پیش‌بینی علت سنگ کلیه در بیماران کلیوی با استفاده از تکنیک‌های جنگل تصادفی، ماشین بردار پشتیبان و شبکه عصبی

مژگان مرتضوی[1]، عبدالامیر عطاپور[1]، مریم محمدی[2]، محمد ستاری[3*]

۱- مرکز تحقیقات بیماری‌های کلیوی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.
۲- گروه مدیریت و فناوری اطلاعات سلامت، دانشکده مدیریت و اطلاع‌رسانی پزشکی، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.
۳- مرکز تحقیقات و فناوری اطلاعات در امور سلامت، دانشگاه علوم پزشکی اصفهان، اصفهان، ایران.

*نویسنده مسئول: اصفهان، دانشگاه علوم پزشکی اصفهان، مرکز تحقیقات و فناوری اطلاعات در امور سلامت.
تلفن: ۰۳۱-۳۷۹۲۵۱۵۲
E-mail: msattarimng.mui@gmail.com

چکیده

دریافت: ۱۴۰۰/۰۶/۰۲ ویرایش: ۱۴۰۰/۰۶/۰۹ پذیرش: ۱۴۰۰/۰۸/۲۴ آنلاین: ۱۴۰۰/۰۹/۰۱

زمینه و هدف: امروزه داده‌های پزشکی با سرعت فزاینده‌ای جمع‌آوری می‌شوند. این مجموعه داده‌ها حاوی اطلاعات قیمتی هستند که دست‌یابی به آن‌ها با استفاده از روش‌های آزمایشگاهی زمان‌بر و هزینه‌بر خواهد بود. بنابراین نیاز به روش‌هایی کم هزینه برای استخراج اطلاعات وجود دارد. این مطالعه بر توسعه یک مدل پیش‌بینی برای طبقه‌بندی علت سنگ کلیه در اصفهان با استفاده از تکنیک‌های داده‌کاوی متمرکز شده است.

روش بررسی: این پژوهش از بهمن ۱۳۹۹ تا مرداد ۱۴۰۰ به‌صورت مقطعی انجام شده است. مجموعه داده‌ای مورد استفاده شامل اطلاعات ۳۵۳ بیمار سنگ کلیه در شهر اصفهان است. در این مطالعه شش صفت هدف سدیم، فسفات، اگزالات، سیترات، سیستئین و اسید اوریک تعیین شده است. تکنیک‌های مورد استفاده برای هر یک از شش صفت به‌صورت جداگانه استفاده می‌شود. تکنیک‌های مورد استفاده در این مطالعه شامل جنگل تصادفی، شبکه عصبی و ماشین بردار پشتیبان خواهد بود.

یافته‌ها: بهترین عملکرد از لحاظ میزان صحت مربوط به تکنیک‌های ماشین بردار پشتیبان در کلاس اسید اوریک، ماشین بردار پشتیبان در کلاس اگزالات و شبکه عصبی در کلاس سیستئین است. بدترین عملکرد هم مربوط به تکنیک جنگل تصادفی در کلاس سیترات است. مطمئن‌ترین قوانین با میزان اطمینان ٦٦٪ مربوط به دو کلاس سیترات و سدیم هست و کم اطمینان‌ترین قاعده با میزان اطمینان ۵۰٪ مربوط به کلاس اگزالات است.

نتیجه‌گیری: سنگ کلیه می‌تواند به‌دلایل مختلفی از جمله پایین بودن سیترات و بالا بودن اگزالات کلسیم باشد. به‌عنوان مثال برای سیترات، عواملی مانند PH خون، قند خون و فشار خون موثر است. برای جلوگیری از هر یک از دلایل سنگ کلیه، باید عوامل آن کنترل شود.

کلمات کلیدی: داده‌کاوی، پیش‌بینی، سنگ کلیه.

مقدمه

امروزه داده‌های پزشکی با سرعت فزاینده‌ای جمع‌آوری و انباشته می‌شوند.[1] استخراج اطلاعات مفید از حجم به‌سرعت در حال رشد داده‌های پزشکی از طریق نسل جدیدی از روش‌های تحلیلی و ابزارهای محاسباتی برای کمک به انسان به یک نیاز فوری تبدیل شده است. در مرکز این فرآیند استفاده از روش‌های خاص داده‌کاوی برای کشف و استخراج دانش وجود دارد.[2] هدف از داده‌کاوی استخراج روش‌مند دانش، الگوها، اطلاعات مفید یا روندهای داده‌های گذشته‌نگر، عظیم و چند بعدی است.[3] استفاده از تکنیک‌های داده‌کاوی می‌تواند منجر به شناسایی قوانین حاکم بر ایجاد، رشد و تسری بیماری‌ها شده و اطلاعات ارزشمندی را به‌منظور شناسایی

علل رخداد بیماری‌ها، تشخیص، پیش‌بینی و درمان بیماری‌ها با توجه به عوامل محیطی حاکم در اختیار متخصصان و دست‌اندرکاران حوزه سلامت قرار دهد. نتیجه این موضوع می‌تواند به‌معنای افزایش عمر و ایجاد آرامش برای افراد جامعه باشد.[٤]

نفرولیتیازیس یا سنگ کلیه، وجود سنگ‌های کلیوی است که به‌دلیل اختلال در تعادل بین حلالیت و رسوب نمک‌ها در مجاری ادرار و کلیه‌ها ایجاد می‌شود.[٥] هنگامی‌که ادرار با ترکیبات نامحلول حاوی کلسیم اگزالات و کلسیم فسفات اشباع می‌شود، سنگ کلیه به‌وجود می‌آید که در نتیجه کمبود آب بدن یا استعداد ژنتیکی برای دفع بیش از حد این یون‌ها در ادرار ایجاد می‌شود.[٦]

