تحقق پیش‌بینی زودهنگام بیماری مزمن کلیوی براساس شناسایی موثرترین روش‌های داده‌کاوی

مهدیه جمشیدی[1]، وحید جمشیدی[2*]

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

[*] نویسنده مسئول: کرمان، بزرگراه امام‌خمینی، میدان پژوهش، دانشگاه شهید باهنر کرمان، گروه مهندسی کامپیوتر.
تلفن: ۳۳۲۵۷۵۰۱-۰۳٤
E-mail: vjamshidi@uk.ac.ir

دریافت: ۱٤۰۲/۰۲/۳۱ ویرایش: ۱٤۰۲/۰۳/۰٦ پذیرش: ۱٤۰۲/۰۳/۲۳ آنلاین: ۱٤۰۲/۰٤/۰۱

چکیده

زمینه و هدف: با توجه به اینکه عوامل مختلفی در پیدایش بیماری‌های مزمن کلیوی دخالت دارند این بیماری با علایم مختلف بالینی و آزمایشگاهی ظاهر می‌شود. این تنوع در تعداد و نوع علایم بالینی، پزشک معالج را در بسیاری از موارد گمراه می‌سازد. پژوهش حاضر با هدف استخراج ویژگی‌های کلیدی بیماری و یافتن بهترین روش‌های داده‌کاوی که باعث بهبود صحت تشخیص و پیش‌بینی بیماری کلیوی می‌شود، صورت گرفته است.

روش بررسی: در این مطالعه، اطلاعات و علایم بالینی مربوط به بیماران مبتلا به بیماری کلیوی، از بیمارستان علی‌بن‌ابیطالب رفسنجان از شهریور ۱۳۹۹ تا اسفند ۱٤۰۱ به‌مدت ۳۰ ماه گردآوری شدند. مدل‌های پیش‌بینی با استفاده از ترکیب‌های مختلف از ویژگی‌های بیماری و هفت روش داده‌کاوی، در نرم‌افزار RapidMiner studio ایجاد شده و مورد آزمایش قرار گرفتند.

یافته‌ها: نتایج آزمایش‌های این مطالعه نشان می‌دهد که مدل پیش‌بینی ارایه شده با استفاده از روش بیز و هشت ویژگی کلیدی شناسایی شده (سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، ورم، شب ادراری و وزن مخصوص ادرار)، می‌تواند ابتلا به بیماری کلیوی را در افراد با ویژگی‌های بالینی مختلف با دقت ۹۹/۳۸٪ تشخیص دهد.

نتیجه‌گیری: با توجه به اینکه تشخیص زودهنگام بیماری کلیوی و اتخاذ روش درمان مناسب، می‌تواند از پیشروی آسیب کلیوی جلوگیری کند، از این‌رو در این مطالعه سعی شده است با بکارگیری شیوه‌های نوین آماری و تکنیک‌های هوش مصنوعی در تحقق این هدف گام برداشت. براساس روش پیشنهادی و آزمایشات انجام شده، مهمترین ویژگی‌ها و بهترین روش داده‌کاوی به‌دست آمد و در نهایت پیش‌بینی بیماری کلیوی با دقت بالا میسر گردید.

کلمات کلیدی: بیماری مزمن کلیوی، داده‌کاوی، تجزیه و تحلیل داده، شبکه عصبی.

مقدمه

یک مشکل بهداشت جهانی که به‌طور پیوسته در حال رشد است، بیماری مزمن کلیوی (Chronic kidney disease, CKD) است. بیماری کلیوی علاوه بر خطر مرگ‌ومیر، با خطر بالای بسیاری از بیماری‌های دیگر از جمله بیماری‌های قلبی- عروقی و همچنین هزینه‌های بالای مراقبت‌های بهداشتی همراه است. سالانه بیش از دو میلیون نفر در سراسر جهان برای زنده ماندن تحت درمان دیالیز یا پیوند کلیه قرار می‌گیرند، اما این تعداد تنها ۱۰٪ از افرادی است که برای زندگی به درمان نیاز دارند.[1] شایان ذکر است اکثریت دو میلیون نفری که برای نارسایی کلیه، تحت درمان قرار می‌گیرند تنها در پنج کشور به‌نسبت ثروتمند هستند که ۱۲٪ از جمعیت جهان را تشکیل می‌دهند. درحالی‌که سالانه بیش از یک میلیون نفر در ۱۱۲ کشور کم درآمد به دلیل نارسایی کلیوی درمان نشده، جان خود را از دست می‌دهند که

دلیل آن بار مالی عظیم درمان دیالیز یا پیوند کلیه است.[۱] در واقع بیماری کلیوی یکی از بیماری‌های خاموش است که در صورت عدم تشخیص و درمان مناسب، پیشروی آسیب کلیوی ادامه یافته و ممکن است در نهایت به آخرین مرحله‌ی نارسایی کلیه منتهی شود که در این وضعیت، فرد بدون انجام فیلتراسیون مصنوعی (دیالیز) یا دریافت پیوند کلیه، جان خود را از دست خواهد داد. بنابراین، تشخیص زودهنگام، کنترل و مدیریت بیماری اهمیت زیادی دارد.

بیماری مزمن کلیوی (CKD) دارای پنج مرحله مختلف رشد است. شدت هر مرحله با پیشرفت از مرحله یک به مرحله پنج افزایش می‌یابد. در مرحله یک، فرد می‌تواند عملکرد زیر نرمال کلیه داشته باشد و حتی کاهش جزیی در عملکرد کلیه را تجربه کند. در طول مرحله ۲، فرد می‌تواند از دست دادن جزیی تا متوسط عملکرد کلیه را تجربه کند. مرحله ۳ بیشتر تشدید می‌شود و فرد با کاهش متوسط تا شدید عملکرد کلیه مواجه می‌شود. در مرحله ٤، فرد دچار کاهش شدید عملکرد کلیه می‌شود. در مرحله ۵، فرد نارسایی کامل کلیه را تجربه خواهد کرد.

با توجه به اینکه عوامل مختلفی در پیدایش بیماری‌های مزمن کلیوی دخالت دارند این بیماری با علایم مختلف بالینی و آزمایشگاهی ظاهر می‌شود. این تنوع در تعداد و نوع علایم بالینی، پزشک معالج را در بسیاری از موارد گمراه می‌سازد.

