مجله مدیریت اطلاعات سلامت
سال هجدهم/ شماره پنجم/ آذر و دی ۱۴۰۰
Published by *Vesnu Publications*
DOI: 10.22122/him.v18i5.4367

مزایا و چالش‌های کاوش کلان‌داده‌های پزشکی

لیلا برادران سرخابی[۱]، فرهاد سلیمانیان قره‌چپق[۲]، جعفر شهام‌فر[۳]

مقاله مروری

چکیده

مقدمه: داده‌کاوی، ابزار کارامدی جهت آشکارسازی دانش نهفته در کلان‌داده‌های پزشکی می‌باشد. اولین قدم داده‌کاوی، شناخت داده و چالش‌های آن است. هدف از انجام پژوهش حاضر، بررسی سرمنشأ، تأثیرات و راهکارهای مواجهه با چالش‌های کاوش کلان‌داده‌های پزشکی و همچنین، تعیین منافع حاصل از کاوش بود.

روش بررسی: در این تحقیق مروری، مطالعات انگلیسی با دو گروه کلید واژه مجزا برای مزایا و چالش‌ها از پایگاه‌های اطلاعاتی PubMed، ScienceDirect، Springer و Google Scholar، طی بازه زمانی سال‌های ۲۰۱۱ تا ۲۰۲۰ جستجو شد. مطالعات تک منظوره حذف و مطالعاتی که به صورت جامع کاوش کلان‌داده‌های پزشکی را مورد بررسی قرار داده بودند، انتخاب شد. سپس هر چالش مورد بررسی دقیق‌تر قرار گرفت و نتایج به صورت طبقه‌بندی شده ارایه گردید.

یافته‌ها: دانش حاصل از کاوش کلان‌داده پزشکی، سبب افزایش کیفیت ارایه خدمات درمانی می‌شود، اما خطا در جمع‌آوری و ثبت اطلاعات، ویژگی‌های ناشی از کلان‌داده بودن و ساختار ذاتی داده‌های پزشکی، چالش‌های بسیاری بر سر راه کاوش قرار داده است که از بین آن‌ها، «ناسازگاری، صحت، امنیت و محرمانگی داده»، دشوارترین مشکلات به شمار می‌روند. استانداردسازی و افزایش دقت و امنیت در جمع‌آوری، ذخیره‌سازی و نمایش داده‌ها، مؤثرترین راهکارهای پیشگیری می‌باشد. طراحی و استفاده از بسترها، الگوریتم‌ها و ساختارهای مناسب کلان‌داده و همچنین، بهره‌گیری از روش‌های یادگیری ماشین و هوش مصنوعی، راهکارهای مناسبی برای مواجهه با چالش‌ها محسوب می‌شوند.

نتیجه‌گیری: عدم آمادگی برای ظهور کلان‌داده‌های پزشکی و رشد بسیار سریع آن‌ها، سرمنشأ بروز چالش‌هایی برای الگوریتم‌های کاوش هستند که برخی قابل پیشگیری، شناسایی و رفع می‌باشند و برخی نیز به روش‌های هوشمند نوینی نیاز دارند که قابلیت مدیریت کلان‌داده‌های پزشکی را داشته باشند.

واژه‌های کلیدی: داده‌کاوی؛ کلان‌داده؛ سلامت

پیام کلیدی: کاوش کلان‌داده‌های پزشکی می‌تواند مرزهای علم پزشکی را جابه‌جا نماید و برای بیماران، دولت‌ها، صنعت بیمه، کادر درمان و جوامع سودمند است. هرچند استخراج دانش پنهان در این داده‌ها، ایده مهیج و قابل دسترسی به نظر می‌رسد، اما هنوز چالش‌های بسیاری در مسیر روش‌های داده‌کاوی وجود دارد که ارایه راهکارهای بهینه برای رفع آن‌ها، بحثی رایج در هر دو حوزه کامپیوتر و پزشکی می‌باشد. در تحقیق حاضر، این چالش‌ها بررسی و دلایل بروز، تأثیرات و راهکارهای رایج برای هر کدام نیز مطرح گردید.

دریافت مقاله: ۱۴۰۰/۴/۲۸ پذیرش مقاله: ۱۴۰۰/۹/۱۴ تاریخ انتشار: ۱۴۰۰/۹/۱۵

ارجاع: برادران سرخابی لیلا، سلیمانیان قره‌چپق فرهاد، شهام‌فر جعفر. **مزایا و چالش‌های کاوش کلان‌داده‌های پزشکی.** مدیریت اطلاعات سلامت ۱۴۰۰؛ ۱۸ (۵): ۲۳۳-۲۲۵

مقدمه

داده‌های پزشکی که در ابتدا جهت مکانیزاسیون پرونده‌های پزشکی ذخیره می‌شدند، اکنون به منبع ارزشمندی برای سیاست‌گذاری، فرهنگ‌سازی، آمارگیری، پیشگیری از همه‌گیری‌ها، الگوهای آموزشی و هوشمندسازی ارایه خدمات درمانی تبدیل شده‌اند. در دو دهه اخیر، برخی از عوامل همچون توسعه سیستم‌های اطلاعاتی بیمارستانی، بهره‌گیری از حسگرهای پزشکی، افزایش زیرساخت‌های ارتباطی، همه‌گیری استفاده از ابزارهای هوشمند (مانند گوشی‌های هوشمند) و گسترش ارایه خدمات پزشکی به صورت اینترنتی، باعث شده‌اند که داده‌های پزشکی به نمونه بارز کلان‌داده تبدیل گردد (۱). با وجود پتانسیل اطلاعاتی بسیار ارزنده‌ای که این کلان‌داده‌ها دارند، به صورت مستقیم قابل بهره‌برداری نیستند و لازم است که ابتدا دانش نهفته آن‌ها استخراج شود. داده‌کاوی ابزار کارامدی جهت آشکارسازی اطلاعات نهفته می‌باشد که اولین مرحله آن، شناخت دقیق چالش‌ها است. علاوه بر چالش‌های معمول کاوش کلان‌داده‌ها (۲)، ویژگی‌های ذاتی داده‌های پزشکی نیز سبب بروز مشکلات عمده‌ای برای روش‌های کاوش می‌شوند (۳).