عوامل خطر سنگ کلیه شامل سن، نژاد، تحصیلات، توده بدن، فشار خون بالا و ادرارآورها در کنار مصرف شیر، قهوه، چای، نوشابه، الکل و مکمل ویتامین C است.[٧]

شیوع مادام‌العمر بیماری سنگ کلیه ١٤٪ است و اطلاعات اپیدمیولوژیک تایید کرده است که این موضوع در حال افزایش است.[٨] علاوه بر این، حداقل ۵۰٪ از بیماران در طی ۱۰ سال دچار عود سنگ کلیه خواهند شد.[٩]

بیماری سنگ کلیه علاوه بر اثرات مخربی که بر کیفیت زندگی بیمار دارد، عواقب اقتصادی گسترده‌ای را نیز در پی دارد. هزینه سالانه این بیماری در ایالات متحده تا سال ۲۰۳۰ بیش از چهار میلیارد دلار تخمین زده می‌شود.[١٠و١١]

یک مطالعه کوهورت آینده‌نگر نشان داد افراد مبتلا به بیماری سنگ کلیه ٦٧٪–۵۰ بیشتر در معرض خطر ابتلا به بیماری مزمن کلیوی هستند. همچنین گروه مبتلا به سنگ کلیه دو برابر خطر ابتلا به بیماری کلیوی در مرحله نهایی را دارند.[١٢]

از این‌رو با استفاده از تکنیک‌های داده‌کاوی می‌توان مدلی جهت پیش‌بینی زود هنگام و کاهش اثرات این بیماری، ارایه نمود.[١٣] علاوه بر این، این روش می‌تواند تعداد آزمایش‌های رادیوگرافی غیر ضروری، مانند CT scan را در محیط مراقبت حاد کاهش دهد.[١٤]

Kaladhar برای پیش‌بینی سنگ کلیه از روش‌های یادگیری ماشین استفاده کرد. در این مطالعه درخت تصمیم C4.5 و جنگل تصادفی صحت ۹۳٪ و ماشین بردار پشتیبانی صحت ۹۱/۹۸٪ را به‌دست آوردند.

الگوریتم درخت تصمیم C4.5 ارتقا یافته (Iterative ID3 Dichotomiser 3) است. این الگوریتم معمولا برای ویژگی‌های اسمی و پیوسته مناسب است. تکنیک درخت تصمیم C4.5 با تقسیم بازگشتی داده‌ها، یک درخت تصمیم برای داده‌ها ایجاد می‌کند.

درخت تصمیم براساس استراتژی اول عمق عمل می‌نماید. همچنین الگوریتم‌های رگرسیون لجستیک و شبکه عصبی نتایج صحت خوبی را نشان دادند. به‌طور کلی رویکردهای یادگیری ماشین نتایج بهتری را در درمان سنگ کلیه ارایه دادند.[١٥]

Oladeji و همکارانش یک مدل پیش‌بینی‌کننده برای خطر ابتلا به بیماری‌های کلیوی با استفاده از سه الگوریتم یادگیری ماشین تحت نظارت درخت تصمیم، پرسپترون چند لایه (Multilayer perceptron) و الگوریتم ژنتیک ارایه دادند. در این مقاله نتیجه‌گیری شد که پرسپترون چند لایه با استفاده از ۳۳ متغیر شناسایی شده توسط متخصصان غدد با صحت ۱۰۰٪ بهترین عملکرد را دارد.[٧]

این مطالعه بر توسعه یک مدل پیش‌بینی برای طبقه‌بندی علت سنگ کلیه در اصفهان با استفاده از تکنیک‌های داده‌کاوی براساس اطلاعات تاریخی استخراج شده در مورد علت سنگ کلیه متمرکز شده است.

روش بررسی

شامل چهار قسمت جمع‌آوری اطلاعات، تبدیلات داده‌ای، مشخص شدن صفات هدف، معیارهای ارزیابی و معرفی تکنیک‌های مورد استفاده است:

جمع‌آوری اطلاعات: مجموعه داده شامل اطلاعات بیمارانی است که در شهر اصفهان زندگی می‌کنند و در معرض سنگ کلیه قرار داشته‌اند. این مجموعه شامل ۳۵۳ بیمار سنگ کلیه است.

مجموعه داده‌ای مورد استفاده شامل ۲۵ صفت جنسیت، شغل، ابتلا به بیماری دیابت، فشار خون، وزن، وزن ایده‌آل، تعداد گلبول سفید خون، هموگلوبین خون، حجم متوسط گلبول قرمز خون، تعداد پلاکت خون، قند خون، اوره، نیتروژن اوره خون، کراتینین خون، نرخ فیلتراسیون گلومرولی، اسید اوریک، فسفر، سدیم، چگالی نسبی یا وزن مخصوص ادرار (SG)، تعداد گلبول قرمز خون، فسفلات، اگزالات، سیترات، سیستئین، اسید یا بازی بودن خون (PH) است.

پس از مشورت با خبرگان، از این ۲۵ صفت، ۱۹ صفت جنسیت، شغل، ابتلا به بیماری دیابت، فشار خون، وزن، وزن ایده‌آل،

تعداد گلبول سفید خون، هموگلوبین خون، حجم متوسط گلبول قرمز خون، تعداد پلاکت خون، قند خون، اوره، نیتروژن اوره خون، کراتینین خون، نرخ فیلتراسیون گلومرولی، فسفر، چگالی نسبی یا وزن مخصوص ادرار SG، تعداد گلبول قرمز خون و اسید یا بازی بودن خون (PH) انتخاب شدند.

آماده‌سازی داده: بسیاری از صفات مورد استفاده از قبیل قند خون، اوره، نیتروژن اوره خون و غیره عددی هستند. اکثر این داده‌ها از قبیل اوره، نیتروژن اوره خون، کراتینین خون، فسفر و تعداد گلبول سفید خون تبدیل به بازه‌های سه‌تایی پایین، نرمال و بالا شدند.

