# کشف الگوهای پنهان موجود در داده‌های بیماران مبتلا به سل

* **فرزاد فیروزی جهانتیغ**: استادیار، گروه مهندسی صنایع، دانشگاه سیستان و بلوچستان، زاهدان، ایران (*نویسنده مسئول). f.firouzi@eng.usb.ac.ir
**بهزاد کیانی**: دانشجوی دکتری تخصصی انفورماتیک پزشکی، کمیته تحقیقات دانشجویی، دانشکده پزشکی، دانشگاه علوم پزشکی مشهد، مشهد، ایران. kiani.Behzad@gmail.com
**ساینا اعتماد**: مرکز تحقیقات سل بالینی و اپیدمیولوژی، دانشگاه علوم پزشکی و خدمات بهداشتی- درمانی شهید بهشتی، تهران، ایران. etemad00@gmail.com

تاریخ دریافت: ۹۴/۹/۲۴ تاریخ پذیرش: ۹۵/۱/۲۹

**چکیده**

**زمینه و هدف**: با تحلیل و بررسی دقیق داده‌های بیماران مبتلا به بیماری خاصی می‌توان به الگوها و دانش غنی موجود در مورد آن بیماری و یا حتی ویژگی‌های خاص خود بیماران مبتلا به آن بیماری د ست یافت. معمولاً در مطالعات پز شکی فر ضیه‌ای در نظر گرفته می شود و سپس به صورت آینده‌نگر برای اثبات یا رد این فر ضیه داده‌هایی جمع آوری می‌گردند، اما در بسیاری از موارد ممکن ا ست بین داده‌های بیماران روابطی وجود دا شته با شد که حتی در مورد آن‌ها تا به حال هیچ حد سی زده ن شده ا ست و طبیعتاً فر ضیه‌ای هم در نظر گرفته ن شده ا ست. لذا، در این مطالعه الگوهای پنهان موجود در مجموعه داده‌های بیماران مبتلا به سل کشف شده است.

**روش کار**: در این مطالعه داده‌های تعداد ۶۰۰ بیمار مبتلا به سل مراجعه کننده به بیمارستان مسیح دانشوری تهران از سال ۱۳۹۲ تا ۱۳۹۳ از طریق خواندن پرونده‌های کاغذی بیماران و استخراج نتایج آزمایش‌های بالینی بیماران از طریق سیستم اطلاعات بیمارستانی جمع‌آوری گردیده است و سپس با استفاده از تکنیک داده‌کاوی APIRIORI با استفاده از ابزار داده‌کاوی WEKA به کشف روابط انجمنی موجود در داده‌ها پرداخته شده است.

**یافته‌ها**: بیماری‌های پرفشاری خون، دیابت بی مزه و ایسکیمیک قلبی، بیماری‌هایی با بیشترین فراوانی در بین مبتلایان به سل در این بانک داده‌ای بوده است. بیمارانی که دیابت بی‌مزه یا تعریق شبانه داشتنه‌اند سرفه مزمن را نیز تجربه کرده‌اند. بیمارانی که آزمایش خلط +BK آن ها یک بوده است و کاهش وزن هم داشته‌اند نیز سرفه مزمن را تجربه کرده‌اند. بیمارانی که خلط خونی داشته‌اند و آزمایش خلط +BK آن ها نیز ۳ بوده است هم سرفه مزمن نیز داشته‌اند. بیماران مردی که تعریق شبانه داشته‌اند دچار کاهش وزن نیز شده‌اند. بیمارانی که هم کاهش وزن و هم تب داشته‌اند تعریق شبانه نیز داشته‌اند.

**نتیجه‌گیری**: قوانین کشف شده می‌توانند برای مطالعات بعدی به خصوص مطالعات کارآزمایی بالینی به عنوان فرضیه‌های اولیه در نظر گرفته شوند. علاوه بر این پزشکان نیز می‌توانند از این قوانین در تحلیل وضعیت بالینی بیماران استفاده نمایند.