پژوهش‌های گسترده‌ای در زمینه تأثیر خصوصیات کلان‌داده‌های پزشکی در بروز چالش‌های کاوش ارایه شده است (۴). در یکی از این مطالعات، ضمن مرور انواع روش‌های مدیریت کلان‌داده‌های سلامت در مجلات معتبر جهان طی

مقاله حاصل پایان‌نامه مقطع دکتری تخصصی به شماره ۱۰۳۴۱۰۰۶۹۷۱۰۰۲ می‌باشد که با حمایت دانشگاه آزاد اسلامی واحد ارومیه انجام شده است.

۱- دانشجوی دکتری تخصصی، مهندسی نرم‌افزار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه، ایران
۲- استادیار، مهندسی نرم‌افزار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی،ارومیه، ایران
۳- استادیار، مهندسی نرم‌افزار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی، ارومیه و پزشکی اجتماعی، گروه پزشکی اجتماعی، دانشکده پزشکی، دانشگاه علوم پزشکی تبریز، تبریز، ایران
نویسنده طرف مکاتبه: فرهاد سلیمانیان قره‌چپق؛ استادیار، مهندسی نرم‌افزار، گروه مهندسی کامپیوتر، واحد ارومیه، دانشگاه آزاد اسلامی،ارومیه، ایران
Email: bonab.farhad@gmail.com

سال‌های ۲۰۰۶ تا ۲۰۱۶، مؤثرترین و بی‌تأثیرترین عوامل دخیل در تحلیل این داده‌ها نیز تعیین شد (۵). در مرور جامع دیگری، چالش‌ها و مزایای تحلیل کلان‌داده‌های پزشکی با جستجو در سه منبع معتبر در بازه سال‌های ۲۰۱۰ تا ۲۰۱۶ ارایه گردید. طبق گزارش موجود، ساختار داده‌ها، امنیت، استانداردسازی، ذخیره‌سازی و مهارت‌های مدیریت داده از چالش‌برانگیزترین مسایل می‌باشد و افزایش کیفیت ارایه خدمات پزشکی، تصمیم‌گیری بهینه، کاهش هزینه‌ها و تشخیص زودهنگام بیماری‌ها از اصلی‌ترین مزایای کاوش کلان‌داده‌های سلامت به شمار می‌روند (۶). طبق گزارش تحقیقی که در آن مزایا و چالش‌های مطرح در کاوش کلان‌داده‌های پزشکی بررسی گردید، ناخالصی و ساختار فعلی داده‌ها، اصلی‌ترین چالش‌ها و کاهش هزینه‌ها و افزایش کیفیت خدمات، عمده‌ترین دستاوردهای کاوش بود (۷).

آگاهی از دلایل بروز و تأثیرات مشکلات، کمک شایانی به ارایه و انتخاب راه‌حل‌های متناسب می‌نماید، اما پژوهش جامعی در خصوص علت و تأثیر چالش‌های ناشی از ویژگی‌های داده‌های پزشکی بر روی روش‌های داده‌کاوی ارایه نشده است. بنابراین در مطالعه حاضر به بررسی چالش‌هایی پرداخته شد که ناشی از خصوصیات کلان‌داده‌های پزشکی هستند. چالش‌ها، عوامل بروز آن‌ها و تأثیراتی که بر روی روش‌های کاوش می‌گذارند، به صورت ساختار یافته بررسی گردید و در راستای تأکید بر اهمیت و ضرورت رفع این چالش‌ها، مزایای کاوش نیز از دیدگاه‌های مختلف ارایه شد.

روش بررسی

این تحقیق از نوع مروری حوزه‌ای بود که در آن پژوهش‌های ارایه شده به زبان انگلیسی در نشریات و پایگاه‌های IEEE Xplore، PubMed، ScienceDirect، Springer و Google Scholar با استفاده از دو گروه واژه کلیدی، در بازه زمانی سال‌های ۲۰۱۱ تا ۲۰۲۰ مورد جستجو قرار گرفت. جهت اطمینان از بررسی نشریات معتبر، پایگاه داده ResearchGate نیز با همان فیلترها مورد جستجو قرار گرفت. گروه اول مربوط به مزایا و گروه دوم مربوط به چالش‌ها بود. تعداد تحقیقات یافته شده در گروه اول، ۸۸۷ و در گروه دوم، ۱۰۵۲ عدد بودند که در جدول ۱ به تفکیک پایگاه داده‌ها ارایه شده‌اند. پژوهش‌های مشترک بسیار زیادی در پایگاه داده‌ها وجود داشتند و همچنین، مطالعات زیادی نیز در راستای یک هدف ارایه شده بودند. پس از بررسی‌های اولیه، موارد تکراری و مطالعات تک منظوره حذف و فقط تحقیقات مروری جامع انتخاب شدند و تعداد پژوهش‌ها در فهرست نهایی به ۹ عدد در گروه اول و ۲۰ عدد در گروه دوم کاهش یافتند. پس از استخراج چالش‌ها، مجدد هر چالش مورد جستجو قرار گرفت و مطالعات مختص به آن استخراج گردید که در جدول ۲ ارایه شده است. در مرحله دوم جستجو نیز جهت پوشش کامل تحقیقات، پایگاه داده ResearchGate مورد جستجو قرار گرفت.