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 باشند تا مهمترین علایم بالینی شناسایی شده و تشخیص و درمان مناسب زودهنگام صورت گیرد و از بروز عواقب وخیم کلیوی جلوگیری شود.[۲]

روش‌های متنوعی توسط پژوهشگران در حوزه داده‌کاوی به‌منظور دستیابی به علایم و تشخیص بیماری‌ها، معرفی شده‌اند.[۳] Baby و همکاران نشان دادند که روش‌های داده‌کاوی می‌توانند به‌طور موثر در کاربردهای پزشکی مورد استفاده قرار گیرند. آنها در مطالعه خود از داده‌های مربوط به بیماران مبتلا به بیماری‌های کلیوی استفاده کردند.[٤]

تعدادی از مطالعات سعی در پیش‌بینی CKD با استفاده از ابزارهای داده‌کاوی کرده‌اند.[۵-۷] به‌عنوان مثال، در مطالعه‌ای از تکنیک‌های شبکه عصبی (ANN) و الگوریتم‌های پشتیبان ماشین‌بردار (SVM) برای تشخیص سنگ کلیه استفاده شده است.[۵] نتایج تجربی آنها نشان داده است که بکارگیری ANN بهبود قابل‌توجهی در دقت تشخیص بسیاری از بیماری‌ها دارد.[٦]

تحقیقاتی برای شناسایی بیماری کلیوی انجام شده است و مقایسه‌ای بین ماشین بردار پشتیبان (SVM) و بیز صورت گرفته است.[۷] داده‌های مورد استفاده خروجی آزمون عملکرد کلیه مصنوعی (Synthetic kidney function test, KFT) بودند و نتایج آزمایشات نشان داد دقت SVM بالاتر از بیز است. Kumar و همکاران، از مدل‌های شبکه‌های عصبی مصنوعی برای پیش‌بینی سنگ کلیه استفاده کرده‌اند.[۸] در این مطالعه، سه الگوریتم شبکه عصبی مختلف استفاده شده، و براساس دقت، زمان برای توسعه مدل و اندازه مجموعه داده‌های آموزش، مورد مقایسه قرار گرفتند.

تکنیک‌های داده‌کاوی بیز (NB) و ماشین‌های بردار پشتیبان (SVM) برای پیش‌بینی بیماری کلیوی استفاده شده است.[۹] در این مقاله نویسندگان دریافتند که عملکرد SVM بهتر از NB است. همچنین این نتیجه حاصل شده است که با در نظر گرفتن پنج مرحله CKD، این بیماری می‌تواند در یک فرآیند غیرقابل برگشت، از مرحله‌ای به مرحله دیگر پیشرفت کند. بنابراین تشخیص زودهنگام بیماری بسیار مهم است. همچنین در تعدادی از مطالعات سعی در پیش‌بینی CKD با استفاده از الگوریتم‌های مختلف یادگیری ماشین کرده‌اند و بیان کردند که در این الگوریتم‌ها برای داشتن دقت بالای ۹۵٪ به میلیون‌ها رکورد با مقادیر گمشده صفر نیاز است.[۱۰و۱۱]

Kunwar و همکاران در مطالعه خود، بیماری مزمن کلیه (CKD) را با استفاده از بیز (Naïve Bayesian) و شبکه عصبی مصنوعی (ANN) پیش‌بینی کردند.[۱۲] نتایج آنها نشان داد که بیز نتایج دقیق‌تری نسبت به شبکه‌های عصبی مصنوعی تولید می‌کند. همچنین مشاهده شد که الگوریتم‌های طبقه‌بندی مختلفی می‌توانند برای بررسی و شناسایی CKD مورد استفاده قرار گیرند. Tangri و همکاران، مدل‌های پیش‌بینی CKD را در سیزده مقاله مورد نقد و بررسی قرار داده‌اند و مدلی را برای پیش‌بینی پیشرفت CKD پیشنهاد کرده‌اند که از روش رگرسیون استفاده می‌کند.[۱۳] یک رویکرد داده‌کاوی نیز برای CKD ارایه شده است.[۱٤] رویکرد پیشنهادی، داده‌ها را از کارهای منتشر شده قبلی استخراج می‌کند. استفاده از تکنیک‌های داده‌کاوی برای پیش‌بینی بقای دیالیز کلیه نیز مورد بحث قرار گرفته است.[۱۵] سه

مدل تحلیل پیش‌بینی، استفاده و مقایسه شده‌اند. برای کمک به تشخیص بیماری، برخی از نویسندگان مدل‌های تحلیلی را برای پیش‌بینی مراحل مختلف CKD استفاده کرده‌اند. یک مدل تحلیلی پیش‌بینی‌کننده مبتنی بر متابولومیک برای شناسایی مراحل CKD ارایه شده است.[۱۶] یک شاخص جدید نیز برای پیش‌بینی مراحل CKD با دقت ۸۱/۳۰٪ معرفی شده است. یک مطالعه موردی برای پیش‌بینی CKD در یک بیمارستان محلی در انگلستان انجام شده است.[۱۷] این مطالعه دو مرحله اصلی یعنی چهار و پنج را در نظر گرفته است: CKD متوسط تا شدید و مرحله نهایی نارسایی کلیه. مدل‌های پیش‌بینی، مبنایی برای شناسایی بالقوه بیماران پرخطر فراهم می‌کنند.

محققان رابطه بین خوشه‌های علایم و کیفیت زندگی بیمار را با تمرکز بر مراحل دو تا چهار بیماری کلیوی، مورد مطالعه قرار دادند.[۱۸] یک مدل پیش‌بینی نیز با استفاده از مجموعه شبکه‌های عصبی مصنوعی برای شناسایی مرحله پایانی (مرحله ۵) بیماری کلیوی معرفی شده است.[۱۹]

البته باید به این نکته توجه داشت که کارایی تکنیک‌های داده‌کاوی، تا حد زیادی به تکنیک مورد استفاده و ویژگی‌های انتخاب شده بستگی دارد. نکته مهم دیگر این است که مجموعه داده‌های پزشکی در حوزه بهداشت و درمان، بسیار زیاد و گهگاه 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.[۲۲-۲۰] 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.[۲۳و۲۴] بنابراین، جهت دستیابی به دقت بالا در شناسایی بیماری کلیوی با استفاده از روش‌های داده‌کاوی، افزون‌بر 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 کلیوی با دقت بالا بپردازند. شایان ذکر است در این مطالعه، از هفت روش معروف داده‌کاوی برای پیش‌بینی بیماری کلیوی و ویژگی‌های کلیدی استفاده شده است.

