‌های مهم در تشخیص بیماری کرونا شناسایی شدند: سن، تب، گلو درد، تنگی نفس، سرفه خشک، درد عضلانی، بیماری زمینه‌ای، خستگی و ESR.

یافته‌ها از طریق آزمایش دوم با استفاده از مجموعه داده‌ای کلینیک عفونی بیمارستان علی‌بن‌ابیطالب مورد بررسی قرار گرفتند. ۹ ویژگی کلیدی و سه تکنیک برتر داده‌کاوی تایید شدند. ویژگی‌های کلیدی عبارتند از: سن، تب، گلودرد، تنگی نفس، سرفه خشک، درد عضلانی، بیماری زمینه‌ای، خستگی و ESR. سه تکنیک برتر داده‌کاوی عبارتند از: شبکه عصبی، بیز و SVM.

همچنین نتایج نشان دادند مدل تشخیص با استفاده از تکنیک داده‌کاوی SVM و ۹ ویژگی کلیدی به بالاترین دقت یعنی ۸۳/۱۹٪ می‌رسد. از آنجا که SVM از دو تکنیک دیگر (شبکه عصبی و بیز) بهتر عمل می‌کند و دقت قابل‌قبولی را نیز در هر دو آزمایش فراهم می‌کند، بنابراین از میان سه تکنیک برتر شناسایی شده، برای تشخیص بیماری کرونا مناسب‌تر است. بنابراین با توجه به نتایج و دقت به‌دست آمده، می‌توان با استفاده از روش SVM و ۹ ویژگی‌های کلیدی شناسایی شده ابتلا به کوید- ۱۹ را در افراد با ویژگی‌های بالینی مختلف با دقت بالا تشخیص داد.

سپاسگزاری: این مقاله حاصل بخشی از پایان‌نامه با عنوان "تشخیص بیماری کرونا بااستفاده از تکنیک‌های داده‌کاوی" در مقطع کارشناسی رشته مهندسی کامپیوتر-گرایش نرم‌افزار در سال ۱۳۹۹ می‌باشد که با حمایت دانشگاه شهید باهنر کرمان اجرا شده است.

References

1. Srinivas K, Rani BK, Govrdhan A. Applications of data mining techniques in healthcare and prediction of heart attacks. *Int J Comput Sci Eng (IJCSE)* 2010;2(02):250-5.
2. Sepehri MM, Rahnama P, Shadpour P, Teimourpour B. A data mining based model for selecting type of treatment for kidney stone patients. *Tehran Univ Med J* 2009;67(6):421-7.
3. Repalli P. Prediction on Diabetes Using Data mining Approach. *Oklahoma State Univ* 2011.
4. Kajabadi A, Saraee MH, Asgari S. Medical Data Mining: An Approach to Discovery Relationships Among Cardiovascular Risk Factors. 3rd Data Mining Conference Amir Kabir University, 2009.

5. Ghafari H, Amini L, Faraahi A. Stroke Diagnosis Using Data Mining Techniques. 5th Data Mining Conference Amir Kabir University; 2011.
6. Li L, Tang H, Wu Z, Gong J, Gruidl M, Zou J, et al. Data mining techniques for cancer detection using serum proteomic profiling. *Artif Intell Med* 2004;32(2):71-83.
7. Miller D, Weinshenker B, Filippi M, Banwell B, Cohen J, Freedman M, et al. Differential diagnosis of suspected multiple sclerosis: a consensus approach. *Mult Scler J* 2008;14(9):1157-74.
8. Forbes SA, Bindal N, Bamford S, Cole C, Kok CY, Beare D, et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 2011;39(Database issue):D945-50.
9. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, et al. ONCOMINE: a cancer microarray database and integrated data-mining platform. *Neoplasia* 2004;6(1):1-6.
10. SA S. Intelligent heart disease prediction system using data mining techniques. *Int J Healthcare Biomed Res* 2013;1:94-101.
11. Shah S, Kusiak A. Cancer gene search with data-mining and genetic algorithms. *Comput Biol Med* 2007;37(2):251-61.
12. Sigurdardottir AK, Jonsdottir H, Benediktsson R. Outcomes of educational interventions in type 2 diabetes: WEKA data-mining analysis. *Patient Educ Couns* 2007;67(1-2):21-31.
13. Verma L, Srivastava S, Negi PC. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. *J Med Syst* 2016;40(7):178.
14. Chaurasia V, Pal S. Early prediction of heart diseases using data mining techniques. *Caribbean J Sci Tech* 2013;1:208-17.
15. Albahri AS, Hamid RA, Alwan JK, Al-Qays ZT, Zaidan AA, Zaidan BB, et al. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. *J Med Syst* 2020;44:1-1.
16. Kavitha R, Kannan E, editors. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. 2016 international conference on emerging trends in engineering, technology and science (icetets); 2016: IEEE.
17. Paul AK, Shill PC, Rabin MRI, Akhand M, editors. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV); 2016: IEEE. pp. 145-50.