پس از بررسی پژوهش‌ها، به‌روزترین مطالعه جامعی که هر سه موضوع دلایل بروز چالش، تأثیرات چالش و راهکارهای مواجهه با چالش را مورد بررسی قرار داده بود، در بخش یافته‌ها معرفی گردید. به دلیل پیشرفت بسیار سریع تکنولوژی و در نتیجه، ارایه راه‌حل برای مشکلات داده‌کاوی، بازه زمانی جستجو در مرحله دوم، سال‌های ۲۰۱۵ تا ۲۰۲۱ در نظر گرفته شد. با تحلیل دقیق هر چالش و عوامل و آثیرات آن بر روی داده‌کاوی، چالش‌ها طبقه‌بندی و راهکارهای رایج برای مقابله با هر کدام نیز به صورت خلاصه ارایه گردید. به تمام منابعی که از نتایج آن‌ها استفاده شده بود، ارجاع داده شد و هیچ داده خصوصی در تحقیق مورد بررسی قرار نگرفت. همچنین، داده‌های مورد بررسی افراد و سازمان‌ها قابلیت شناسایی نداشتند و تلاشی نیز در شناسایی آن‌ها انجام نشده است.

یافته‌ها

در این بخش، ابتدا مزایای کاوش کلان‌داده‌های پزشکی از دیدگاه‌های مختلف عنوان شد و سپس چالش‌ها به صورت طبقه‌بندی شده، ارایه گردید.

مزایای کاوش کلان‌داده پزشکی: محققان بر این باور هستند که هزینه کردن درصد اندکی از مبلغ درمان برای پیشگیری از بیماری، می‌تواند باعث کاهش چشمگیر هزینه‌های آتی فرد و جامعه شود (۸). همان‌طور که در شکل ۱ مشاهده می‌شود، بیمار، دولت، بیمه، کادر درمان و در نهایت، کل جامعه از این مزایا بهره می‌برند. بیمار خدمات باکیفیت بیشتر و هزینه کمتری دریافت می‌کند (۱۰، ۹). دولت می‌تواند با پیشگیری به‌موقع از اپیدمی‌ها، هزینه‌ها را کاهش دهد (۱۲، ۱۱). شرکت‌های بیمه‌ای با ارایه طرح‌های مناسب، در هزینه‌های خود صرفه‌جویی می‌کنند. احتمال خطای کادر درمانی کم می‌شود (۱۳) و در نهایت، میزان رضایتمندی و سلامت جامعه و همچنین، درصد امید به زندگی افزایش خواهد یافت (۱۵، ۱۴). همچنین، کاوش کلان‌داده پزشکی می‌تواند منجر به گسترش کاربردهای پزشکی نوین شود (۱۶).

چالش‌های کلان‌داده پزشکی: داده‌های پزشکی، گستره وسیعی از رکوردهای الکترونیک سلامت گرفته تا داده‌های دریافتی از حسگرهای تجهیزات پوشیدنی پزشکی را در برمی‌گیرند. مشکلات رایجی در اغلب این داده‌ها وجود دارد که در پژوهش حاضر بر اساس منشأ بروز به سه دسته تقسیم شده‌اند. دسته اول شامل موارد مرتبط با خود رکورد ذخیره شده است (شکل ۲). در دسته دوم، مشکلات مرتبط با ساختار داده ذخیره شده قرار می‌گیرند که در شکل ۳ چالش‌های این گروه و عوامل، تأثیرات و راهکارهای رایج برای رفع آن‌ها نمایش داده شده است. دسته سوم مشکلات معنایی و ذاتی را در برمی‌گیرد که ناشی از ذات مفاهیم کلان‌داده و کاربردهای داده‌های بیماران در حوزه پزشکی می‌باشد.

جدول ۱: تحقیقات مرتبط با کلید واژه‌های مزایا و چالش‌ها به تفکیک پایگاه داده

گروه	کلید واژه	IEEE Xplore	PubMed	ResearchGate	ScienceDirect	Springer	Google Scholar
مزایا	Medical|healthcare big data + mining + benefits| advantages + survey|review	۱۷۳	۱۶۲	۱۴۱	۱۶۴	۱۱۵	۱۳۲
چالش‌ها	Medical| Healthcare big data + mining + challenges|obstacles| problems + survey| review	۱۷۰	۱۸۲	۱۹۵	۱۲۴	۲۲۴	۱۵۷