روش بررسی

طرح حاضر با کد اخلاق IR.RUMS.REC.1401.117 به تصویب کمیته اخلاق دانشکده پزشکی، دانشگاه علوم پزشکی رفسنجان رسیده است. در این مطالعه، داده‌های بیماری کلیوی از بیمارستان علی‌ابن‌ابیطالب رفسنجان از شهریور ۱۳۹۹ تا اسفند ۱۴۰۱ به‌مدت ۳۰ ماه گردآوری شده‌اند. مجموعه داده‌ای تنظیم شده، شامل ۲۵۳ رکورد می‌باشد.

با بررسی پرونده‌های بیماران مبتلا به بیماری کلیوی و آنالیز آنها، جدول ۱ استخراج شده است. همانگونه که در جدول ۱ نشان داده شده است ۱۶ ویژگی در نظر گرفته شده که ۱۵ ویژگی آن مربوط به تشخیص بیماری کلیوی و یک ویژگی به‌عنوان خروجی (ویژگی پیش‌بینی شده) برای نشان دادن شدت بیماری کلیوی در فرد می‌باشد. ویژگی «خروجی» عددی از صفر تا پنج است که «صفر» برای نشان دادن عدم وجود بیماری کلیوی و مقادیر یک تا پنج برای نشان دادن شدت بیماری هستند (پنج بیشترین شدت بیماری است).

در این تحقیق، برای انجام آزمایش از RapidMiner studio 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.

روند کار بدین صورت است که مجموعه داده‌ای تهیه شده از بیمارستان علی‌ابن‌ابیطالب به RapidMiner داده می‌شود. فرآیند داده‌کاوی از مرحله پیش‌پردازش آغاز می‌شود، سپس 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 پیاده سازی مدل برای شناسایی بیماری می‌شود.

در فرآیند داده‌کاوی تشخیص کلیوی، زیر مجموعه‌ای حاوی حداقل سه ویژگی از ۱۵ ویژگی انتخاب می‌شود سپس ایجاد مدل برای هر ترکیب از آنها، انجام می‌شود. عملکرد هر مدل ساخته شده، براساس ویژگی‌ها و تکنیک داده‌کاوی در طول هر تکرار، ثبت می‌شود و

جدول ۱: شرح ویژگی‌های مجموعه داده‌ای بیمارستان علی‌بن‌ابیطالب

ویژگی	نوع	توضیحات
سن	عددی	سن بیمار به سال
آلبومینوری	اسمی	عدد دو: مقادیر آلبومین ادرار بیش از ۵ g عدد یک: مقادیر آلبومین ادرار تا ۵ g عدد صفر: تراس
اورمی	اسمی	عدد یک: افزایش اوره خون عدد صفر: عدم افزایش اوره خون
ورم	اسمی	عدد یک: افزایش وزن روزانه عدد صفر: عدم افزایش وزن روزانه
بیوپسی کلیه	اسمی	عدد یک: داشتن نتایج غیرطبیعی عدد صفر: داشتن نتایج طبیعی
سدیمانتاسیون	اسمی	عدد یک: وجود التهاب عدد صفر: عدم وجود التهاب
نارسایی قلب	اسمی	عدد یک: داشتن نارسایی قلبی عدد صفر: نداشتن نارسایی قلبی
تب	اسمی	عدد یک: داشتن تب عدد صفر: نداشتن تب
فشارخون	اسمی	عدد یک: افزایش فشارخون‌های سیستولی و دیاستولی عدد صفر: عدم افزایش فشارخون‌های سیستولی و دیاستولی
وزن مخصوص ادرار	اسمی	عدد یک: داشتن کاهش وزن مخصوص ادرار عدد صفر: نداشتن کاهش وزن مخصوص ادرار
کم‌خونی	اسمی	عدد یک: داشتن کم خونی عدد صفر: نداشتن کم خونی
آدیس کانت	عددی	مقدار موجود در نتایج آزمایش‌ها
شب ادراری	اسمی	عدد یک: داشتن شب ادراری عدد صفر: نداشتن شب ادراری
رادیوگرافی کلیه‌ها	اسمی	عدد یک: داشتن تغییرات در شکل و اندازه کلیه‌ها عدد صفر: نداشتن تغییرات در شکل و اندازه کلیه‌ها
الکتروفورز خون	عددی	بررسی نوع و میزان هموگلوبین موجود در خون
خروجی	اسمی	تشخیص بیماری کلیوی صفر به معنای عدم وجود و یک به معنای وجود

نتیجه پس از اتمام کل فرآیند نشان داده می‌شود.

شرح مراحل فرآیند پیشنهاد شده برای آزمایش‌ها به‌صورت زیر است:

پیش‌پردازش داده‌ها: داده‌ها پس از جمع‌آوری، پیش‌پردازش می‌شوند. در مجموعه داده‌ای بیمارستان علی‌ابن‌ابیطالب، مقدار ویژگی «خروجی» برای داشتن بیماری کلیوی از رکوردهای چند مقداری (صفر برای عدم وجود و ٤–١ برای وجود) به رکوردهای دو مقداری (صفر برای عدم وجود و یک برای وجود بیماری کلیوی) تبدیل می‌شود. کار پیش‌پردازش داده‌ها با تبدیل تمام مقادیر تشخیص (دو، سه و چهار) به یک انجام می‌شود. مجموعه داده‌ای حاصل فقط شامل مقادیر صفر و یک است که صفر معرف فقدان و یک معرف ابتلا بیمار به بیماری کلیوی است.

انتخاب ویژگی: از میان ۱۵ ویژگی مورد استفاده در تشخیص بیماری کلیوی، فقط «سن» به اطلاعات شخصی هر بیمار اشاره دارد. ١٤ ویژگی باقیمانده شامل خصوصیات بالینی جمع‌آوری شده از معاینات پزشکی می‌باشند.