**کلیدواژه‌ها**: اصل APIRIORI، داده‌کاوی، سل، قوانین انجمنی، کشف الگو

## مقدمه

مجموعه داده های بیماران در علوم پزشکی بسیار گسترده شده است و داده‌های غنی از بیماران و نشانه‌های بالینی آن‌ها وجود دارد. با تحلیل و بررسی موشکافانه داده‌های بیماران مبتلا به بیماری خاصی می‌توان به الگوها و دانش غنی موجود در مورد آن بیماری و یا حتی ویژگی‌های خاص خود بیماران مبتلا به آن بیماری دست یافت. معمولاً در مطالعات پزشکی فرضیه‌ای در نظر گرفته می‌شود و سپس به‌صورت آینده‌نگر برای اثبات یا رد این فرضیه داده‌هایی جمع‌آوری می‌گردند، اما در بسیاری از موارد ممکن است بین داده‌های بیماران روابطی وجود داشته باشد که حتی در مورد آن‌ها تا به‌حال هیچ حدسی زده نشده است. برای کشف این‌گونه الگوها در داده‌ها می‌توان از تکنیک‌های داده‌کاوی استفاده نمود. داده‌کاوی و کشف دانش در داده‌ها، رویکردی برای یافتن روابط و الگوهای پنهان در داده‌ها است (۱). امروزه داده کاوی در بسیاری از مطالعات علوم پزشکی شامل تشخیص بیماری‌ها (۲ و ۳)، کشف الگوهای پنهان موجود در داده‌ها (۴) و غیره ا ستفاده شده ا ست. ایده‌های جدید مانند ک شف دانش از پایگاه داده (Knowledge Discovery and Data Mining) که شامل تکنیک های

[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

داده‌کاوی هستند، امروزه محبوبیت بیشتری یافته و به یک ابزار تحقیقاتی مطلوب برای پژوهشگران مبدل شده‌اند. به کمک آن‌ها پژوهشگران می‌توانند الگو ها و روابط بین تعداد زیادی از متغیرها را شناسایی کرده و پیش‌بینی نتایج حاصل از یک بیماری با استفاده از ذخایر اطلاعاتی موجود در پایگاه‌های داده برای آن‌ها امکان‌پذیر گشته است (۵).

رویکردهای اصلی در حوزه داده‌کاوی شامل دسته‌بندی (Classification)، خوشه‌بندی (Clustering) و کشف روابط انجمنی (Association rule discovery) است. دسته‌بندی شامل تکنیک‌هایی از داده‌کاوی است که در آن‌ها مجموعه‌ای از رکوردها با یک فیلد برچسب موجود است. در دسته‌بندی به دنبال مدلی هستیم که به رکوردی با برچسب ناشناخته برچسب مناسب را نسبت دهیم (۶). در خوشه‌بندی برخلاف دسته‌بندی، رکوردها دارای فیلد برچسب نیستند و مدل داده‌کاوی به دنبال این است که رکوردهای مشابه را شناسایی نماید (۷). قوانین انجمنی نیز وابستگی‌ها و روابط بین مجموعه‌ی داده‌ها را در یک بانک داده‌ای مشخص می‌نمایند. یافتن چنین قوانینی در حوزه های مختلفی مانند پزشکی کاربردهای فراوانی دارد (۸). برای نمونه با تحلیل داده‌های بیماران و کشف روابط انجمنی می‌توان مشاهده نمود که بیمارانی که علامت x را داشته‌اند، علامت y را نیز بروز داده‌اند و یا بین شغل بیماران و نوع بیماری آن‌ها رابطه وجود داشته است که در ابتدا اصلاً چنین فرضیه‌هایی مطرح نبوده است.

بیماری سل یکی از شایع‌ترین بیماری‌های مزمن موجود در جهان است که سالانه جان تعداد زیادی از انسان‌ها را می‌گیرد (۹). در سال ۱۳۹۲ هجری شمسی در ایران یک مطالعه داده‌کاوی برای تشخیص بیماری سل انجام شده است که البته این مطالعه بر روی داده‌های استاندارد انجام شده است (۱۰) و به کشف قوانین انجمنی نپرداخته است، بلکه یک مدل تشخیصی ارائه کرده است. در مورد مطالعات داده‌کاوی که در حوزه تشخیص یک بیماری انجام می‌گردد باید به این نکته مهم توجه نمود که هرگز چنین مدل‌هایی جایگزین پزشک نخواهند بود بلکه این‌ها نقش یک سیستم تصمیم‌یار (Decision Support System) را برای پزشک ایفا می‌نمایند. مطالعات خارجی متعددی نیز در حوزه داده‌کاوی روی داده‌های بیماری سل انجام گرفته است (۱۳-۱۱). در مطالعه حاضر برخلاف بسیاری از مطالعات داده‌کاوی دیگر، داده‌ها توسط پژوهشگران صرفاً برای انجام این مطالعه جمع‌آوری گردیده است. لذا، داده‌های تمیزی برای انجام داده‌کاوی جمع‌آوری شده است و مانند بسیاری از مطالعات داده‌کاوی دیگر که به صورت گذشته‌نگر و بر روی داده‌های موجود و جمع‌آوری شده برای سایر اهداف بوده، نخواهد بود و به همین دلیل بسیاری از مراحل پیش‌پردازش داده ها حذف می‌گردند. علاوه بر این در این مطالعه سایر بیماری‌هایی که بیماران مبتلا به سل به آن‌ها مبتلا بوده‌اند نیز در نظر گرفته شده است تا روابط احتمالی موجود بین انواع بیماری‌ها نیز در صورت وجود به دست آیند. در این مطالعه با استفاده از تکنیک‌های داده کاوی، الگوهای پنهان موجود بین داده های بیماران مبتلا به سل کشف خواهد شد. این الگوها می‌توانند در اختیار محققان قرار بگیرند و مطالعات تکمیلی برای رد یا اثبات آن‌ها انجام گیرد؛ زیرا داده کاوی قابلیت اثبات ندارد و فقط الگوهای پنهان را کشف می‌نماید. از سویی دیگر این الگوها می‌توانند در اختیار پزشکان قرار گرفته و در فرآیند تحلیل وضعیت بیماران به کار گرفته شوند. لذا، هدف این مطالعه بررسی داده‌های بیماران مبتلا به سل و کشف روابط پنهانی احتمالی موجود در بین این داده‌ها است.