جدول ۲: تحقیقات مرتبط با هر یک از چالش‌های به دست آمده به تفکیک پایگاه داده

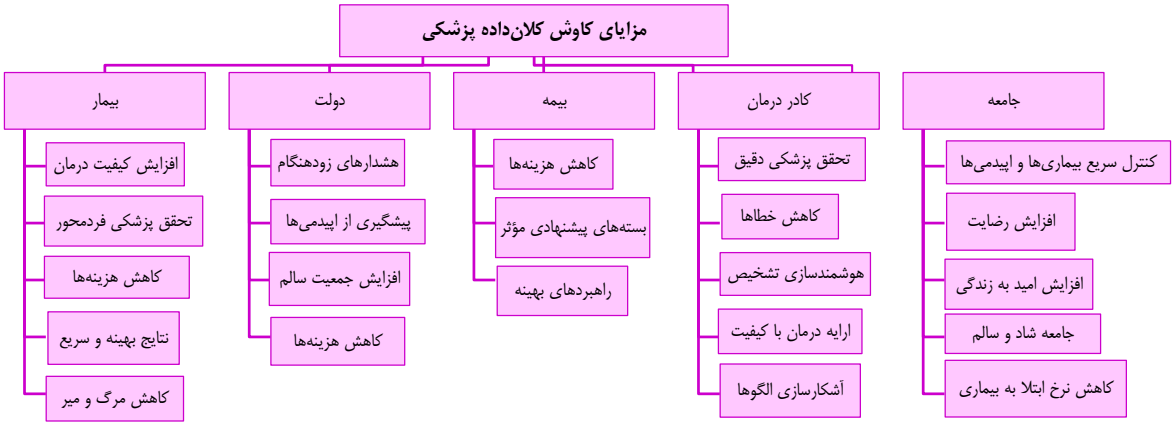
چالش	کلید واژه	IEEE Xplore	PubMed	ResearchGate	ScienceDirect	Springer	Google Scholar
داده ناقص	Missing value + medical|health data	۱۶۵	۹۶	۱۰۸۴	۱۱۵	۱۰۳	۱۹۹
داده خارج از دامنه	Outliers + medical|health data	۸۴	۳۵	۷۶۳	۴۶	۷۵	۶۹
سایر نویزها	Noises + medical|health data	۲۲۵	۷۴	۲۵۸	۱۰۵	۷۸	۹۲
ناهمگنی	Heterogeneity + medical|health data	۵۴	۳۲	۱۳۱	۱۶	۶۶	۷۴
افزونگی	Redundancy|duplication + medical|health data	۱۸۷	۶۴	۶۵۶	۴۷	۵۹	۱۴۰
حجم و رشد زیاد	Volume|velocity + medical|health data	۸۲	۳۵	۸۱۲	۵۶	۶۷	۵۴
ابعاد زیاد	Dimensionality|multi-aspect + medical|health data	۶۰	۴۷	۳۴۷	۴۹	۳۷	۵۵
ناسازگاری	Inconsistency + medical|health data	۴۵	۲۳	۴۳۸	۵۵	۳۰	۷۲
امنیت و محرمانگی	Security|privacy + medical|health data	۱۸۶	۲۷۱	۱۷۰۴	۱۷۴	۹۸	۱۶۳
عدم صحت	Veracity + medical|health data	۳۹	۳۲	۴۷۰	۶۵	۷۴	۴۳
تنوع و عدم ثبات	variability|variety + medical|health data	۸۳	۳۴	۲۷۹	۵۹	۶۰	۸۶

شکل‌های ۴ و ۵، چالش‌های مطرح در این گروه و عوامل، تأثیرات و راهکارهای رایج برای مواجهه با آن‌ها را نشان می‌دهد.

چالش‌های مرتبط با رکوردهای پزشکی: چالش‌های این گروه جزء تأثیرگذارترین مشکلات هستند، اما با روش‌های پاکسازی قابل رفع می‌باشند. هرچند که رفع این مشکلات هزینه‌بر است، اما تأثیر عمیقی که در نتایج کاوش دارند، باعث می‌گردد ارزش هزینه کردن داشته باشند.

چالش‌های مرتبط با ساختار: چالش‌های ساختاری بیشتر هنگام یکپارچه‌سازی منابع داده، ارزیابی چندین منبع داده مختلف و انتقال اطلاعات منبع داده به محیط دیگری بروز پیدا می‌کنند و حل این چالش‌ها نیز به زمان یکپارچه‌سازی و انتقال موکول می‌گردد که هزینه سنگینی برای روش کاوش دارد (۱۷). چنانچه این مشکلات برطرف نشود، سبب بروز افزونگی خواهد شد. همچنین، به دلیل ذخیره داده‌های چند وجهی و چند بعدی به صورت طولی در بیشتر سیستم‌ها، تفکیک و تحلیل چند جانبه این داده‌ها چالش عظیمی ایجاد می‌نماید و در بسیاری از مواقع، از آن‌جایی که ابعاد مختلف شناسایی نمی‌شوند، تحلیل دقیقی نمی‌تواند صورت گیرد (۱۸).

مواجهه با چالش‌های ساختاری کلان‌داده بسیار زمان‌بر و هزینه‌بر است و راهکارهای ارایه شده فعلی نه مبتنی بر یکپارچه‌سازی، بلکه مبتنی بر بسط و توزیع می‌باشند (۱۹).



شکل ۱: مزایای کاوش کلان‌داده‌های پزشکی از دیدگاه‌های مختلف

شکل ۲: دلایل بروز چالش، تأثیرات و راهکارهای مواجهه با چالش‌های مرتبط با رکورد

چالش‌های ذاتی و معنایی: چالش‌های ذاتی و معنایی مشکل‌سازترین چالش‌ها برای داده‌کاوی محسوب می‌شوند. این مشکلات به صورت نامحسوس، تأثیر عمیقی در کاوش می‌گذارند و رفع بسیاری از آن‌ها مشکل و در برخی مواقع غیر ممکن است. هزینه رفع برخی از این چالش‌ها به قدری زیاد است که متخصصان ترجیح می‌دهند روش‌هایی برای کاوش ابداع نمایند که با وجود این چالش‌ها نتایج نسبتاً مطلوبی ارایه نمایند (۲۵).