قند خون تبدیل به چهارتایی پایین، نرمال، پیش‌دیابت و دیابت شد. فشار خون و دیابت به بازه دوتایی برحسب این‌که فرد دیابت یا فشار خون دارد یا ندارد، تقسیم شد.

بازه‌های ٦٠–٤٠، ٨٠–٦١ و ٩٩–٨١ و بیشتر از ١٠٠ برای وزن و بازه‌های ٦٠–٤٠، ٨٠–٦١ برای وزن ایده‌آل در نظر گرفته شد. نرخ فیلتراسیون گلومرولی به سه بازه نارسایی کلیوی، افت عملکرد کلیه و عملکرد طبیعی کلیه‌ها) تبدیل شد. تعداد گلبول قرمز خون تبدیل به سه بازه کم، نرمال و زیاد شدند.

مشخص شدن صفات هدف: در این مطالعه شش صفت هدف تعیین شده است که شامل صفات سدیم، فسفات، اگزالات، سیترات، سیستئین و اسید اوریک است .هر شش صفت هدف از علل ابتلای افراد به سنگ کلیه هستند. همچنین هر شش صفت عددی هستند. هر شش صفت به بازه‌های سه‌تایی پایین، نرمال و بالا تقسیم شدند.

معیارهای ارزیابی: این مطالعه از مجموعه داده‌های آزمون برای ارزیابی روش‌ها استفاده کرده است. ابتدا یک ماتریس درهم‌ریختگی محاسبه می‌شود که شامل معیارهای مختلفی از قبیل مثبت واقعی، منفی واقعی، مثبت کاذب و منفی کاذب است. در زمینه یادگیری ماشین و به‌طور خاص موضوع طبقه‌بندی آماری، ماتریس درهم‌ریختگی به‌عنوان ماتریس خطا نیز شناخته می‌شود.[١٦]

ماتریس درهم‌ریختگی، یک جدول‌بندی خاص است که اجازه می‌دهد تا عملکرد یک الگوریتم، به‌طور معمول یادگیری نظارت شده، مجسم شود. مثبت واقعی، تعداد رکوردهایی را مشخص می‌کند که عضو کلاس مثبت هستند و دسته‌بند آن‌ها را به‌طور صحیح عضو کلاس مثبت قرار داده است. منفی واقعی، رکوردهایی را مشخص می‌کند که عضو کلاس منفی هستند و دسته‌بند آن‌ها را به‌طور صحیح عضو کلاس منفی قرار داده است.

مثبت کاذب، رکوردهایی را مشخص می‌کند که عضو کلاس منفی هستند اما دسته‌بند آن‌ها را به اشتباه عضو کلاس مثبت قرار داده است. منفی کاذب، رکوردهایی را مشخص می‌کند که عضو کلاس مثبت هستند اما دسته‌بند آن‌ها را به اشتباه عضو کلاس منفی قرار داده است.[١٧]

گروه هدف شامل هفت کلاس با سه مقدار پایین، نرمال و بالا است. در واقع برای هر یک از کلاس‌ها به‌صورت جداگانه تکنیک‌ها استفاده خواهند شد. معیارهای مختلفی برای ارزیابی استفاده می‌شود. یکی از این معیارها Accuracy است که هرچه دقت این معیار نزدیک‌تر باشد، نتیجه بهتری خواهد داشت.[١٨] این معیار براساس فرمول زیر محاسبه می‌شود :

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}.$$

معیار دیگر، معیار Confidence است.[١٩] این معیار مشخص می‌نماید که برحسب فرضیات، چه میزان می‌توان به حکم اطمینان داشت. این معیار هرچه به یک نزدیک‌تر باشد، عملکرد بهتری دارد.

معرفی تکنیک‌های داده‌کاوی: تکنیک‌های مورد استفاده برای هر یک از شش صفت به‌صورت جداگانه استفاده می‌شود. نرم‌افزار مورد استفاده، رپیدماینر خواهد بود. (Software, Germany RapidMiner). تکنیک‌های مورد استفاده در این مطالعه شامل جنگل تصادفی، شبکه عصبی و ماشین بردار پشتیبان خواهد بود.

طبق پژوهش‌های انجام شده در خصوص تکنیک‌های داده‌کاوی بر روی مجموعه داده‌های بیماری‌های مختلف، تکنیک شبکه عصبی از متداول‌ترین روش‌های توسعه مدل‌های پیش‌بینی نتایج در زمینه پزشکی است.[٢٠و٢٣]

شبکه‌های عصبی، الگوریتم‌هایی هستند که می‌توانند برای انجام مدل‌سازی آماری غیر خطی مورد استفاده قرار گیرند و جایگزینی جدید برای رگرسیون لجستیک (Logistic regression) ارایه دهند. شبکه‌های عصبی مزایای زیادی از جمله نیاز به آموزش آماری کمتر، توانایی تشخیص ضمنی روابط غیر خطی پیچیده بین متغیرهای وابسته و مستقل، توانایی تشخیص همه تعاملات احتمالی بین متغیرهای پیش‌بینی‌کننده و در دسترس بودن الگوریتم‌های آموزشی متعدد را ارایه می‌دهند.[٢٤]

تکنیک جنگل تصادفی نیز از جمله تکنیک‌های پر کاربرد در زمینه داده‌کاوی است.[۲۱] این تکنیک میزان پیش‌برازش را که یکی از مشکلات درختان تصمیم است کاهش داده و به بهبود میزان صحت کمک می‌کند. همچنین این تکنیک با هر دو نوع مقادیر اسمی یا دسته‌ای و مقادیر پیوسته به‌خوبی کار می‌کند. از طرف دیگر به‌دلیل استفاده از رویکرد مبتنی بر قاعده در این تکنیک نیاز به نرمال‌سازی داده‌ها وجود ندارد. تکنیک جنگل تصادفی اغلب دقیق‌تر از یک طبقه‌بندی‌کننده واحد است. این تکنیک دارای توانایی مدیریت داده‌ها بدون پیش‌پردازش و عدم نیاز به کوچک‌سازی و تبدیل داده‌ها است.[۲۵]