## روش کار

**جمع آوری داده ها:** در این پژوهش داده های تعداد ۶۰۰ بیمار مبتلا به سل در بیمارستان مسیح دانشوری شهر تهران از سال ۱۳۹۲ تا



[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

جدول ۱- فیلدهای جمع‌آوری شده از بیماران مبتلا به سل در بیمارستان مسیح دانشوری تهران

| فیلد داده‌ای | توضیح | فیلد داده‌ای | توضیح |
|---|---|---|---|
| سن | سن بیمار را مشخص می‌کند (سال) | بیمار مبتلا به ویروس نقص ایمنی انسانی | آیا بیمار مبتلا به ایدز است؟(بله-خیر) |
| شغل | شغل بیمار را مشخص می‌کند. | خلط + BK | تعداد آزمایش‌های اسمیر خلط مثبت بیمار را نشان می‌دهد (عددی بین صفر تا ۳) |
| جنسیت | جنسیت بیمار را مشخص می‌کند (مرد-زن) | تعداد گلبول‌های سفید | تعداد گلبول‌های سفید در یک میلی‌لیتر خون بیمار را نمایش می‌دهد. |
| سرفه مزمن | آیا بیمار سرفه مزمن داشته است (بله- خیر) | هموگلوبین خون | مقدار هموگلوبین موجود در خون برحسب گرم در دسی لیتر را نشان می‌دهد. |
| خلط آغشته به خون | آیا بیمار خلط آغشته به خون داشته است (بله-خیر) | PLATELET | تعداد پلاکت‌ها در هر میلی‌لیتر مکعب خون را نشان می‌دهد. |
| کاهش وزن | آیا بیمار کاهش وزن داشته است (بله- خیر) | Erythrocyte Sedimentation Rate | سرعت رسوب گلبول‌های قرمز خون را برحسب میلی‌متر در ساعت نشان می‌دهد. |
| تعریق شبانه | آیا بیمار تعریق شبانه داشته است (بله-خیر) | Fasting Blood Sugar | میزان قند خون ناشتا بیمار را برحسب میلی‌گرم بر میلی‌لیتر نشان می‌دهد. |
| تب | آیا بیمار تب داشته است (بله-خیر) | کراتین | مقدار کراتینین موجود در خون را برحسب میلی‌گرم در دسی لیتر نشان می‌دهد. |
| سابقه تماس با بیمار مبتلا به سل | آیا بیمار در گذشته با فردی مبتلا به سل تماس داشته است (بله-خیر) | آلبومین | مقدار آلبومین موجود در خون بیمار را برحسب گرم بر دسی لیتر نشان می‌دهد. |
| سابقه استفاده از سیگار | آیا بیمار سیگاری است؟ (بله-خیر) | سایر بیماری‌های بیمار | اگر بیمار به بیماری‌های دیگری نیز مبتلا است در اینجا نام این بیماری‌ها ذکر شده است. |
| سابقه استفاده از الکل | آیا بیمار الکل استفاده کرده است (بله-خیر) | | |

۱۳۹۳ جمع‌آوری گردیده است. داده‌ها از طریق مراجعه به پرونده‌های کاغذی بیماران در بایگانی بیمارستان و استخراج شرح‌حال، خلاصه پرونده بیمار و فرم ارجاع به دست آمده است. علاوه بر این تک تک جواب آزمایش های بیماران نیز از طریق سیستم اطلاعات بیمارستانی (Health Information System) بیمارستان استخراج گردیده است. فیلدهای جمع‌آوری شده در جدول شماره یک توضیح داده شده است.

**گسسته سازی داده‌ها:** برای کشف الگوهای پنهان موجود در داده‌ها ابتدا باید داده‌ها به فرمت مناسب تبدیل شوند. برای این منظور تمامی فیلدهای عددی را به فیلدهای گسسته تبدیل نموده‌ایم. از روش گسسته‌سازی بازه‌هایی به طول یکسان (Equal-width interval discretization) برای گسسته سازی فیلدهایی که مقادیر پیوسته داشته‌اند استفاده شده است (۱۴).

