



Review Article

Applications of data mining in healthcare with a focus on disease prediction and diagnosis: A systematic review

* Mohammad Hossein Ronaghi¹ , Ahmadreza Malekpour¹

1. Department of Management, School of Economics Management and Social Sciences, Shiraz University, Shiraz, Iran.

Use your device to scan
and read the article online



Citation: Ronaghi MH, Malekpour A. Applications of data mining in healthcare with a focus on disease prediction and diagnosis: A systematic review. *Journal of Modern Medical Information Science*. 2025; 11(2):233-250. [In Persian]



10.48312/jmis.11.2.998.1

Article Info:

Received: 29 July 2025

Accepted: 4 Sep 2025

Available Online: 20 Jun 2025

ABSTRACT

Introduction: This study aimed to systematically review the applications of data mining in disease prediction and diagnosis within the healthcare domain.

Information sources or data: A systematic search was conducted in PubMed, Scopus, ScienceDirect, and Magiran databases for studies published between 2010 and 2025.

Selection methods for study: This study was conducted based on the PRISMA 2020 guidelines. After removing duplicate studies, the titles, abstracts, and full texts of articles were screened according to the inclusion and exclusion criteria, and finally, 44 studies were included in the analysis. Study quality was assessed using the Joanna Briggs Institute checklist, and data synthesis was performed using a qualitative narrative approach.

Combine content and results: The findings showed that data mining was most frequently applied to the prediction and diagnosis of cardiovascular diseases, diabetes, kidney diseases, cancers, and other chronic conditions. The data used mainly consisted of structured clinical data and electronic health records. Commonly used algorithms included decision trees, Naïve Bayes, support vector machines, neural networks, and random forests. More recent studies, in addition to predictive accuracy, focused on model interpretability, feature selection, and the clinical applicability of results using methods such as SHAP and LIME. Furthermore, no single algorithm demonstrated absolute superiority across all problems, and model performance depended on disease type, data quality, and preprocessing methods.

Conclusion: The results of this systematic review showed that data mining can play an effective role in early disease diagnosis and clinical decision-making support. However, challenges such as data heterogeneity, information quality, and limitations in clinical implementation still remain. The development of data infrastructures, standardization of health information, and greater attention to model interpretability are essential for the effective use of data mining in healthcare systems.

Key Words:

Data Mining, Data Analysis, Healthcare, Disease Prediction.

* Corresponding Author:

Dr Mohammad Hossein Ronaghi

Address: Shiraz University, Shiraz, Iran.

E-mail: mh_ronaghi@shirazu.ac.ir



Copyright © 2025 The Author[s];

This is an open access article distributed under the terms of the Creative Commons Attribution License [CC-BY-NC: <https://creativecommons.org/licenses/by-nc/4.0/legalcode>], which permits use, distribution, and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.



Extended Abstract

Introduction:

The rapid growth of electronic health records [EHRs] and clinical databases has generated large volumes of healthcare data with significant potential for improving disease prediction and diagnosis. However, the complexity and heterogeneity of these data limit the effectiveness of traditional analytical methods. Data mining, as a major branch of artificial intelligence and machine learning, provides techniques for extracting meaningful patterns and predictive knowledge from large-scale healthcare datasets.

In recent years, data mining methods have been widely applied to support early disease detection, clinical decision-making, and risk assessment across various medical domains. Applications have been reported for cardiovascular diseases, diabetes, chronic kidney disease, cancers, and other chronic conditions, demonstrating the growing role of these techniques in healthcare. Despite this expansion, the existing literature is highly diverse in terms of disease focus, data sources, algorithms, and evaluation approaches, which makes it difficult to draw general conclusions or identify dominant methodological trends.

Most previous review studies have addressed specific diseases or individual algorithms, rather than providing a comprehensive overview of data mining applications for disease prediction and diagnosis using clinical and EHR-based data. Therefore, the aim of this study was to systematically review published research on the use of data mining techniques in healthcare, with a particular focus on disease prediction and diagnosis. This review aims to summarize current applications, highlight methodological developments, and identify research gaps to guide future investigations in this field.

Methods:

This study was conducted as a systematic review in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses [PRISMA] guidelines. A comprehensive literature search was performed in four electronic databases, including PubMed, Scopus, Science Direct, and Magiran, to identify relevant studies published between 2010 and March 2025. The search strategy

was developed using a combination of controlled vocabulary terms and free-text keywords related to data mining, machine learning, disease prediction, disease diagnosis, clinical data, and electronic health records. Boolean operators were applied to appropriately combine search terms.

Original research articles were eligible for inclusion if they applied data mining or machine learning techniques to predict or diagnose diseases using clinical data, electronic health records, demographic information, or patient registries. Studies focusing exclusively on medical imaging, genomic data, signal processing, or non-clinical settings were excluded. Only studies published in English or Persian with accessible full texts were considered.