بحث

مطالعات صورت گرفته شان می‌دهد که چالش‌های کاوش کلان‌داده‌های پزشکی در سه گروه «رکوردی، ساختاری، معنایی و ذاتی» طبقه‌بندی می‌شوند. بیشتر مشکلات رکوردی، هنگام جمع‌آوری و ثبت نمودن اطلاعات ایجاد می‌گردد. تشخیص و پیشگیری مشکلات این گروه آسان است و با هشدار دهنده‌ها و بازدارنده‌ها می‌توان با آن‌ها مقابله کرد و با روش‌های پاکسازی می‌توان آن‌ها را رفع نمود. بیشتر مشکلات ساختاری در اثر یکپارچه‌سازی، ارزیابی و انتقال داده‌ها نمایان می‌گردد و علت آن، فقدان فرمت واحد است. شناسایی این چالش‌ها آسان، اما مواجهه با آن‌ها پرهزینه می‌باشد. راه مقابله با چالش‌های ساختاری، طراحی استانداردسازی ذخیره و نمایش داده‌ها است، اما در حال حاضر به دلیل عدم امکان وجود فرمت استاندارد، جایگزین نمودن یکپارچه‌سازی با بسترهای منعطف، توزیع شده و هوشمند می‌باشد. این راهکارها سعی بر این دارند که پالایش اولیه در هر منبع به صورت مستقل انجام گیرد و سپس منابع آماده شده وارد فاز تحلیل شوند.

چالش‌های گروه سوم که مشکل‌سازترین چالش‌ها هستند، ریشه در ذات داده پزشکی و کلان‌داده دارند و کاوش را به امر زمان‌بر و هزینه‌بری تبدیل می‌کنند. هشدار دهنده‌ها و بازدارنده‌ها در پیشگیری از نویزهای این گروه ناکارامد هستند و نیاز به سیستم هوشمندی است تا آن‌ها را شناسایی نماید. روش حل این نویزها نیز باید هوشمند باشد و با تکنیک‌های یادگیری ماشین شناسایی و رفع گردد. به دلیل سربار زمانی و هزینه بالای شناسایی، این نوع نویزها در بیشتر موارد نادیده گرفته می‌شوند. مواجهه با مشکلات این گروه ساختار، روش‌های خاصی می‌طلبد و نمی‌توان با ساختارهای کلاسیک کلان‌داده را به صورت مطلوب تحلیل کرد و به صورت تفسیرپذیر نمایش داد.

شرح چالش		
عوامل بروز چالش	تأثیرات چالش بر روی داده‌کاوی	راهکارهای رایج برای مواجهه با چالش
ناهمگنی: تا زمانی که از منابع به صورت مستقل استفاده می‌شود، چالشی وجود ندارد. چالش ناهمگنی زمانی آغاز می‌شود که داده‌هایی که از منابع مختلف و ناهمگن، برای اهداف تحلیل یا نمایش یکپارچه‌سازی شوند (۲۶).		
فقدان فرمت و ادبیات استاندارد توزیع داده‌ها در منابع ناهمگن	عدم یا کاهش کارایی الگوریتم کاوش نتایج غیر دقیق و غیر قابل اعتماد کاهش تفسیرپذیری	تطبیق ادبیات داده استانداردسازی تعیین فرمت ارایه بسترهای سخت‌افزاری و نرم‌افزاری که قابلیت تحمل داده‌های ناهمگن را دارند.
افزونگی و تکرار: داده به صورت‌های یکسان یا مختلف در منابع تکرار می‌شود (۲۷).		
وجود نویز توزیع و تکرار داده در منابع مختلف فقدان فرمت و ساختار استاندارد و یکسان ناهمگنی منابع	نتایج غیر دقیق افزایش هزینه زمانی و حافظه مورد نیاز کاوش	روش‌های پردازش متن و تصویر برای داده‌های افزونه غیر مبهم روش‌های هوشمند یادگیری ماشین برای داده‌های افزونه مبهم (داده‌هایی با ظاهر متفاوت ولی ذات یکسان)
تبدیل داده‌های چند بعدی به داده‌های طولی: بیشتر سیستم‌های فعلی وجوه مختلف داده را نادیده می‌گیرند و داده را به صورت طولی در یک رکورد ذخیره می‌نمایند (۲۸).		
نمایش و ثبت داده‌ها بدون تجزیه و تفکیک	نتایج غیر دقیق به علت عدم رعایت توالی و اولویت ابعاد عدم امکان یا مشکل بودن آشکارسازی روابط بین ابعاد	ارایه ساختار و مدل‌های نوین برای نمایش و کاوش بهره‌گیری از الگوریتم‌های یادگیری عمیق برای استخراج ویژگی‌ها
تنوع و عدم ثبات: در کلان‌داده‌ها نه تنها فرمت و ساختار استانداردی برای یک نوع داده وجود ندارد، بلکه این داده‌ها انواع مختلفی از متن تا صوت را شامل می‌شوند و داده‌ها مرتب تغییر می‌کنند (۲۹).		
وجود انواع مختلفی از داده‌ها در حوزه پزشکی مانند داده‌های حسگرها، تصاویر، داده‌های طولی و...	عدم امکان یا مشکل بودن کشف روابط افزایش پیچیدگی الگوریتم افزایش پیچیدگی تفسیر و نمایش نتایج	کاوش داده‌ها با استفاده از ساختارها و روش‌های منعطف