**ایجاد ویژگی های جدید از داده ها:** فیلد مشخص‌کننده‌ی سایر بیماری‌هایی که بیمار به آن‌ها مبتلا است را باید به چندین فیلد دیگر تبدیل نماییم. برای این منظور تمامی بیماری‌های موجود در این فیلد را استخراج نمودیم و سپس به تعداد این بیماری‌ها، فیلدهای جدیدی با نام هر بیماری ایجاد نمودیم. برای هر بیمار، اگر بیماری موردنظر را دارد درون این فیلد کلمه yes و در غیر این صورت این فیلد را خالی قرار دادیم چرا که اگر مقدار این فیلد را در حالتی که بیماری موجود نیست no قرار دهیم تعداد no ها بسیار بیشتر از تعداد yes ها می‌گردد و بیشتر قوانین کشف شده بر پایه مقادیر no به وجود خواهند آمد و معنی‌دار نخواهند بود. جدول شماره ۲ متغیرهای دیگری را که بر پایه این فیلد به وجود



[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

جدول ۱- سایر بیماری‌های بیماران مبتلا به سل در مجموعه داده‌های بیماران

| | |
|---|---|
| پرفشاری خون | Hypertension |
| دیابت بی‌مزه | Diabetes mellitus |
| نارسایی مزمن کلیوی | Chronic Renal Failure |
| انواع سرطان | Cancer |
| نارسایی احتقانی قلب | Congestive heart failure |
| بیماری‌های روانی | Psychological disease |
| انواع هپاتیت | Hepatitis |
| بیماری‌های مرتبط با تیروئید | Thyroid Diseases |
| بیماری ایسکیمیک قلبی | ischemic heart disease |
| بیماری مزمن ریوی | Asthma |
| بیماری هپاتیت C | hepatitis C disease |
| بیماری انسدادی مزمن ریوی | Chronic obstructive pulmonary disease |
| سکته مغزی | cerebral vascular accident |
| بیماری هپاتیت B | hepatitis B disease |
| بیماری‌های قلبی عروقی | Cardiovascular disease |
| سندرم نقص ایمنی (ایدز) | acquired immunodeficiency syndrome |
| بیماری کورپولمونل | Corpulmonale disease |

آورده‌ایم نشان می‌دهد.

**اعمال الگوریتم APPRIORI جهت کشف روابط انجمنی:** در مرحله بعدی از الگوریتم APPRIORI برای کشف روابط انجمنی استفاده می‌گردد (۱۵). در این الگوریتم دو معیار Confidence و Support کیفیت یک قانون انجمنی کشف شده از درون داده‌ها را مشخص می‌نمایند. Support در واقع احتمال حضور همزمان X,Y را در تراکنش X→Y مشخص می‌کند و Confidence هم احتمال وجود Y را در صورت وجود X نشان می‌دهد (۱۶).

الگوریتم APPRIORI توسط ابزار داده کاوی WEKA بر روی داده‌ها اعمال گردیده است و تعداد ۱۰۰ قانون انجمنی با بیشترین مقدار Confidence از مجموعه داده‌ها کشف گردیده است. ازآنجایی‌که برخی از این قوانین به لحاظ پزشکی بی‌معنی هستند، این قوانین به پزشک متخصص در حوزه ریه نشان داده شد و در نهایت تعداد ۶ قانون به‌عنوان قوانین انجمنی موجود و مرتبط در داده‌ها استخراج گردید. جهت ارزیابی قوانین انجمنی و میزان مناسب بودن آن‌ها از معیار ارزیابی Confidence و Support استفاده شده است.