در انجام آزمایش، ترکیبی از ویژگی‌ها برای استفاده در هفت روش داده‌کاوی (k-nearest neighbors algorithm (k-NN), decision tree, Bayes, regression, voting, Support vector machines (SVM) and neural network) مورد استفاده قرار می‌گیرد. ابتدا تمام ترکیبات احتمالی سه ویژگی از ۱۵ 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 چهار ویژگی از ۱۵ ویژگی انجام می‌شود. تعداد کل ترکیب‌های قابل دستیابی از مجموعه‌ای با ۱۵ ویژگی، به استثنای مجموعه تهی، ۱–۲^n است. در این تحقیق، یک زیر مجموعه واحد از ویژگی‌ها نمی‌تواند کمتر از سه ویژگی داشته باشد. بنابراین، تمام زیر مجموعه‌های ترکیبی که با داشتن یک یا دو ویژگی حاصل می‌شوند، حذف می‌شوند. معادله‌ای که برای محاسبه تعداد کل ترکیب‌ها حاصل می‌شود عبارت است از:

تعداد کل ترکیب‌ها (۱)

$$2^n - \left(\frac{n!}{1!(n-1)!}\right) - \left(\frac{n!}{2!(n-2)!}\right) - 1$$

$$2^n - n - \frac{n(n-1)}{2} - 1$$

$$2^n - \left(\frac{2n+n^2-n}{2} + 1\right)$$

$$2^n - \left(\frac{n^2-n}{2} + 1\right)$$

n تعداد کل ویژگی‌هایی است که برای تولید زیر مجموعه‌های ترکیبی به‌کار می‌رود که در این آزمایش ۱۵ است. بنابراین، در مجموع ۳۲٦٦۲ ترکیب از ویژگی‌ها در این آزمایش انتخاب و آزمایش شده‌اند.

ایجاد مدل: پس از انتخاب ویژگی‌ها، مدل‌ها با استفاده از هفت 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 10–Folds Cross Validation 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.

سه معیار برای ارزیابی کارایی مدل‌ها وجود دارد: دقت (Accuracy)، صحت (Precision) و معیار F (F-measure).

منظور از «دقت» درصد موارد صحیح تشخیص داده شده در بین همه موارد است و از رابطه ۲ به‌دست می‌آید.

$$\text{Accuracy}= \frac{TP+TN}{TP+TN+FP+FN} \quad (۲$$

«معیار F» میانگین وزنی دقت و فراخوانی است و از رابطه ۵ به‌دست می‌آید.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (۳$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (٤$$

$$\text{F-Measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} \quad (۵$$

پارامترهای ذکر شده در روابط (۲) تا (۵) به شرح زیر، تعریف می‌گردند:

TP: تعداد رکوردهایی که به‌درستی، مثبت تشخیص داده می‌شوند.

TN: تعداد رکوردهایی که به‌درستی، منفی تشخیص داده می‌شوند.

FP: تعداد رکوردهایی که به غلط، مثبت تشخیص داده می‌شوند.

FN: تعداد رکوردهایی که به غلط، منفی تشخیص داده می‌شوند.

برای شناسایی بهترین ویژگی‌های کلیدی، از سه معیار دقت، صحت و معیار F استفاده شده است، و برای شناسایی تکنیک‌های داده‌کاوی برتر برای ایجاد بهترین مدل، معیارهای دقت و صحت مورد استفاده قرار گرفته‌اند.

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 کلیوی فراهم کند زیرا دقت و صحت بهترین معیارهای ارزیابی کارایی هستند.

برای هر روش داده‌کاوی، معیارها به‌صورت جداگانه اندازه گیری شدند و تمام نتایج برای تجزیه و تحلیل ثبت شدند.

یافته‌ها

جدول ۲ کارایی هفت تکنیک داده‌کاوی را در سه معیار مختلف نشان می‌دهند. براساس جدول ۲ سه تکنیک برتر بیز، شبکه عصبی، و SVM هستند که دقتی بیش از ۸۵٪ دارند. همچنین نتایج نشان می‌دهند که دو روش درخت تصمیم و k-NN کمترین میزان دقت را نسبت به سایر تکنیک‌ها دارند. این در حالی است که براساس جدول ۲، روش درخت تصمیم بالاترین میزان صحت را دارا می‌باشد.

براساس تجزیه و تحلیل نشان داده شده در جداول، می‌توانیم مشاهده کنیم که بالاترین میزان دقت (۹۹/۳۸٪) توسط بیز با ۹ ویژگی انجام شده است. از طرف دیگر، بالاترین میزان صحت (۱۰۰/۰۰٪) با استفاده از ۳ ویژگی (یعنی سن، بیوپسی کلیه، آلبومینوری) با درخت تصمیم حاصل شده است، درحالی‌که بیشترین میزان F (۹۸/۷۸٪) با بیز با هفت ویژگی انجام شده است. با استفاده از هر روش داده‌کاوی، میانگین دقت، میانگین صحت، و میانگین F بر روی ۳۲٦٦۲ ترکیب مختلف از ویژگی‌ها نیز محاسبه شده است. بالاترین میانگین دقت در روش‌های داده‌کاوی مربوط به بیز، شبکه عصبی و SVM بترتیب با داشتن میانگین ۹٦/۱۵٪ و ۸۵/۱۷٪ و ۸٤/۱۵٪ می‌باشد. بالاترین میانگین صحت مربوط به بیز، SVM و شبکه عصبی به‌ترتیب با داشتن میانگین ۹۵/۷۱٪ و ۷۷/۳٦٪ و ۷٦/۲۵٪ می‌باشد. درحالی‌که بالاترین میانگین F مربوط به بیز، SVM و رگرسیون به‌ترتیب با داشتن میانگین

۹۵/۶۸٪ و ۹۳/۳۵٪ و ۷۷/۲۷٪ 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 کلیوی 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 ۳۲۶۶۲ ترکیب ویژگی‌ها، ترکیب‌هایی که منجر به بالاترین کارایی یک تکنیک خاص شده‌اند، مشخص شدند.