تکنیک ماشین بردار پشتیبان نیز یکی از موفق‌ترین الگوریتم‌های طبقه‌بندی در حوزه داده‌کاوی است.[۲۶] ماشین بردار پشتیبان از یک فضا با ابعاد بالا برای انجام طبقه‌بندی باینری استفاده می‌کند.[۲۷]

ماشین بردار پشتیبان یک روش طبقه‌بندی غیر خطی و غیر پارامتری نویدبخش جدید است که در حال حاضر نتایج خوبی در زمینه تشخیص پزشکی، و سایر زمینه‌ها نشان داده است. ماشین بردار پشتیبان حتی با وجود مقداری سوگیری در مجموعه داده آموزشی، می‌تواند نتایج طبقه‌بندی با صحت بالا و قوی را براساس مبانی نظری تولید کند، حتی اگر داده‌های ورودی غیر یکنواخت و غیر خطی باشند. بنابراین این تکنیک می‌تواند به ارزیابی راحت‌تر اطلاعات مرتبط کمک کند.[۲۸] در واقع این روش به‌دنبال رابطه خطی با حاشیه اطمینان بالا بین متغیرهای مستقل و وابسته است.

ارزیابی در مرحله ارزیابی ۳۵۳ نمونه و ۲۵ صفات در نظر گرفته شده است. تقسیم‌بندی مجموعه داده‌ای به مجموعه آموزشی و تست، از روش 10-fold cross validation استفاده می‌شود. بر این اساس ۳۵۳ نمونه موجود به ۱۰ گروه تقسیم می‌شود. طی ۱۰ بار تکرار آزمایش هر بار نه گروه که ۹۰٪ مجموعه داده‌ای اصلی را تشکیل می‌دهد به‌عنوان مجموعه داده آموزشی و یک گروه باقی‌مانده، که ٪۱۰ مجموعه داده‌ای اصلی را تشکیل می‌دهد به‌عنوان مجموعه داده‌ای تست در نظر گرفته می‌شود. از مجموعه داده‌ای آموزشی برای تولید الگوهای ورودی جهت ساخت مدل طبقه‌بند و از مجموعه داده‌ای تست برای ارزیابی صحت آن استفاده می‌شود.

یافته‌ها

۱۹ صفت به‌عنوان پیش‌بینی‌کننده در نظر گرفته شده است. با توجه به جدول ۱، تعداد مردان دو برابر زنان است. همچنین کارمندان و بازنشستگان تنها ۱۵٪ افراد و خانه‌دارها۳۰٪ افراد را تشکیل می‌دهند و بقیه شغل‌ها جزو مشاغل آزاد بوده و ۵۵٪ شغل‌ها را تشکیل می‌دهند. همچنین ۱۱۹ نفر از افراد فشار خون دارند و غالب افراد وزنشان بین ۹۹-۸۱ است.

متغیر هدف شامل شش کلاس دو مقداری سدیم، سیترات، اسید اوریک، اگزالات، فسفات و سیستئین تعریف شده است. ۱۳۵ نفر سنگ کلیه از نوع اگزالات دارند و ۱۲۳ نفر از آن‌ها فسفات، در مقابل تنها ۲۲ نفر سنگ کلیه از نوع سدیم دارند. (جدول ۲).

طبق جدول ۳، بهترین عملکرد مربوط به تکنیک ماشین بردار پشتیبان در کلاس اسید اوریک، ماشین بردار پشتیبان در کلاس اگزالات و شبکه عصبی در کلاس سیستئین است. بدترین عملکرد هم مربوط به تکنیک جنگل تصادفی در کلاس سیترات است. جنگل تصادفی در کلاس سیستئین با ۸۹/٦٪ صحت نسبت به سایر کلاس‌ها عملکرد بهتری داشته است.

با استفاده از تکنیک جنگل تصادفی، در مجموع برای هر یک از کلاس‌ها، یک قانون استخراج شده که نسبت به بقیه میزان اطمینان بیشتری داشته است. بنابراین در نهایت شش قانون استخراج شده است که این قوانین برحسب میزان اطمینان به‌صورت نزولی مرتب شده‌اند. مطمئن‌ترین قوانین با میزان اطمینان ٦٦٪ مربوط به دو کلاس سیترات و سدیم هست و کم اطمینان‌ترین قاعده با اطمینان ۵۰٪ مربوط به کلاس اگزالات است. (جدول ٤).

تکنیک ماشین بردار پشتیبان برای هر یک از شش کلاس به‌صورت جداگانه اعمال شود سپس وزن مستخرج برای صفات موجود در قوانین ٦-۱ استخراج شده است. در نهایت میانگین این وزن‌ها محاسبه شده است و نتایج آن در جدول ۵ آمده است. میانگین صفاتی که کمتر از ۰/۰۵ است در جدول ۵ نیامده است. با استفاده از روش ماشین بردار پشتیبان، عواملی مانند فشار خون، اوره و شغل، سه عامل تاثیرگذار بودند.