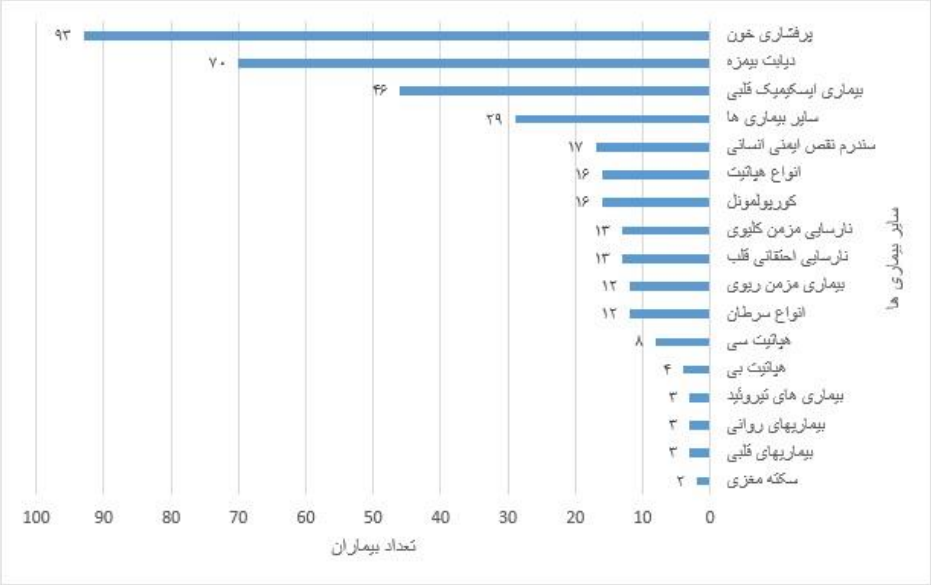
## یافته‌ها

پس از دسته‌بندی سایر بیماری‌هایی که مبتلایان به سل در مجموعه داده‌ها به آن‌ها مبتلا بوده‌اند، نتایج زیر به دست آمده است که در شکل شماره یک نشان داده شده است. پس از اعمال الگوریتم APIRIORI نیز به داده های بیماران تعداد ۱۰۰ قانون انجمنی با بیشترین مقدار Confidence به‌عنوان قوانین انجمنی منتخب جهت بررسی توسط متخصص انتخاب شده است. این تعداد قانون انجمنی همگی در اختیار متخصص قرار گرفته است و متخصص مربوطه پس از ارزیابی و بررسی این قوانین تعداد ۶ قانون را به‌عنوان قوانین مرتبط و با معنی در نظر گرفته است. قوانین مذکور در جدول شماره ۳ نشان داده شده است.

بر طبق شکل فوق مشخص است که بیماری‌های پرفشاری خون، دیابت بی‌مزه و بیماری ایسکیمیک قلبی بیماری‌هایی با بیشترین فراوانی هستند که مبتلایان به سل به آن‌ها نیز مبتلا بوده‌اند. در جدول شماره ۳ قوانین انجمنی کشف شده در



[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]




شکل ۱- فراوانی مطلق سایر بیماری‌ها در بیماران مبتلا به سل

جدول ۲- قوانین انجمنی بین داده‌های بیماران مبتلا به سل مراجعه کننده به بیمارستان مسیح دانشوری تهران

| الگوی کشف شده | مقدار Confidence الگو |
|---|---|
| بیمارانی که خلط خونی داشته‌اند و آزمایش خلط +BK آن‌ها نیز سه بوده است، سرفه مزمن نیز داشته‌اند. | ۱ |
| بیمارانی که کاهش وزن داشته‌اند و آزمایش خلط +BK آن‌ها نیز یک بوده است، سرفه مزمن را نیز تجربه کرده‌اند. | ۱ |
| بیمارانی که دیابت بی‌مزه داشته‌اند، سرفه مزمن را نیز تجربه کرده‌اند. | ۰/۹۹ |
| بیمارانی که تعریق شبانه داشته‌اند، سرفه مزمن را نیز تجربه کرده‌اند. | ۰/۹۶ |
| بیمارانی که مرد بوده‌اند و تعریق شبانه داشته‌اند، کاهش وزن را نیز تجربه کرده‌اند. | ۰/۹۲ |
| بیمارانی که کاهش وزن و تب داشته‌اند، تعریق شبانه نیز داشته‌اند. | ۰/۹۱ |

بین داده‌های بیماران مبتلا به سل آمده است.

طبق جدول شماره ۳ مشخص است که بیمارانی که دیابت بی‌مزه یا تعریق شبانه دا شته‌اند، سرفه مزمن را نیز تجربه کرده‌اند. بیمارانی که آزمایش خلط +BK آن‌ها یک بوده است و کاهش وزن هم داشته‌اند، نیز سرفه مزمن را تجربه کرده‌اند. بیمارانی که خلط خونی دا شته‌اند و آزمایش خلط +BK آن‌ها نیز ۳ بوده است هم سرفه مزمن نیز داشته‌اند. بیماران مردی که تعریق شبانه داشته‌اند دچار کاهش وزن نیز شده‌اند؛ و بیمارانی که هم کاهش وزن و هم تب داشته‌اند تعریق شبانه نیز داشته‌اند.

## بحث و نتیجه‌گیری

مطالعات داده‌کاوی در پزشکی نسبتاً جدید است و در جستجویی که پژوهشگران این مطالعه انجام داده‌اند، مطالعه‌ی مشابه که به بررسی قوانین انجمنی در بین داده‌های بیماران مبتلا به سل انجام شده باشد، یافت نشده است.