جدول ۳ 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 ۳ نشان می دهد که چند بار هر یک از این ویژگی‌ها در بین ترکیب‌هایی که به بالاترین دقت در بین هفت 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 ۱۵ ویژگی، «آلبومینوری»، 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 ۱۵ بار ظاهر شده‌اند و بیشترین کارایی را داشته‌اند، به‌عنوان ویژگی‌های کلیدی در تشخیص بیماری کلیوی شناخته شده‌اند.

براساس آنالیز جدول ۳، هشت ویژگی به‌عنوان ویژگی‌های مهم در تشخیص بیماری کلیوی شناخته شده‌اند: سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، ورم، شب ادراری و وزن مخصوص ادرار.

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 جدول ۲ نیز مورد بررسی قرار گرفتند و تایید شدند.

باید توجه داشت که از میان هشت ویژگی، هفت مورد از آنها ویژگی‌های بالینی جمع‌آوری 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 کلیوی موثر هستند.

برای سنجش اعتبار ویژگی‌های کلیدی، یافته‌ها از طریق دانش پزشکی نیز مورد بررسی قرار گرفتند. شکل ۱، نمودار نتایج پزشکی حاصل از بررسی علایم بالینی و آزمایشگاهی به‌دست آمده از بیمارهای کلیوی را نشان می‌دهد. نکته قابل توجه این است که اطلاعات این نمودار، نتایج داده کاوی را تایید می‌کند.

آلبومینوری، شایعترین و در بسیاری از موارد زودرس‌ترین علامت آزمایشگاهی در بیماری‌های کلیوی می‌باشد. آلبومینوری همیشه یک علامت ثابت بیماری است. شدت پروتیینوری و ترکیب پروتیینوری در ادرار ارتباط مستقیم با مکانیسم صدمه کلیوی دارد. در افراد نرمال کمتر از ۱۵۰ mg/d و کمتر از ۳۰ mg/d آلبومین دفع می‌شود. اگر سد گلومرولی تخریب شود نشت پروتیین‌های پلاسما به ادرار روی خواهد داد و با اندازه‌گیری آن، آسیب کلیوی قابل تشخیص است.

گاهی از ابتدای بیماری، مقادیر قابل‌ملاحظه‌ای پروتیین از ادرار دفع می‌گردد و بالاخره زمانی وجود آلبومین در ادرار در مراحل اشتداد بیماری و یا هنگامی که بیماری کلیوی به‌ویژه از نوع گلومرولونفریت به مرحله حاد یا مزمن می‌رسد پدیدار می‌گردد.

وجود آلبومین حقیقی در ادرار حتی اگر همراه با از بین رفتن سایر علایم بیماری کلیه نیز باشد معرف وجود قطعی بیماری کلیوی که در واقع همان خراب شدن مامبران بازال (Basal membranes) باشد خواهد بود. بنابراین اگر بیماری یک آلبومین پایدار و ثابت داشته باشد وجود ضایعات مزمن یا حاد کلیه به‌صورت حتم وجود دارد و هیچگاه این کلیه از نظر بالینی و آسیب‌شناسی التیام یافته تلقی نمی‌گردد.

اورمی یکی دیگر از علایم مهم در تشخیص بیماری‌های کلیوی تلقی می‌گردد. اورمی در بسیاری از موارد از همان ابتدای بیماری ظاهر می‌شود (مانند اشکال متعدد گلومرولونفریت حاد) و در مواردی نیز

در مراحل مزمن و حاد بیماری پدیدار می‌گردد. ورم یکی از علایم بالینی در تشخیص بیماری‌های کلیوی بشمار آید. ورم کلیوی از نواحی صورت و پلک شروع می‌شود ولی این اصل همیشه عمومیت ندارد و گاهی تشخیص آن فقط با تعیین وزن روزانه بیماران امکان‌پذیر می‌شود.

نتایج حاصل از بیوپسی کلیه کمک شایانی در تشخیص بیماری کلیه می‌کند و با استفاده از آن می‌توان حتی ضایعات لوپوسی، مولتیپل میلوما و آمیلوز کلیه را کشف کرد.

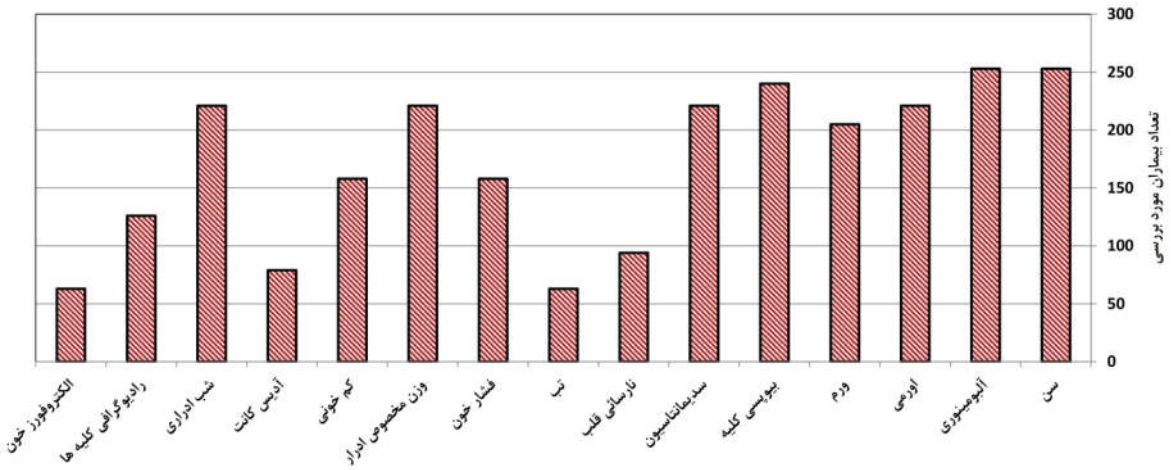
سدیمانتاسیون در ضایعات حاد، لوپوس، مولتیپل میلوما و ضایعات عفونی کلیه‌ها افزایش می‌یابد. ازاین‌رو افزایش سدیمانتاسیون

جدول ۲: بهترین ترکیب ویژگی‌ها و بالاترین میزان دقت، صحت، و معیار F در هر روش داده‌کاوی