جدول ۱: خصوصیات، مقادیر و تعداد مربوط به مجموعه داده‌ای بیماران سنگ کلیه

خصوصیات	مقادیر (تعداد)
جنسیت	مرد (۲۱۰) و زن (۱۱۲)
شغل	آزاد (شیشه‌بر (٦)، فروشنده (۲۳)، راننده (١٤)، کارگر (۲۵)، بیکار (۲۰)، منشی (۸)، آرایشگر (۷)، پیمانکار (۹)، کشاورز (۵)، زرگر (٦)، پلیس (۳)، خانه‌دار (۷۸)، کارمند(٤۵) و بازنشسته (١٤)
دیابت	بله (٤۲) و خیر (۲۱۵)
فشار خون	دارد (بیشتر از ۱۲ و کمتر از ۱۱) (۱۱۹) و ندارد (۱۱–۱۲) (١٤۵)
وزن	٦۰–٤۰ (۲۱)، ۸۰–٦۱ (۱۲٦)، ۹۹–۸۱ (١٤۰) و بیشتر از ۱۰۰ (۲۸)
وزن ایده‌آل	٦۰ –٤۰ (٦۵) و ۸۰–٦۱ (١٤۰)
قند خون	پایین (کمتر از ۷۰) (٤۸)، نرمال (۹۰–۷۰) (۱۲۵)، پیش‌دیابت (۱۰۰–۱۲۵) (۱۰۵)، دیابت (بیشتر از ۱۲۵) (۵۲)
تعداد گلبول قرمز خون	کم (کمتر از ٤/۷ میلیون سلول در µl) (۱۸۰)، نرمال(٦/۱–٤/۷ میلیون سلول در µl) (۹۳)، زیاد (بیشتر از ٦/۱ میلیون سلول در µl) (٦۲)
هموگلوبین خون	پایین (کمتر از ۱۲) (۱۱۲)، نرمال(۱٦–۱۲) (۱۷۵)، بالا(بیشتر از ۱٦) (٤۲)
فسفر	پایین (کمتر از ۳) (٦۰)، نرمال (٤/۵–۳) (۲۰۱)، بالا (بیشتر از ٤/۵) (۱۷)
اوره	پایین (کمتر از ۱۰) (۱۰۵)، نرمال (۱۰–٤۰) (١٤۳)، بالا (بیشتر از ٤۰) (۷۵)
نیتروژن اوره خون	پایین (کمتر از ۷) (۷۱)، نرمال (۷–۲۰) (١٦۱)، بالا (بیشتر از ۲۰)(۷۷)
کراتینین خون	پایین (کمتر از ۰/۳) (٦۳)، نرمال(١/٤–۰/۳) (۱۵۰)، بالا (۷۹) (بیشتر از ١/٤)
نرخ فیلتراسیون گلومرولی	نارسایی کلیوی (کمتر از ۱۵) (١٤۰)، افت عملکرد کلیه‌ها (۱۵–۹۰) (۷۰)، عملکرد طبیعی کلیه‌ها (بیشتر از ۹۰) (۹۵)
تعداد پلاکت خون	پایین (کمتر از ۱۵۰) (۸۹)، نرمال (٤۵۰–۱۵۰) (١٦۱)، بالا (بیشتر از ٤۵۰) (۹٤)
حجم متوسط گلبول قرمز	پایین (کمتر از ۸۰) (۸٤)، نرمال (۸۰–۹٦) (۲۰٤)، بالا(بیشتر از ۹٦) (۲۱)
تعداد گلبول‌های سفید	پایین (کمتر از ۵۰۰۰) (۱۰۷)، نرمال (۵۰۰۰–۱۰۰۰۰) (۲۱۱)، بالا (بیشتر از ۱۰۰۰۰)(۱۷)
PH	اسیدی (کمتر از ۷/۳۵) (۳۵)، نرمال (۷/٤۵–۷/۳۵) (۳۹)، قلیایی (بیشتر از ۷/٤۵) (۵۰)
چگالی نسبی SG	پایین (کمتر از ۱/۰۰۲) (۷۵)، نرمال (۱/۰۰۳ تا ۱/۰۰۲) (۱٦۰)، بالا (بیشتر از ۱/۰۰۳) (۳۰)

جدول ۲: کلاس‌های مختلف سنگ کلیه و تعداد اعضای آن‌ها

کلاس	سدیم	سیترات	اسید اوریک	اگزالات	فسفات	سیستئین
سنگ کلیه از نوع مشخص شده است	۲۲	۱۱۲	٦۳	۱۳۵	۱۲۳	۷٤
سنگ کلیه از نوع مشخص شده نیست	۲۰۳	۱۸۵	۲۲۷	۱۵۳	١٦٤	۱۵۹

جدول ۳: میزان Accuracy تکنیک‌های مختلف داده‌کاوی در کلاس‌های مختلف سنگ کلیه

روش‌ها	سدیم	سیترات	اسید اوریک	اگزالات	فسفات	سیستئین
جنگل تصادفی	٪۸۳/۳	٪۷۹/۲	٪۸۵/٤	٪۸۸	٪۸۳/۳	٪۸۹/۵
ماشین بردار پشتیبان	٪۸۳/۳	٪۸۷/۷	٪۹۵/۸	٪۹۵/۸	٪۸۱/۳	٪۹۳/٦
شبکه عصبی	٪۸۷/۷	٪۸۵/٤	٪۹۳/٦	٪۹۳/٦	٪۸۰/٤	٪۹۵/۸

جدول٤: قوانین مستخرج از اعمال تکنیک‌های مختلف داده‌کاوی بر روی مجموعه داده‌ای

Rule	Condition	Result	Confidence
Rule 1	اگر PH خون اسیدی باشد و فشار خون فرد بالا باشد و قند خون از نوع پیش‌دیابت باشد	علت سنگ کلیه پایین بودن سیترات خواهد بود	۰/٦٦
Rule 2	اگر چگالی نسبی SG پایین باشد و اوره پایین باشد	علت سنگ کلیه بالا بودن سدیم خواهد بود	۰/٦٦
Rule 3	اگر PH خون اسیدی باشد و شغل فرد آزاد باشد و وزن فرد بین ۱۰۰–۸۰ باشد	علت سنگ کلیه بالا بودن فسفات خواهد بود	۰/٦٣
Rule 4	اگر نرخ فیلتراسیون گلومرولی متناظر با افت عملکرد کلیه‌ها باشد و تعداد گلبول قرمز خون پایین باشد و شغل آزاد باشد	علت سنگ کلیه بالابودن سیستئین خواهد بود	۰/٦٣
Rule 5	اگر تعداد گلبول سفید از نوع نرمال باشد و نرخ فیلتراسیون گلومرولی بالا باشد و جنسیت مرد باشد	علت سنگ کلیه بالا بودن اسید اوریک خواهد بود	۰/٥٨
Rule 6	اگر نیتروژن اوره خون بالا باشد و وزن فرد بین ۱۰۰–۸۰ باشد	علت سنگ کلیه بالا بودن اگزالات خواهد بود	۰/٥