ا ستفاده از تکنیک‌های داده‌کاوی به‌خ صوص در داده‌های پزشکی با توجه به اینکه معمولاً حجم بالایی دارند و روابط ناشناخته زیادی بین علل بیماری‌ها و یا مشخصات دموگرافیک افراد و ریزفاکتورهای خطر ابتلا به بیماری‌ها وجود دارد، مفید است. در مطالعات زیادی در حوزه پزشکی علاوه بر روش های آ ماری از تکنیک های متعدد داده‌کاوی مانند دسته‌بندی، خوشه‌بندی و یا قوانین انجمنی استفاده گردیده است (۱۷). در پژوهش‌های داده‌کاوی مواردی به دست می‌آیند که ممکن است در مورد آن‌ها تابه‌حال هیچ فکری



[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

نشـده باشـد و پس از به دسـت آمدن این نتایج می‌توان آن ها را از طریق تکنیک های آماری و طراحی مطالعات دیگری اثبات یا رد نمود؛ چرا که داده کاوی فقط روابط بین مجموعه‌ی داده های یک بانک داده‌ای مشـخص را به دسـت می‌آورد و لزوماً این روابط در داده های دیگر ممکن اسـت برقرار نباشد. در مطالعه حاضر هم به کشف روابط انجمنی موجود بین داده‌های بیماران مبتلا به سل که به بیمارستان مسیح دانشوری مراجعه نموده بودند، پرداخته شده است. قوانین انجمنی کشف شده در این مطالعه در این مجموعه داده‌ها برقرار اسـت ولی می‌تواند به‌عنوان فرضـیه های برای مطالعات دیگر نیز اسـتفاده گردد و در مطالعات دیگری به بررسی این روابط پرداخته شود.

پس از انجام تکنیک داده کاوی APIRIORI و استخراج قوانین با بیشـترین مقدار Confidence مشاهده شد که بسیاری از قوانین از طریق اصل APIRIORI قابل حذف شـدن هسـتند. اصـل APIRIORI می‌گوید اگر قانون انجمنی $X \rightarrow Y$ موجود باشـد و X زیرمجموعه Z باشـد، آنگاه بدیهی اسـت که قانون $Z \rightarrow Y$ برقرار اسـت و این قانون باید از مجموعه قوانین کشـف شـده حذف گردند (۱۸). در نتایج اولیه این مطالعه قوانین این‌چنینی فراوان وجود داشـت که با توجه به اینکه خود تکنیک APIRIORI توانایی حذف این قوانین را نداشت به‌صورت دستی این قوانین حذف گردیدند.

علامت‌های بیماری سل بستگی به این دارد که میکروب به کدام عضـو حمله کرده اسـت. معمولاً میکروب سل ریه‌ها را درگیر می‌کند و علامت‌های سل ریوی شامل: سرفه شدید بیش از ۳ هفته، درد قفسه سینه و خلط خونی. سایر علائم عبارتند از: ضـعف یا خسـتگی زیاد، کاهش وزن، از دست دادن اشتها، تب، لرز و عرق شبانه که البته گاهی ممکن اسـت با هیچ کدام از علائم فوق همراه نباشـد؛ بلکه بدون علامت و تنها از طریق عکس اتفاقی ریه تشـخیص داده شـود و با اینکه بدون سـرفه و علائم ریوی و تنها با کاهش وزن همراه باشد، ممکن است فرد علائمی مثل دردی که در ناحیه قفسه سینه است و با نفس کشیدن و سرفه ایجاد و تشدید می‌شود، داشته باشد. با نگاهی به جدول شماره ۳ در این پژوهش مشخص است که تقریباً تمامی علائم اصلی بیماری سل شامل خلط خونی، تب، کاهش وزن و تعریق شـبانه همراه با سـرفه مزمن نیز خواهند بود. لذا، پزشـکان در تشخیص بیماری سل در صورت وجود این علائم می‌توانند با در نظر گرفتن احتمال بالای به وجود آمدن سـرفه مزمن در آینده‌ای نزدیک برای بیمار به تشخیص این بیماری بپردازند.

یکی از محدودیت‌های روش داده‌کاوی کشـف قوانین انجمنی این است که قوانین کشف شده در این مجموعه از داده‌ها برقرار است و ممکن است این قوانین در مجموعـه‌ی دیگری از داده‌هـای بیماران مبتلا به سل برقرار نباشد یا تغییر نماید. برای این منظور بهتر اسـت که چندین بانک داده‌ای با هم تلفیق شـوند تا قوانین انجمنی قابل اطمینانی به دست آیند.

یکی دیگر از محدودیت های اصـلی دیگر این مطالعه تعداد بیماران مورد مطالعه اسـت که در این مطالعه ۶۰۰ بیمار بوده اسـت، اما معمولاً در مطالعات داده‌کاوی تعداد داده‌ی زیاد لازم اسـت. برای این منظور سیستم‌های اطلاعات پزشکی و بیمارسـتانی در کشـور ما باید تقویت شـود و به بحث ذخیره‌سـازی داده‌های بیماران و پرونده‌های الکترونیک بسیار بیشتر توجه شود چرا که وقتی چنین سـیسـتم‌هایی برای ذخیره‌سـازی داده‌های بالینی بیماران وجود نداشـته باشـد، عملاً امکان انجام مطالعات داده کاوی با حجم عظیمی از داده‌ها امکان‌پذیر نیست.