شکل ۳: دلایل بروز چالش، تأثیرات چالش و راهکارهای مواجهه با چالش‌های ساختاری

نویزهای این گروه در بسیاری از مواقع به صورت دستی تعیین تکلیف می‌شوند که با توجه به حجم زیاد داده‌های پزشکی، این امر بسیار مشکل است. برخی از نویزهای این گروه با استفاده از روش‌های هوشمندی که بسیار خوب هستند و با داده بسیار زیاد آموزش دیده‌اند نیز قابل شناسایی نیستند؛ چرا که ممکن است واقعاً اولین باری باشد که چنین مقداری به وجود آمده است. چالش‌های حجم و سرعت رشد بالا، نیازمند روش‌های مقیاس‌پذیر و چالش وجود ابعاد و وجوه زیاد، نیازمند ساختار و الگوریتم‌های مناسب برای کاوش و نمایش است. چالش‌های ناسازگاری، صحت داده، امنیت و محرمانگی دشوارترین چالش‌ها برای رفع می‌باشند. در سال‌های اخیر، الگوریتم‌های یادگیری عمیق موفقیت زیادی در رفع چالش‌های معنایی و ذاتی به دست آورده‌اند. دانش حاصل از کاوش کلان‌داده‌های پزشکی برای بیماران، دولت‌ها، صنعت بیمه، کادر درمان و جوامع به اندازه‌ای سودمند است که این چالش‌ها ارزش هزینه کردن دارند.

تحقیق حاضر بر اساس نتایج مطالعات روی پایگاه داده‌های خارجی ارایه گردید. پژوهش بر روی منابع و پایگاه داده‌های بومی می‌تواند سایر چالش‌های احتمالی موجود در داده‌های پزشکی داخل کشور را آشکار نماید.

نتیجه‌گیری

نتایج کاوش داده‌های پزشکی تأثیر بسزایی در درمان مناسب، پیشگیری به‌موقع، افزایش کیفیت درمان، کاهش هزینه‌ها و کاهش میزان اثرات منفی جسمی، روحی و اجتماعی دارد، اما عدم آمادگی برای ظهور این کلان‌داده‌ها، سرمنشأ چالش‌هایی است که روش‌های کاوش را با مشکل مواجه می‌سازد. خطاهای کاربران، ذات داده‌های پزشکی، عدم وجود فرمت یکسان و عدم بهره‌گیری از ساختارهای مناسب برای ذخیره و نمایش کلان‌داده، از جمله اصلی‌ترین عوامل انحراف الگوریتم‌های کاوش و کیفیت پایین نتایج به شمار می‌روند. چالش‌های حاصل از این عوامل را می‌توان با استانداردسازی، طراحی و استفاده از الگوریتم‌ها و ساختارهای مناسب و هوشمندسازی ثبت داده، پیشگیری و رفع نمود. حجم و سرعت رشد بالای داده با استفاده از روش‌های مقیاس‌پذیر و توزیع شده قابل حل است. تنوع داده‌های پزشکی را می‌توان با روش‌های هوشمند و منعطف مدیریت نمود.

شرح چالش		
عوامل بروز چالش	**تأثیرات چالش بر روی داده‌کاوی**	**راهکارهای رایج برای مواجهه با چالش**
داده غیر معمول: این نوع داده‌ها در نگاه اول نویز به نظر می‌رسند و متأسفانه توسط بسیاری از روش‌های پاکسازی غیر هوشمند از بین می‌روند، اما آن‌ها نویز نیستند و نشان دهنده موارد استثنایی و حاوی اطلاعات ارزشمندی هستند (۳۰).		
وجود موارد نادر و خاص	کاهش صحت کاوش با تشخیص داده غیر معمول به عنوان نویز و حذف آن از چرخه کاوش بر هم زدن روال عادی و تعادل کاوش	همکاری متخصصان حوزه پزشکی و داده برای شناسایی و تعیین تکلیف بهره‌گیری از روش‌های هوشمند.
ناسازگاری: ناسازگاری جزء مواردی است که شناسایی آن مشکل و حل آن مشکل‌تر است. با وجود پیشرفت چشمگیر الگوریتم‌های یادگیری ماشین، این مورد نیز همچون داده‌های غیر معمول، هنوز هم نیاز به نظارت و تأیید کاربر دارند (۲۶).		
خطاهای عمدی یا غیر عمدی افزونگی عدم شناسایی ارتباطات بین داده‌ها تفسیر نادرست از داده ورودی فقدان یا تعداد بالای محدودیت در ورود اطلاعات	نتایج غیر دقیق یا نادرست انحراف مسیر آموزش در تکنیک‌های یادگیری ماشین	بهره‌گیری از روش‌های هوشمند رفع ناسازگاری به صورت دستی (در حال حاضر هیچ روش ۱۰۰ درصد مطمئنی برای مواجهه با داده‌های ناسازگار وجود ندارد و تمام روش‌ها موجود برخی داده‌ها را قربانی برخی دیگر می‌نمایند).
امنیت و محرمانگی: این چالش به معنی ویرایش و یا استفاده داده توسط افراد غیر مجاز است (۳۱).		
توزیع یا به اشتراک‌گذاری داده	کاهش دقت و کارایی کاوش به دلیل محدودیت‌ها و موانع امنیتی در دسترسی به داده	جایگزین نمودن روش‌های توزیع‌پذیر با یکپارچه‌سازی منابع استفاده از الگوریتم‌های رمزگذاری و رمزگشایی است. سطح بندی کاربران و پنهان نمودن برخی از اطلاعات از برخی سطوح
قابلیت اعتماد به داده: رقابت گسترده بین سازمان‌ها و دولت‌ها در استخراج دانش از داده، عدم اعتماد به برخی داده‌ها از جمله داده‌های پزشکی را افزایش می‌دهد. این ویژگی از دو جهت «یکی خود داده و دیگری منبع تولید داده» قابل بررسی است (۳۲).		
وجود نویزهای عمدی یا غیر عمدی وجود خطا در جمع‌آوری و یکپارچه‌سازی داده تفسیر نادرست داده جمع‌آوری شده هنگام ثبت	عدم یا کاهش اطمینان به نتایج تصمیم‌گیری نادرست به دلیل نتایج غیر دقیق	بهره‌گیری از روش‌های هوشمندی که بتواند داده صحیح و غیر واقعی را تشخیص دهد. فعلاً راهکار جامعی برای آن وجود ندارد.