جدول ٥: میانگین وزن مستخرج با استفاده ازتکنیک ماشین بردار پشتیبان برای هر یک از صفات

صفت	وزن
نیتروژن اوره خون	۰/۰۷۷
تعداد گلبول سفید	۰/۱۱۰
جنسیت	۰/۰۸۱
وزن	۰/۰۸٦
نرخ فیلتراسیون گلومرولی	۰/۰٥۸
تعداد گلبول قرمز خون	۰/۰٥۸
شغل	۰/۱۸٤
PH	۰/۰٥۷
چگالی نسبی SG	۰/۰٤۲
اوره	۰/۱٤۲
فشار خون	۰/۲۰٤
قند خون	۰/۰٥۷

بحث

سیترات: قاعده اول بیان می‌نماید که اگر PH خون اسیدی باشد و فشار خون فرد بالا باشد و قند خون از نوع پیش‌دیابت باشد آن‌گاه علت سنگ کلیه، پایین بودن سیترات خواهد بود. غذاهایی که بار اسیدی بالایی دارند شامل گوشت، ماهی، مرغ، پنیر و تخم مرغ است.[۲۹]

رژیم غذایی با پروتئین حیوانی بالا نه‌تنها کلسیم ادرار و غلظت اسید اوریک را افزایش می‌دهد، بلکه باعث کاهش سطح سیترات و PH ادرار می‌شود. در قاعده اول هم رابطه بین اسیدی بودن PH ادرار و پایین بودن سطح سیترات بیان شده است. با این‌حال، رژیم کم پروتئین توصیه نمی‌شود. در عوض، یک رژیم پروتئینی مناسب، متشکل از gr ۸–۰/۱ پروتئین در kg است. پروتئین گیاهی نسبت به پروتئین حیوانی، ادرار را کمتر اسیدی می‌کند و منبع پروتئین ترجیحی است.[۳۰]

یکی دیگر از علل ایجاد سنگ کلیه پایین بودن اسید اوریک بود. افزایش سن، شرایط گرم و خشک محیطی، جنسیت مرد، کاهش حجم ادرار و کاهش PH ادرار از عوامل مهم در بروز نفرولیتیاز اسید اوریک است. تقریبا ۲/۳ سنگ‌های کلیه اسید اوریک را می‌توان با افزایش PH و حجم ادرار همراه با کاهش هیپراوریکوزوری

(Hyperuricosuria) حل کرد.[۳۱] طبق قاعده پنجم هم، جنسیت مرد در بروز سنگ کلیه از نوع اسید اوریک موثر است.

سنگ‌های کلسیم: اگزالات کلسیم و فسفات کلسیم. بالا بودن اگزالات یکی دیگر از عوامل ایجاد سنگ کلیه بود. عود سنگ کلسیم بیشتر از سایر انواع سنگ کلیه است. برای پیشگیری از اگزالات کلسیم، باید با خوردن رژیم غذایی حاوی میوه‌ها و سبزیجات، مصرف سیترات مکمل یا تجویز شده یا نوشیدن آب‌های معدنی قلیایی، ادرار قلیایی شود. برای جلوگیری از سنگ‌های فسفات کلسیم، ادرار باید اسیدی شود.[۳۲]

طبق قاعده سوم هم اسیدی بودن PH و اضافه وزن در بروز سنگ فسفات کلسیم موثر است. همچنین طبق قاعده ششم، افزایش وزن در بروز سنگ کلسیم اگزالات موثر است.

یکی دیگر از علل ایجاد سنگ کلیه در این مطالعه، بالا بودن سدیم بود. در یک مطالعه نشان داده شد که دریافت سدیم در رژیم غذایی با افزایش ۶۱٪–۱۱ خطر ابتلا به سنگ کلیه، پس از تنظیم سایر عوامل خطر نفرولیتیازیس همراه بود. این اثر در زنانی که بیشترین میزان سدیم در روز را بیش از ۳۲۴۹ mg دریافت کرده‌اند، بارزتر است.[۳۳] طبق قاعده شماره دو، کاهش وزن مخصوص ادرار می‌تواند در بالا بردن سدیم موثر باشد.

سیستئین: بالا بودن سیستئین یکی دیگر از عوامل ایجاد سنگ کلیه بود. در یک مطالعه گذشته‌نگر که از ۴۴۲ بیمار مبتلا به سیستینوریا جمع‌آوری شده بود، ۲۷٪ بیماران نرخ فیلتراسیون گلومرولی کمتر از ۶۰ میلی لیتر در دقیقه تخمین زده شده بودند که نشان‌دهنده درجه بیماری مزمن کلیه است.

فشار خون بالا نیز در این مطالعه شایع بود، ۲۸/۶٪ از بیماران دارای فشار خون بالا بودند.[۳۴] طبق قاعده چهارم، اگر نرخ فیلتراسیون گلومرولی متناظر با افت عملکرد کلیه‌ها باشد می‌تواند در ایجاد سنگ کلیه از نوع سیستئین موثر باشد.