تکنیک	بهترین ترکیب ویژگی‌ها برای داشتن بالاترین دقت	دقت	بهترین ترکیب ویژگی‌ها برای داشتن بالاترین صحت	صحت	بهترین ترکیب ویژگی‌ها برای داشتن بالاترین معیار F	معیار F
بیز	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، کم خونی، ورم، نارسایی قلب، شب ادراری	٪۹۹/۳۸	سن، بیوپسی کلیه، ورم، آلبومینوری، نارسایی قلب	٪۹۸/۷۵	سن، یوپسی کلیه، کم‌خونی، آلبومینوری، ورم، نارسایی قلب، شب ادراری	٪۹۸/۷۸
شبکه عصبی	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، کم خونی، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، آدیس کانت	٪۸۸/۷٤	سن، بیوپسی کلیه، شب ادراری	٪۸۲/٦۷	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، کم خونی، نارسائی قلب، ورم، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، فشارخون، آدیس کانت	٪۷۹/۵۹
SVM	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها	٪۸۷/۵۸	سن، بیوپسی کلیه، اورمی، آلبومینوری، رادیوگرافی کلیه‌ها	٪۸۵/٦۲	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، کم خونی، ورم، نارسایی قلب، شب ادراری	٪۹۷/٦۸
رگرسیون	سن، بیوپسی کلیه، ورم، سدیمانتاسیون، آلبومینوری، نارسایی قلب، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، الکتروفورز خون	٪۷۳/٦٤	سن، بیوپسی کلیه، نارسایی قلب، کم‌خونی، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، شب ادراری، فشارخون، الکتروفورز خون	٪۷۳/٤۸	سن، بیوپسی کلیه، کم‌خونی، آلبومینوری، ورم، نارسایی قلب، شب ادراری، رادیوگرافی کلیه‌ها، الکتروفورز خون	٪۸۲/٤۷
رأی‌گیری	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، کم خونی، ورم، نارسایی قلب، رادیوگرافی کلیه‌ها، تب	٪۷۳/۷۵	سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، کم خونی، رادیوگرافی کلیه‌ها، نارسایی قلب، تب، شب ادراری، تب	٪۹۷/٦۳	کم‌خونی، ورم، نارسایی قلب، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، فشارخون، تب	٪٦۹/٤٤
درخت تصمیم	سن، فشارخون، ورم، سدیمانتاسیون، آلبومینوری، نارسایی قلب، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها	٪٦۱/۸۳	سن، بیوپسی کلیه، آلبومینوری	٪۱۰۰/۰۰	اورمی، نارسایی قلب، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، فشارخون	٪۵۸/۳٦
k-NN	سن، فشارخون، ورم، سدیمانتاسیون، آلبومینوری، نارسایی قلب، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها	٪۵٤/۹۲	سن، بیوپسی کلیه، اورمی	٪٦٦/۵۸	اورمی، نارسایی قلب، شب ادراری، وزن مخصوص ادرار، رادیوگرافی کلیه‌ها، فشارخون	٪٦۸/۵۷

جدول ۳: مقایسه ویژگی‌ها براساس تعداد دفعات رخداد در روش‌های داده‌کاوی با بالاترین میزان کارایی

ویژگی / رخداد	سن	آلبومینوری	اورمی	ورم	بیوپسی کلیه	سدیمانتاسیون	نارسایی قلب	تب	فشار خون	ادرار وزن مخصوص	کم‌خونی	آدیس کانت	شب ادراری	رادیوگرافی	الکتروفورز خون
رخداد با بالاترین میزان دقت	۷	۷	۷	٦	۷	۷	٥	۲	٥	۷	٥	۳	۷	٥	۳
رخداد با بالاترین میزان F	۷	۷	۷	٦	۷	۷	٤	۲	٦	۷	٦	٤	۷	٦	۳
رخدا با بالاترین میزان صحت	٥	٦	٤	۳	٦	٤	۲	۱	۳	٤	۳	۲	٤	۲	۲
تعداد کل رخدادها	۱۹	۲۰	۱۸	۱٥	۲۰	۱۸	۱۱	٥	۱٤	۱۸	۱٤	۹	۱۸	۱۳	۸



شکل ۱: اهمیت ویژگی‌های استخراج شده از دانش پزشکی جهت مقایسه با ویژگی‌های استخراج شده از روش‌های داده‌کاوی

یکی از علایم ارزشمند در تشخیص بیماری کلیوی محسوب می‌شود. چون نشان می‌دهد که سد فیلتراسیون خراب شده و قادر نیست از ورود اجزاء خون به ادرار جلوگیری کند. کم شدن وزن مخصوص ادرار یک علامت ثابت و پایدار در تفسیر بیماری مزمن کلیوی بشمار می‌آید. هر قدر وزن مخصوص ادرار کمتر باشد حدس نارسایی کلیه بیشتر و شدیدتر خواهد بود. شب ادراری علامت بسیار مهم و با ارزش در نارسایی مزمن کلیه می‌باشد. چه بسا اغلب بیماران مبتلا به نارسایی کلیه در همان ابتدا که اوره خون و آزمایش ادرار آنها طبیعی است دچار شب ادراری می‌شوند. که علت آن کاهش قدرت تغلیظ ادرار توسط توبول‌های کلیه است و نشان‌دهنده آسیب کلیه می‌باشد.

نتایج نشان دادند، دقت مدل‌های تشخیص با استفاده از هشت ویژگی کلیدی بهتر از مدل‌های با استفاده از ۱۵ ویژگی کلیدی است. بالاترین دقت (۹۹/۱۲٪) برای مدل داده‌کاوی با ۱۵ ویژگی مربوط به بیز می‌باشد و همچنین بالاترین دقت (۹۹/۳۸٪) برای مدل داده‌کاوی با هشت ویژگی نیز مربوط به بیز می‌باشد.

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. هرچند 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، زیرا کارایی و دقت تصمیمات زمانی که انسان‌ها با حجم و فشار کاری مواجه هستند کاهش می‌یابد. امروزه بیماری کلیوی 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 سرعت و دقت بالا در تشخیص بیماری کلیوی، در ابتدا یک فرآیند کامل و موثر برای استفاده از روش‌های داده کاوی پیشنهاد گردید که شامل مراحل 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 داده کاوی، معیارها به صورت جداگانه اندازه‌گیری شدند و تمام نتایج برای تجزیه و تحلیل ثبت شدند. سرانجام ویژگی‌های کلیدی و بهترین روش‌های داده‌کاوی که باعث تشخیص بیماری کلیوی با دقت بالا می‌شوند، استخراج شدند.