شکل ۴: دلایل بروز چالش، تأثیرات چالش و راهکارهای مواجهه با چالش‌های معنایی

صحت، امنیت و محرمانگی، چالش‌برانگیزترین خصوصیات کلان‌داده‌های پزشکی می‌باشد که راهکار بهینه‌ای برای مواجهه با آن‌ها وجود ندارد. در حال حاضر، تکنیک‌های یادگیری ماشین و ساختارها و الگوریتم‌های مقیاس‌پذیر، موفق‌ترین راهکارها را برای مواجهه با بیشتر چالش‌های کلان داده ارایه می‌دهند. در مطالعه حاضر، به علت محدودیت دسترسی و جستجوی بسیاری از منابع بومی، نتایج تحقیقات غیر بومی مورد بررسی قرار گرفت که سبب عدم آگاهی از چالش‌های احتمالی داده‌های بومی می‌شود.

پیشنهادها

طراحی فرمت واحد و ساختار مناسب برای ثبت و نمایش داده‌های پزشکی، یکی از ضروری‌ترین نیازهای این حوزه است. همچنین، طراحی و پیاده‌سازی بسترها و روش‌های مقیاس‌پذیر و امنی که قابلیت مدیریت داده‌های حجیم، چند بعدی، متنوع و پیچیده‌ای همچون داده‌های پزشکی را داشته باشد، اقدام مؤثری در دستیابی به نتایج کاوش قابل اطمینان و دقیق می‌باشد.

تشکر و قدردانی

بدین وسیله از کلیه افرادی که در انجام پژوهش حاضر همکاری نمودند، تشکر و قدردانی به عمل می‌آید.

تضاد منافع

در انجام پژوهش حاضر، نویسندگان هیچ‌گونه تضاد منافعی نداشته‌اند.

شکل ۵: دلایل بروز چالش، تأثیرات چالش و راهکارهای مواجهه با چالش‌های ذاتی

References

1. Raghupathi W, Raghupathi V. Big data analytics in healthcare: Promise and potential. Health Inf Sci Syst 2014; 2: 3.
2. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. Yearb Med Inform 2014; 9: 97-104.
3. Oussous A, Benjelloun FZ, Ait Lahcen A, Belfkih S. Big data technologies: A survey. Journal of King Saud University - Computer and Information Sciences 2018; 30(4): 431-48.
4. Islam MS, Hasan MM, Wang X, Germack HD, Noor-E-Alam. A systematic review on healthcare analytics: application and theoretical perspective of data mining. Healthcare (Basel) 2018; 6(2).
5. Pashazadeh A, Navimipour NJ. Big data handling mechanisms in the healthcare applications: A comprehensive and systematic literature review. J Biomed Inform 2018; 82: 47-62.
6. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: A systematic review. JMIR Med Inform 2016; 4(4): e38.
7. Lee CH, Yoon HJ. Medical big data: Promise and challenges. Kidney Res Clin Pract 2017; 36(1): 3-11.
8. Jothi N, Rashid NA, Husain W. Data Mining in Healthcare: A Review. Procedia Comput Sci 2015; 72: 306-13.
9. Divaris K. Fundamentals of precision medicine. Compend Contin Educ Dent 2017; 38(8 Suppl): 30-2.
10. Chawla NV, Davis DA. Bringing big data to personalized healthcare: A patient-centered framework. J Gen Intern Med 2013; 28(Suppl 3): S660-S665.
11. Huang L, Shea AL, Qian H, Masurkar A, Deng H, Liu D. Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. J Biomed