Kazemi در پژوهش خود نوع سنگ کلیه با استفاده از تکنیک‌های داده‌کاوی شناسایی نمود. او عواملی مانند فشار خون، سطح کلسیم، جنسیت، سطح اسید اوریک، دیابت، حالت تهوع را به‌عنوان عوامل موثر بر سنگ کلیه برشمرد.[۳۵] در مطالعه ما هم عواملی مانند فشار خون و دیابت در قواعد مستخرج ذکر شده است.

محدودیت خاصی در پژوهش وجود ندارد. پیشنهاد می‌شود در مطالعات دیگر از مجموعه‌های داده‌ای شهرهای دیگر ایران استفاده شده و نتایج آن با نتایج اصفهان مقایسه شود تا بررسی شود آیا موقعیت جغرافیایی می‌تواند بر نتایج مستخرج تاثیر بگذارد یا خیر.

نتیجه‌گیری: سنگ کلیه می‌تواند به‌دلایل مختلفی از جمله پایین بودن سیترات، بالا بودن اسید اوریک، بالا بودن اگزالات کلسیم باشد و برای هر یک از این دلایل، عواملی مختلفی می‌تواند موثر باشد. در این پژوهش با استفاده از تکنیک‌های داده‌کاوی، این عوامل در قالب قواعد استخراج شده‌اند.

با وجود این‌که بیشتر داده‌ها غیر عددی بودند ولی قابل تبدیل به عدد بودند، تکنیک‌های مانند ماشین بردار پشتیبان و شبکه عصبی عملکرد مناسبی داشتند. مطمئن‌ترین قانون بیان می‌کند که اگر PH خون اسیدی باشد و فشار خون فرد بالا باشد و قند خون از نوع پیش‌دیابت باشد علت سنگ کلیه، پایین بودن سیترات است. با استفاده از روش ماشین بردار پشتیبان، عواملی مانند فشار خون، اوره و شغل سه عامل تاثیرگذار بودند. همچنین اضافه وزن و دیابت در سه قانون از شش قانون مستخرج به‌عنوان عوامل حضور دارند. بنابراین افراد باید بیشتر مراقب وزن خود باشد چون علاوه بر دیابت آن‌ها را در معرض بیماری سنگ کلیه هم قرار خواهد داد.

سپاسگزاری: این مقاله حاصل از طرح تحقیقاتی تحت عنوان "پیش‌بینی نوع سنگ کلیه با استفاد از تکنیک‌های داده‌کاوی "مصوب دانشگاه علوم پزشکی اصفهان در سال ۱۳۹۹ با کد اخلاق IR.MUI.RESEARCH.REC.1399.520 می‌باشد که با حمایت دانشگاه علوم پزشکی اصفهان اجرا شده است.

References

1. Jothi N, Husain W. Data mining in healthcare–a review. *Procedia Comput Sci* 2015;72:306-13.
2. Agrawal R, Psaila G. Active Data Mining. In: KDD. 1995.
3. Iavindrasana J, Cohen G, Depeursinge A, Müller H, Meyer R, Geissbuhler A. Clinical data mining: a review. *Yearb Med Inform* 2009;18(01):121–33.
4. Jurian N. Predicting the effectiveness of preeclampsia medications based on dose and method of drug consumption using data mining. *Iran J Obstet Gynecol Infertil* 2014;17(123):13-22.html
5. Han H, Segal AM, Seifter JL, Dwyer JT. Nutritional Management of Kidney Stones (Nephrolithiasis). *Clin Nutr Res* 2015;4(3):137-52.