در این مطـالعـه قوانین انجمنی پنهـان در داده‌های بیماران مبتلا به سـل مراجعه‌کننده به بیمارستان مسیح دانشوری تهران با استفاده از تکنیک APIRIORI کشـف گردیدند. از دید فنی همین که قانونی Support و Confidence حداقل را داشـته باشـد یک قانون انجمنی معنی‌دار در مجموعه داده‌ها اسـت اما قوانینی که در ابتدا در



[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

این مطالعه کشف شدند کاملاً نشان دادند که نظر کاربر انسانی و به‌خصوص فرد متخصص در حوزه موردنظر نیز بعد از معیارهای تکنیکی برای پالایش قوانین لازم است. برای نمونه یکی از قوانینی که مقدار Confidence آن نیز ۰/۹۱ بوده است نشان داده که افرادی که سیگار می‌کشیده‌اند، مرد بوده‌اند. کاملاً واضح است که چنین قانونی ارزش خاصی به لحاظ پزشکی ندارد و توسط فرد خبره باید فیلتر گردد. پس از حذف چنین قوانینی تعداد ۶ قانون انجمنی مورد تأیید در مجموعه داده‌های بیماران شناسایی شد.

سرفه مزمن در افرادی که بیماری سل دارند تقریباً با تمامی علائم دیگر بیماری همراه است. در صورت به وجود آمدن علائم دیگر بیماری سل می‌توان سرفه مزمن را نیز به آن علائم اضافه کرد؛ چرا که با احتمال بالایی به وجود خواهد آمد و سپس با در نظر گرفتن احتمال بالای به وجود آمدن سرفه مزمن به تشخیص بیماری پرداخت.

در مطالعه حاضر از داده‌های ۶۰۰ بیمار استفاده گردیده است. در مطالعات داده‌کاوی هر چه تعداد داده‌ها بیشتر باشد نتایج الگوریتم‌های کشف دانش بهتر خواهند بود. لذا، پیشنهاد می‌شود تا در مطالعات بعدی از داده‌های بیشتری با ترکیب داده‌های مختلف از بیمارستان‌های مختلف کشور بهره گرفته شود.

مطالعه حاضر بر روی بیماران مبتلا به سل انجام شده بود و ازآنجایی‌که همه آن‌ها مبتلا به سل بوده‌اند امکان ایجاد مدل‌های داده‌کاوی بر پایه تکنیک‌های دسته‌بندی میسر نبوده است. پیشنهاد می‌گردد یک مطالعه جهت ایجاد مدل دسته‌بندی برای تشخیص بیماری سل در بین افراد مشکوک به این بیماری از طریق در نظر گرفتن داده‌های بیماران مبتلا به سل و همچنین داده‌های بیمارانی که مشکوک به سل بوده‌اند اما مبتلا نیستند، انجام شود.

ازآنجایی که نتایج حاصل از یک مطالعه داده‌کاوی به دلیل اینکه بر روی مجموعه خاصی از داده‌ها انجام می‌گردد را به سختی می‌توان تعمیم داد، پیشنهاد می‌کنیم مطالعات مرور نظام‌مند (Systematic Review) و در صورت کمی بودن نتایج، مطالعات متا آنالیز (Meta-Analysis) بر روی مطالعات داده‌کاوی که در دنیا در حوزه یک بیماری یا بیماران خاصی انجام شده است انجام گیرد تا با تحلیل نتایج حاصل از مطالعات متعددی بتوان نتایج کلی را به دست آورد. برای نمونه می‌توان یک نمونه مطالعه مرور نظام‌مند بر روی کلیه مطالعات داده‌کاوی انجام شده در حوزه بیماری سل طراحی و اجرا نمود.