Inform 2019; 99: 103291.
12. Prosperi M, Min JS, Bian J, Modave F. Big data hurdles in precision medicine and precision public health. BMC Med Inform Decis Mak 2018; 18(1): 139.
13. Viceconti M, Hunter P, Hose R. Big data, big knowledge: Big data for personalized healthcare. IEEE J Biomed Health Inform 2015; 19(4): 1209-15.
14. Luo J, Wu M, Gopukumar D, Zhao Y. Big data application in biomedical research and health care: A literature review. Biomed Inform Insights 2016; 8: 1-10.
15. Sun J, Reddy CK. Big data analytics for healthcare. Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining; 2013 Aug 11-14; Chicago, IL, USA.
16. Huang BE, Mulyasasmita W, Rajagopal G. The path from big data to precision medicine. Expert Rev. Precis Med Drug Dev 2016; 1(2): 129-43.
17. Installe AJ, Van den Bosch T, De Moor B, Timmerman D. Clinical data miner: An electronic case report form system with integrated data preprocessing and machine-learning libraries supporting clinical diagnostic model research. JMIR Med Inform 2014; 2(2): e28.
18. Papalexakis EE, Faloutsos C. Unsupervised tensor mining for big data practitioners. Big Data 2016; 4(3): 179-91.
19. Jeon I, Papalexakis EE, Faloutsos C, Sael L, Kang U. Mining billion-scale tensors: Algorithms and discoveries. The VLDB Journal 2016; 25(4): 519-44.
20. Jain S, Jain K, Chodhary N. A survey paper on missing data in data mining. Int J Innov Eng Res Technol 2016; 3(12): 45-50.
21. Bansal R, Gaur N, Singh SN. Outlier Detection: Applications and techniques in Data Mining. Proceedings of the 6th International Conference - Cloud System and Big Data Engineering; 2016 Jan 14-15; Noida, India. p. 373-7.
22. Garc+¡a S, Luengo J, Herrera F. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowl Based Syst 2016; 98: 1-29.
23. Zhou PY, Wong AKC. Explanation and prediction of clinical data with imbalanced class distribution based on pattern discovery and disentanglement. BMC Med Inform Decis Mak 2021; 21(1): 16.
24. Idri A, Benhar H, Fernandez-Aleman JL, Kadi I. A systematic map of medical data preprocessing in knowledge discovery. Comput Methods Programs Biomed 2018; 162: 69-85.
25. Benhar H, Idri A, Fernandez-Aleman JL. Data preprocessing for heart disease classification: A systematic literature review. Comput Methods Programs Biomed 2020; 195: 105635.
26. García S, Ramírez -Gallego S, Luengo J, Benítez JM, Herrera F. Big data preprocessing: Methods and prospects. Big Data Analytics 2016; 1(1): 9.
27. Giordani P, Kiers HAL. A review of tensor-based methods and their application to hospital care data. Stat Med 2018; 37(1): 137-56.
28. Lin JH, Haug PJ. Data preparation framework for preprocessing clinical data in data mining. AMIA Annu Symp Proc 2006; 2006: 489-93.
29. Atzmueller M, Schmidt A, Hollender M. Data preparation for big data analytics: Methods and experiences. In: Atzmueller M, Oussena S, Roth-Berghofer T, editors. Enterprise big data engineering, analytics, and management. Hershey, PA: GI Global; 2016. p. 157-70.
30. Ortega Jn, Iturbide E, Olivares Peregrino V, Hidalgo M, Almanza N, Martinez-Rebollar A. A data preparation methodology in data mining applied to mortality population databases. J Med Syst 2015; 39: 152.
31. Rashid A, Mohd Yasin N. Generalization technique for privacy preserving of medical information. Int J Eng Technol 2014; 6(4): 262-4.
32. Gkoulalas-Divanis A, Loukides G, Sun J. Publishing data from electronic health records while preserving privacy: A survey of algorithms. J Biomed Inform 2014; 50: 4-19.
33. Rajinder Sandhu, Navroop Kaur, Sandeep K. Sood, and Rajkumar Buyya. 2018. TDRM: Tensor-based data representation and mining for healthcare data in cloud computing environments. J Supercomput 2018; 74(2): 592-614.
34. Henderson J, He H, Malin BA, Denny JC, Kho AN, Ghosh J, et al. Phenotyping through Semi-Supervised Tensor Factorization (PSST). AMIA Annu Symp Proc 2018; 2018: 564-73.
35. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018; 15(141): 20170387.
36. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: A systematic review. Int J Med Inform 2018; 114: 57-65.

Health Information Management

Published by ***Vesnu Publications***

Vol 18/ No 5/ December 2021

DOI: 10.22122/him.v18i5.4367

Advantages and Challenges of Medical Big Data Mining

Leila Baradaran-Sorkhabi[1], Farhad Soleimanian-Gharehchopogh[2], Jafar Shahmfar[3]

Original Article

Abstract

Introduction: Data mining seems to be a good tool for showing underlying knowledge of Medical Big Data (MBD). Understanding characteristics of data and possible challenges are the first steps of the journey. This study endeavors to inspect reasons, effects, and solutions of challenges as well as benefits of MBD mining.

Methods: In so doing, PubMed, ScienceDirect, Springer, and Google Scholar databases were scrutinized using two groups of keywords for benefits and challenges in the years 2011-202. The search language was English. Single-purpose studies were excluded and those studies that were focused on MBD mining were included. Then, challenge was examined separately and the results were categorized.

Results: Extracted knowledge from MBD enhances quality of care. However, low-quality performance in gathering and storing the data, properties of big data, and inherent structure of medical data cause many problems for mining methods. Inconsistency, veracity, privacy, and security issues are the major challenging problems. Standardization and enhancing quality of data gathering, storing, and representing tasks are the effective problem prevention strategies. Designing and using appropriate frameworks, algorithms, and structures as well as utilizing machine learning and artificial intelligence techniques are the most effective solutions for dealing with the challenges.

Conclusion: MBD was appeared and expanded when the world was not ready for it. Thus, it caused many challenges for mining methods. Some of them are traceable, preventable, and manageable. However, some challenges need novel and intelligent methods that are able to handle MBD.

Keywords: Data Mining; Big Data; Health

Received: 19 July, 2021 Accepted: 05 Dec., 2021 Published: 06 Dec., 2021

Citation: Baradaran-Sorkhabi L, Soleimanian-Gharehchopogh F, Shahmfar J. **Advantages and Challenges of Medical Big Data Mining.** Health Inf Manage 2021; 18(5): 225-33.

Article resulted from PhD thesis No. 10341006971002 funded by Urmia Branch, Islamic Azad University.

1- PhD Student, Software Engineering, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran

2- Assistant Professor, Software Engineering, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran

3- Assistant Professor, Software Engineering, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia AND Department of Community Medicine, Tabriz University of Medical Sciences, Tabriz, Iran

Address for correspondence: Farhad Soleimanian-Gharehchopogh; Assistant Professor, Software Engineering, Department of Computer Engineering, Urmia Branch, Islamic Azad University, Urmia, Iran; Email: bonab.farhad@gmail.com