TUMJ Tehran University Medical Journal, January 2022; Vol. 79, No. 10: 822-830 Brief Report

Automated detection of coronavirus disease (COVID-19) by using data-mining techniques: *a brief report*

Seyed Ali Akbar Arabzadeh B.Sc.[1]
Vahid Jamshidi Ph.D.[1*]
Masoud Saeed Ph.D.[1]
Rostam Yazdani M.D.[2]
Mahdieh Jamshidi M.D.[3]

1- Department of Computer Engineering, Faculty of Engineering, Shahid Bahonar University of Kerman, Kerman, Iran.
2- Department of Internal Medicine, Faculty of Medicine, Kerman University of Medical Sciences, Kerman, Iran.
3- Department of Internal Medicine, Faculty of Medicine, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.

[*]Corresponding author: Department of Computer Engineering, Shahid Bahonar University of Kerman, Pajoohesh Sq., Emam Khomeini Highway, Kerman, Iran.
Tel: +98-34-33257501
E-mail: vjamshidi@uk.ac.ir

Abstract

Received: 11 Sep. 2021 Revised: 18 Sep. 2021 Accepted: 14 Dec. 2021 Available online: 22 Dec. 2021

Background: The clinical field has vast sick data that has not been analyzed. Discovering a way to analyze this raw data and turn it into an information treasure can save many lives. Using data mining methods is an efficient way to analyze this large amount of raw data. It can predict the future with accurate knowledge of the past, providing new insights into disease diagnosis and prevention. Several data mining methods exist but finding a suitable one is very important. Today, coronavirus disease (COVID-19) has become one of the causing deadly diseases in the world. The early diagnosis of pandemic coronavirus disease has a significant impact in preventing death. This study aims to extract the key indications of the disease and find the best data mining methods that enhance the accuracy of coronavirus disease diagnosis.

Methods: In this study, to obtain high accuracy in diagnosing COVID-19 disease, a complete and effective workflow over data mining methods was proposed, which includes these steps: data pre-analyzing, indication selection, model creation, the measure of performance, and display of results. Data and related indications of patients with COVID-19 were collected from Kerman Afzalipour Hospital and Rafsanjan, Ali Ebn Abi Taleb Hospital. Prediction structures were made and tested via different combinations of the disease indications and seven data mining methods. To discover the best key indications, three criteria including accuracy, validation and F-value were applied and to discover the best data mining methods, accuracy and validation criteria were considered. For each data mining method, the criteria were measured independently and all results were reported for analysis. Finally, the best key indications and data mining methods that can diagnose COVID-19 disease with high accuracy were extracted.

Results: 9 key indications and 3 data mining methods were obtained. Experimental results show that the discovered key indications and the best-operating data mining method (i.e. SVM) attain an accuracy of 83.19% for the diagnosis of coronavirus disease.

Conclusion: Due to key indications and data mining methods obtained from this study, it is possible to use this method to diagnose coronavirus disease in different people of different clinical indications with high accuracy.

Keywords: data mining, diagnosis, clinical symptoms, coronavirus, COVID-19, pandemics.