## منابع


1. Wu X, Zhu X, Wu GQ, Ding W. Data mining with big data. Knowledge and Data Engineering, IEEE Transactions on; 2014. 26(1):97-107.
2. Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications–A decade review from 2000 to 2011. Expert Systems with Applications; 2012. 39(12): 11303-11.
3. Shouman M, Turner T, Stocker R. Using data mining techniques in heart disease diagnosis and treatment. In: Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on. IEEE; 2012.
4. Sim DYY, the CS, Ismail AI. Adaptive apriori and weighted association rule mining on visual inspected variables for predicting obstructive sleep apnea. Australian Journal of Intelligent Information Processing Systems; 2014.14(1):17-25.
5. Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine; 2005. 34(2): 113-۱27.
6. Buczak AL. Fuzzy association rule mining and classification for the prediction of malaria in South Korea. BMC Med Inform Decis Mak; 2015.15(1): 47.
7. Banjari I, Banjari I, Kenjerić D, Šolić K, Mandić ML. Cluster analysis as a prediction tool for pregnancy outcomes. Coll Antropol; 2015. 39(1):247-52.
8. Wei L, Scott J. Association rule mining in the US Vaccine Adverse Event Reporting System (VAERS). Pharmacoepidemiol Drug Saf; 2015.
9. Sulis G. Tuberculosis: Epidemiology and Control. Mediterranean journal of hematology and infectious diseases; 2014. 6(1):27-34.
10. Daliri Shams Abadi H, EbrahimPour Kumle H. Check the forecast tuberculosis using smart data mining algorithms. Natinal Conference on




[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

Computer Engineering and Information Technology, Islamic Azad university, Shushtar, Iran; 1392. [Persian]

11. Shukla M, Agarwal S. Hybrid approach for tuberculosis data classification using optimal centroid selection based clustering. in Engineering and Systems (SCES), 2014 Students Conference on. IEEE; 2014.

12. Jain A, Pardasani KR. Mining fuzzy amino acid associations in peptide sequences of mycobacterium tuberculosis complex (MTBC). Network Modeling Analysis in Health Informatics and Bioinformatics; 2015. 4(1):1-14.

13. Rastogi N, Couvin D. Phylogenetic associations with demographic, epidemiological and drug resistance characteristics of Mycobacterium tuberculosis lineages in the SITVIT2 database: Macro-and micro-geographical cleavages and phylogeographical specificities. International Journal of Mycobacteriology; 2014. 5(3):65-72.

14. Joiţa D. Unsupervised static discretization methods in data mining. Titu Maiorescu University, Bucharest, Romania; 2010.

15. Abdullah Z, Herawan T, Chiroma H, Mat Deris M. A sequential data preprocessing tool for data mining, in computational science and its applications–ICCSA; 2014. 2014, Springer. p. 734-746.

16. Baralis E. Generalized association rule mining with constraints. Information Sciences; 2012. 194: 68-84.

17. Tseng WT, Chiang WF, Liu ShY, Roan J, Lin ChN. The application of data mining techniques to oral cancer prognosis. J Med Syst; 2015. 39(5): 59.

18. Prasenna P. Network programming and mining classifier for intrusion detection using probability classification. in Pattern Recognition, Informatics and Medical Engineering (PRIME), 2012 International Conference on. IEEE; 2012.

[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]

# Discovering hidden patterns available in data of patients with tuberculosis

**Behzad Kiani**, PhD candidate in Medical Informatics, Student Research Committee, Faculty of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran.
***Farzad Firouzi Jahantigh**, Assistant Professor, Department of Industrial Engineering, Faculty of Engineering, University of Sistan and Baluchestan, Zahedan, Iran (*Corresponding author). f.firouzi@eng.usb.ac.ir
**Saina Etemad**, Clinical Tuberculosis and Epidemiology Research Center Shahid Beheshti University of Medical Science, Tehran, Iran.

## Abstract

**Background**: Having precisely analyzed the data of patients with specific diseases we can obtain the patterns and knowledge of these disease or even specific characteristics of patients. A hypothesis is usually considered in medical studies when some data are gathered prospectively to prove or deny this hypothesis, but in many cases there may be relationships between the data of the patients which have never been attended and no hypothesis has been considered. Thus, in this study available hidden patterns in the data of patients with tuberculosis have been discovered.

**Methods**: Data of the study included 600 patients with tuberculosis who had referred to Masih Daneshvari hospital of Tehran. Data were gathered by reading patients files and observing clinical tests of patients from hospital data system. APPIRIORI data mining technique and WEKA tool of data mining were utilized to discover the associative relationships of the data.

**Results**: Hypertension diseases, diabetes insipidus and ischemic heart disease have had the most frequency in patients with tuberculosis. Patients with diabetes insipidus or night sweats had also experienced chronic cough. Patients who have had weight loss and had BK+ test one result, had also experienced chronic cough. Patients who have been coughing up blood and had BK+ sputum tests (3), had also experienced chronic cough. Male patients who had night sweats, had also experienced weight loss. Patients who have had weight loss and fever, had also experienced night sweats.

**Conclusion**: Discovered rules can be considered as primary hypothesis for the upcoming studies especially those of clinical trials, In addition to this, physicians can use these rules to analyze the clinical condition of patients.

**Keywords**: APIRIORI principle, Associative rules, Data mining, Pattern discovery, Tuberculosis

[ Downloaded from rjms.iums.ac.ir on 2025-05-17 ]