آزمایش برروی مجموعه داده‌ای استخراج شده از بیمارستان علی‌ابن‌ابیطالب رفسنجان با استفاده از فرآیند پیشنهاد شده انجام شد. بالاترین میزان دقت (۹۹/۳۸٪) توسط بیز با هشت ویژگی به‌دست آمد. از طرف دیگر، بالاترین میزان صحت (۱۰۰/۰۰٪) با استفاده از سه ویژگی (یعنی سن، بیوپسی کلیه، آلبومینوری) با درخت تصمیم حاصل شد، در حالی که بیشترین میزان F (۹۸/۷۸٪) با بیز با هفت ویژگی به‌دست آمد. همچنین نتایج نشان دادند که دو روش درخت تصمیم و k-NN کمترین میزان دقت را نسبت به سایر تکنیک‌ها دارند و روش درخت تصمیم دارای بالاترین میزان صحت می‌باشد.

همچنین با اعمال تکنیک‌های داده‌کاوی بر روی ۳۲۶۶۲ ترکیب مختلف از ویژگی‌ها، مشاهده گردید بالاترین میانگین دقت در روش‌های داده‌کاوی مربوط به بیز، شبکه عصبی و SVM به‌ترتیب با داشتن میانگین ۹۶/۱۵٪ و ۸۵/۱۷٪ و ۸٤/۱۵٪ می‌باشد. بالاترین میانگین صحت مربوط به بیز، SVM و شبکه عصبی به‌ترتیب با داشتن میانگین ۹۵/۷۱٪ و ۷۷/۳٦٪ و ۷٦/۲۵٪ می‌باشد. درحالی‌که بالاترین میانگین F مربوط به بیز، SVM و رگرسیون بترتیب با داشتن میانگین ۹۵/٦۸٪ و ۹۳/۳۵٪ و ۷۷/۲۷٪ 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: بیز، SVM و شبکه عصبی.

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، هشت ویژگی به‌عنوان ویژگی‌های مهم در تشخیص بیماری کلیوی شناسایی شدند: سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، ورم، شب ادراری و وزن مخصوص ادرار. یافته‌ها از طریق دانش پزشکی مورد بررسی قرار گرفتند و هشت ویژگی کلیدی استخراج شده، تایید شدند. ویژگی‌های کلیدی به‌دست آمده از دانش پزشکی نیز عبارتند از: سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، ورم، شب ادراری و وزن مخصوص ادرار.

سه تکنیک برتر داده‌کاوی عبارتنداز: شبکه عصبی، بیز و SVM. همچنین نتایج نشان دادند مدل تشخیص با استفاده از تکنیک داده‌کاوی بیز و هشت ویژگی کلیدی به بالاترین دقت یعنی ۹۹/۳۸٪ می‌رسد. از آنجا که بیز از دو تکنیک دیگر (شبکه عصبی و SVM) بهتر عمل می‌کند و دقت قابل قبولی را فراهم می‌کند، بنابراین از میان

سه تکنیک برتر شناسایی شده، برای تشخیص بیماری کلیوی مناسب‌تر است.

محدودیت‌های اصلی مطالعه به شرح زیر است: ۱- اطلاعات بعضی از پرونده‌های بیماران مبتلا به بیماری کلیوی ناقص هستند که پس از بررسی و شناسایی کنار گذاشته شدند. ۲- مدل‌ها براساس پرونده‌های بیماران گروه سنی ٤٠ سال به بالا بودند که ممکن است تعمیم نتایج را به گروه سنی وسیعتر محدود کند. ۳- هرچند یک روش جامع و دقیق ارایه شده است اما مدل‌ها، تنها بر اطلاعات پرونده‌های بیماران مبتلا به بیماری کلیوی بیمارستان علی‌ابن‌ابیطالب رفسنجان تکیه داشتند. ٤- پارامتر اقلیم در مجموعه داده تحت بررسی، برای کشف روابط پنهانی این پارامتر با بیماری کلیوی در نظر گرفته نشده است.

یافته‌های این مطالعه نشان می‌دهد که با استفاده از روش بیز و هشت ویژگی کلیدی شناسایی شده (سن، بیوپسی کلیه، اورمی، سدیمانتاسیون، آلبومینوری، ورم، شب ادراری و وزن مخصوص ادرار)، می‌توان ابتلا به بیماری کلیوی را در افراد با ویژگی‌های بالینی مختلف با دقت بالا تشخیص داد.

سپاسگزاری: این مقاله حاصل بخشی از طرح تحقیقاتی تحت عنوان "تشخیص خودکار بیماری‌ها با استفاده از روش‌های نوین آماری و شبکه عصبی" مصوب دانشگاه علوم پزشکی و خدمات بهداشتی و درمانی رفسنجان در سال ۱۳۹۹ و کد ٦٥٤٥٢ می‌باشد که با حمایت دانشگاه علوم پزشکی و خدمات بهداشتی و درمانی رفسنجان انجام شده است.