6. T Taylor EN, Curhan GC. Dietary calcium from dairy and nondairy sources, and risk of symptomatic kidney stones. *J Urol* 2013;190(4):1255–9.
7. Oladeji F, Idowu P, Egejuru N, Faluyi S, Balogun J. Model for predicting the risk of kidney stone using data mining techniques. 2019.
8. Rukin NJ, Siddiqui ZA, Chedgy ECP, Somani BK. Trends in Upper Tract Stone Disease in England: Evidence from the Hospital Episodes Statistics Database. *Urol Int* 2017;98(4):391-6.
9. Knoll T. Epidemiology, pathogenesis, and pathophysiology of urolithiasis. *Eur Urol Suppl* 2010;9(12):802-6.001545
10. Jones P, Karim Sulaiman S, Gamage KN, Tokas T, Jamnadass E, Somani BK. Do Lifestyle Factors Including Smoking, Alcohol, and Exercise Impact Your Risk of Developing Kidney Stone Disease? Outcomes of a Systematic Review. *J Endourol* 2021;35(1):1–7.
11. Antonelli JA, Maalouf NM, Pearle MS, Lotan Y. Use of the National Health and Nutrition Examination Survey to calculate the impact of obesity and diabetes on cost and prevalence of urolithiasis in 2030. *Eur Urol* 2014;66(4):724-9.
12. Rule AD, Bergstralh EJ, Melton LJ 3rd, Li X, Weaver AL, Lieske JC. Kidney stones and the risk for chronic kidney disease. *Clin J Am Soc Nephrol* 2009;4(4):804-11.
13. Kazemi Y, Mirroshandel SA. A novel method for predicting kidney stone type using ensemble learning. *Artif Intell Med* 2018;84:117-26.
14. C Chen Z, Bird VY, Ruchi R, Segal MS, Bian J, Khan SR, et al. Development of a personalized diagnostic model for kidney stone disease tailored to acute care by integrating large clinical, demographics and laboratory data: the diagnostic acute care algorithm - kidney stones (DACA-KS). *BMC Med Inform Decis Mak* 2018;18(1):1-14.
15. Kaladhar D, Apparao R, Varahalarao V. Statistical and data mining aspects on kidney stones. Statistical and data mining aspects on kidney stones: A systematic review and meta-analysis. *Open Access Sci Rep* 2012;1:543.
16. Stehman SV. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens Environ* 1997;62(1):77-89.
17. Powers DM. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *ArXiv Preprint ArXiv:201016061* 2020.
18. Swets JA. Measuring the accuracy of diagnostic systems. *Science* 1988;240(4857):1285-93.
19. Mac Aodha O, Humayun A, Pollefeys M, Brostow GJ. Learning a confidence measure for optical flow. *IEEE Trans Pattern Anal Mach Intell* 2013;35(5):1107-20.
20. Bagherian H, Haghjooy Javanmard S, Sharifi M, Sattari M. Using data mining techniques for predicting the survival rate of breast cancer patients: a review article. *Tehran Univ Med J* 2021;79(3):176-86.
21. Han X, Zheng X, Wang Y, Sun X, Xiao Y, Tang Y, et al. Random forest can accurately predict the development of end-stage renal disease in immunoglobulin a nephropathy patients. *Ann Transl Med* 2019;7(11).
22. Almansour NA, Syed HF, Khayat NR, Altheeb RK, Juri RE, Alhiyafi J, et al. Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. *Comput Biol Med* 2019;109:101-11.
23. Moeinzadeh F, Rouhani MH, Mortazavi M, Sattari M. Prediction of chronic kidney disease in Isfahan with extracting association rules using data mining techniques. *Tehran Univ Med J* 2021;79(6):459-67.
24. Tu JV. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. *J Clin Epidemiol* 1996;49(11):1225-31.
25. Shaik AB, Srinivasan S, editors. A brief survey on random forest ensembles in classification model. International Conference on Innovative Computing and Communications; 2019: Springer.
26. Soumaya Z, Taoufiq BD, Benayad N, Yunus K, Abdelkrim A. The detection of Parkinson disease using the genetic algorithm and SVM classifier. *Appl Acoust* 2020;171:107528.
27. Bhavsar YB, Waghmare KC. Intrusion detection system using data mining technique: Support vector machine. *Int J Emerg Technol Adv Eng* 2013;3(3):581-6.
28. Auria L, Moro RA. Support vector machines (SVM) as a technique for solvency analysis. 2008.
29. Phillips R, Hanchanale VS, Myatt A, Somani B, Nabi G, Biyani CS. Citrate salts for preventing and treating calcium containing kidney stones in adults. *Cochrane Database Syst Rev* 2015(10).
30. Gul Z, Monga M. Medical and dietary therapy for kidney stone prevention. *Korean J Urol* 2014;55(12):775-9.
31. Zegarra M, Ruchi R, Hammer K, Lo TS. Uric acid nephrolithiasis. In: Nephrolithiasis: Risk Factors,Treatment and Prevention.Nova Science Publishers. 2015. p.35-54.
32. Gottlieb M, Long B, Koyfman A. The evaluation and management of urolithiasis in the ED: A review of the literature. The American journal of emergency medicine. 2018 Apr 1;36(4):699-706.
33. Park S, Pearle MS. Pathophysiology and management of calcium stones. *Urol Clin North Am* 2007;34(3):323.34.
34. Prot-Bertoye C, Lebbah S, Daudon M, Tostivint I, Bataille P, Bridoux F, et al. CKD and its risk factors among patients with cystinuria. *Clin J Am Soc Nephrol* 2015;10(5):842-51.
35. Kazemi Y, Mirroshandel SA. A novel method for predicting kidney stone type using ensemble learning. *Artif Intell Med* 2018;84:117-26.

.

TUMJ Tehran University Medical Journal, December 2021; Vol. 79, No. 9: 706-714 Original Article

Predicting the cause of kidney stones in patients using random forest, support vector machine and neural network

Mojgan Mortazavi Ph.D.[1]
Abdolamir Atapour Ph.D.[1]
Maryam Mohammadi M.D.[2]
Mohammad Sattari Ph.D.[3*]

1- Isfahan Kidney Diseases Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.
2- Department of Health Information Technology and Management, School of Medical Management and Information Sciences, Isfahan University of Medical Sciences, Isfahan, Iran.
3- Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.

[*]Corresponding author: Health Information Technology Research Center, Isfahan University of Medical Sciences, Isfahan, Iran.
Tel: +98–31-37925152
E-mail: msattarimng.mui@gmail.com

Abstract

Received: 24 Aug. 2021 Revised: 31 Aug. 2021 Accepted: 15 Nov. 2021 Available online: 22 Nov. 2021

Background: Today, with the advancement of technology in various fields, the importance of recording data in the field of health is increasing so much that for many diseases around the world, including kidney disease, registration systems have been set up. This is happening in our country and in the future, the number of these systems will increase. The medical data set contains valuable information that will be time-consuming and costly to obtain using laboratory methods, so there is a need for low-cost methods for extracting information. This study focuses on developing a predictive model for classifying the cause of kidney stones in Isfahan using three data mining techniques.

Methods: This cross-sectional research has been done from February 2021 to May 2021. The used medical data set includes information of 353 kidney stone patients in Isfahan. In this study, six target attributes of sodium, phosphate, oxalate, citrate, cysteine and uric acid were identified. The techniques for each of the 6 attributes are used separately. The techniques used in this study were three data mining techniques including random forest (RF), artificial neural network (ANN) and support vector machine (SVM).

Results: The best performance in terms of accuracy is related to support vector machine techniques in uric acid class, support vector machine in oxalate class and neural network in cysteine class. The worst performance is related to the random forest technique in the citrate class. The safest rules with a 66% confidence level are for the citrate and sodium classes, and the least reliable rule with a 50% confidence level is for the oxalate class.

Conclusion: Kidney stones can occur due to various reasons such as low citrate and high calcium oxalate. For example, for citrate, factors such as blood pH (potential of hydrogen), blood sugar and blood pressure are effective. To prevent any of the causes of kidney stones, factors should be controlled.

Keywords: data mining, forecasting, kidney stones.