References

1. Couser WG, Remuzzi G, Mendis S, Tonelli M. The contribution of chronic kidney disease to the global burden of major noncommunicable diseases. *Kidney international* 2011;80(12):1258-70.
2. Fereydouni T, Hajihashemi S, Yousefichaijan P, Rahbari A. Protective Effects of Vitamin C Concomitant Treatment on Deferasirox-induced Renal Toxicity in Rats. *Journal of Arak University of Medical Sciences* 2020;23(6):926-43.
3. Arabzadeh SA, Jamshidi V, Saeed M, Yazdani R, Jamshidi M. Automated detection of coronavirus disease (COVID-19) by using data-mining techniques: a brief report. *Tehran University Medical Journal TUMS Publications* 2021;79(10):822-30.
4. Baby PS, Vital TP. Statistical analysis and predicting kidney diseases using machine learning algorithms. *International Journal of Engineering Research and Technology* 2015;4(7):206-10.
5. Sinha P, Sinha P. Comparative study of chronic kidney disease prediction using KNN and SVM. *International Journal of Engineering Research and Technology* 2015;4(12):608-12.
6. Vijayarani S, Dhayanand S, Phil M. Kidney disease prediction using SVM and ANN algorithms. International *Journal of Computing and Business Research (IJCBR)* 2015;6(2):1-2.
7. Abhishek GS, Gupta D. Proposing efficient neural network training model for kidney stone diagnosis. *International Journal of Computer Science and Information Technologies* 2012;3(3):3900-4.
8. Kumar K, Abhishek B. Artificial neural networks for diagnosis of kidney stones disease. *Germany: GRIN Verlag* 2012; 7(3), 20-25.
9. Ramya S, Radha N. Diagnosis of chronic kidney disease using machine learning algorithms. *International Journal of Innovative Research in Computer and Communication Engineering* 2016;4(1):812-20.
10. Salekin A, Stankovic J. Detection of chronic kidney disease and selecting important predictive attributes. *In2016 IEEE International Conference on Healthcare Informatics (ICHI)* 2016 (pp. 262-270). IEEE.
11. Jena L, Kamila NK. Distributed data mining classification algorithms for prediction of chronic-kidney-disease. *International Journal of Emerging Research in Management &Technology* 2015;4(11):110-8.
12. Kunwar V, Chandel K, Sabitha AS, Bansal A. Chronic Kidney Disease analysis using data mining classification techniques. *In2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)* 2016 (pp. 300-305). Ieee.
13. Tangri N, Stevens LA, Griffith J, Tighiouart H, Djurdjev O, Naimark D, Levin A, Levey AS. A predictive model for progression of chronic kidney disease to kidney failure. *Jama* 2011;305(15):1553-9.
14. Kostoff RN, Patel U. Literature-related discovery and innovation: Chronic kidney disease. *Technological Forecasting and Social Change* 2015;91:341-51.
15. Lakshmi KR, Nagesh Y, Krishna MV. Performance comparison of three data mining techniques for predicting kidney dialysis survivability. *International Journal of Advances in Engineering & Technology* 2014;7(1):242.
16. Kobayashi T, Yoshida T, Fujisawa T, Matsumura Y, Ozawa T, Yanai H, Iwasawa A, Kamachi T, Fujiwara K, Kohno M, Tanaka N. A metabolomics-based approach for predicting stages of chronic kidney disease. *Biochemical and biophysical research communications* 2014;445(2):412-6.
17. Hippisley-Cox J, Coupland C. Predicting the risk of Chronic Kidney Disease in Men and Women in England and Wales: prospective derivation and external validation of the QKidney® Scores. *BMC family practice* 2010;11(1):1-3.
18. Lee SJ, Jeon J. Relationship between symptom clusters and quality of life in patients at stages 2 to 4 chronic kidney disease in Korea. *Applied Nursing Research* 2015;28(4):e13-9.doi: 10.1016/j.apnr.2015.03.004
19. Di Noia T, Ostuni VC, Pesce F, Binetti G, Naso D, Schena FP, Di Sciascio E. An end stage kidney disease predictor based on an artificial neural networks ensemble. *Expert systems with applications* 2013;40(11):4438-45.
20. Verma L, Srivastava S, Negi PC. A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data. *Journal of medical systems* 2016;40(7):1-7.
21. Chaurasia DV, Pal S. Early prediction of heart diseases using data mining techniques. *Caribbean journal of Science and Technology* 2013;1:208-17.
22. Albahri AS, Hamid RA, Alwan JK, Al-Qays ZT, Zaidan AA, Zaidan BB, Albahri AO, AlAmoodi AH, Khlaf JM, Almahdi EM,

Thabet E. Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. *Journal of medical systems* 2020;44:1-1.

23. Kavitha R, Kannan E. An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. *In2016 international conference on emerging trends in engineering, technology and science (icetets)* 2016 (pp. 1-5). IEEE.
24. Paul AK, Shill PC, Rabin MR, Akhand MA. Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. In2016 5th International Conference on Informatics, *Electronics and Vision (ICIEV)* 2016 (pp. 145-150). IEEE

TUMJ Tehran University Medical Journal, July 2023; Vol. 81, No. 4: 307-318 Original Article

Realizing the early prediction chronic kidney disease based-on identifing

Mahdieh Jamshidi M.D.[1]
Vahid Jamshidi Ph.D.[2*]

1- Department of Internal Medicine, Faculty of Medicine, Rafsanjan University of Medical Sciences, Rafsanjan, Iran.
2- Department of Computer Engineering, Faculty of Engineering, Shahid Bahonar University of Kerman, Kerman, 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: 21 May 2023 Revised: 27 May 2023 Accepted: 13 Jun 2023 Available online: 22 Jun. 2023

Background: Due to the fact that various factors are involved in the development of chronic kidney disease, this disease appears with different clinical and laboratory symptoms. The variety in type and number of clinical symptoms often misguides the treating physician. The aim of this study is to extract the key features of the disease and find the best data mining methods to improve the accuracy of kidney disease diagnosis.

Methods: This cross-sectional study was conducted from September 2021 to March 2023 for 30 months at Rafsanjan Ali Ebn Abi Taleb Hospital. Predictive models were developed and tested using different combinations of disease characteristics and seven data mining methods in RapidMiner Studio software. The limitations of the study are as follows: 1) The models were based on 40-year-old and older patients records, which may limit the generalization of results to a wider age group. 2) Despite the high accuracy and comprehensiveness of the method, the models were based only on the information of kidney disease patients at Ali Ibn Abi Talib Rafsanjan Hospital. 3) The climate parameter has not been considered in the data set of the investigation to discover the hidden relationships of this parameter with the kidney disease.

Results: The results of the experiments in this study showed that the proposed prediction model using the Bayes method and eight identified key features (age, renal biopsy, uremia, sedimentation, albumin, edema, nocturnal enuresis, and urine-specific gravity), can detect kidney disease in people of different clinical characteristics, with 99.38% accuracy.

Conclusion: Considering that the early diagnosis of kidney disease and the adoption of appropriate treatment methods can prevent the progression of kidney damage, in this study, an attempt has been made to achieve this goal by using new statistical methods and artificial intelligence techniques. Based on the proposed method and the conducted experiments, the most important features and the best data mining method were obtained, and finally, kidney disease prediction was possible with high accuracy.

Keywords: chronic kidney disease, data analysis, data mining, neural network.