After removing duplicate records, titles and abstracts were independently screened by two reviewers based on predefined inclusion and exclusion criteria. Full-text screening was subsequently conducted to determine final eligibility. Any disagreements between reviewers were resolved through discussion or consultation with a third reviewer. Data extraction was performed using a standardized checklist capturing study characteristics, data sources, disease domains, algorithms used, and performance evaluation metrics. Study quality was assessed using the Joanna Briggs Institute critical appraisal checklist. Due to heterogeneity across studies, findings were synthesized using a qualitative narrative approach rather than meta-analysis.

Results:

A total of 1,264 records were initially identified through database searching. After removing duplicates and completing the screening process, 44 studies met the inclusion criteria and were included in the final analysis. The number of published studies increased notably over time, with a marked rise observed after 2020, reflecting growing interest in the application of data mining and machine learning in healthcare.

The findings indicated that data mining techniques were most frequently applied to the prediction and diagnosis of cardiovascular diseases, followed by diabetes, chronic kidney disease, cancers, and other chronic conditions. Most studies relied on structured clinical data and electronic health records, while fewer studies incorporated demographic or



population-based health survey data. Sample sizes varied substantially, ranging from several hundred patients to large-scale datasets involving hundreds of thousands or millions of records.

Commonly used data mining and machine learning algorithms included decision trees, Naïve Bayes, support vector machines, artificial neural networks, random forests, and logistic regression. More recent studies increasingly adopted advanced ensemble and gradient boosting methods, such as XGBoost and LightGBM. From 2021 onward, several studies emphasized model interpretability and clinical applicability, employing explainable artificial intelligence techniques, including SHAP and LIME, to identify influential predictors and enhance transparency.

Across the reviewed studies, no single algorithm consistently outperformed others across all disease domains. Model performance was strongly influenced by disease type, data quality, feature selection, and preprocessing strategies. Overall, most studies reported moderate to high predictive performance, supporting the potential value of data mining techniques for disease prediction and diagnosis in clinical settings.

Conclusion:

The findings of this systematic review demonstrate the growing importance of data mining and machine learning techniques in disease prediction and diagnosis within healthcare settings. The strong focus on cardiovascular and metabolic diseases reflects both their high disease burden and the availability of structured clinical data suitable for predictive modeling. Traditional machine learning algorithms, such as decision trees, random forests, support vector machines, and Naïve Bayes classifiers, continue to play a central role due to their robustness and relatively straightforward implementation.

In recent years, there has been a clear shift from solely accuracy-oriented models toward approaches that emphasize interpretability and clinical usability. The increasing use of explainable machine learning methods, including SHAP and LIME, highlights the need for transparency to support clinician trust and facilitate real-world adoption. Despite promising predictive performance, several challenges remain,

particularly related to data quality, heterogeneity of data sources, and limited external validation across diverse healthcare environments.

From a methodological perspective, the lack of a universally superior algorithm suggests that model selection should be context-dependent, considering disease characteristics, data quality, and clinical objectives. Effective data preprocessing, feature selection, and interdisciplinary collaboration are critical for improving model performance and applicability.

In conclusion, data mining techniques offer substantial potential to enhance early disease detection, support clinical decision-making, and optimize healthcare resource allocation. To fully realize these benefits, future research should prioritize standardized data infrastructures, rigorous validation, and integration of interpretable models into clinical workflows. Addressing these challenges will be essential for translating data mining research into meaningful improvements in healthcare practice.



مقاله مروری

کاربردهای داده کاوی در حوزه سلامت با تمرکز بر پیش بینی و تشخیص بیماری‌ها: مرور نظام مند

* محمدحسین رونقی^۱ ID، احمدرضا ملک پور^۱ ID

۱. گروه مدیریت، دانشکده اقتصاد مدیریت و علوم اجتماعی، دانشگاه شیراز، شیراز، ایران.



Citation: Ronaghi MH, Malekpour A. Applications of data mining in healthcare with a focus on disease prediction and diagnosis: A systematic review. *Journal of Modern Medical Information Science*. 2025; 11(2):233-250. [In Persian]

doi 10.48312/jmis.11.2.998.1

چکیده

اطلاعات مقاله :

تاریخ دریافت: ۸ مرداد ۱۴۰۴

تاریخ پذیرش: ۱۴ شهریور ۱۴۰۴

تاریخ انتشار: ۳۰ شهریور ۱۴۰۴

هدف: این مطالعه باهدف مرور نظام مند کاربردهای داده کاوی در پیش بینی و تشخیص بیماری‌ها در حوزه سلامت انجام شد.

منابع اطلاعات یا داده‌ها: این مطالعه براساس دستورالعمل پریزما ۲۰۲۰ انجام شد. پس از حذف مطالعات تکراری، عناوین، چکیده‌ها و متون کامل مقالات براساس معیارهای ورود و خروج بررسی شدند و در نهایت ۴۴ مطالعه وارد تحلیل شدند. کیفیت مطالعات با استفاده از چک لیست مؤسسه جو آنا بری گز ارزیابی شد و سنتز داده‌ها به صورت کیفی و روایت محور انجام گرفت.

روش‌های انتخاب برای مطالعه: استخراج داده‌ها در پنج محور اصلی نوع الگوریتم‌های پردازش تصویر، ارزیابی بالینی و حضور پزشک در تیم تحقیق و نیازهای فنی و اقتصادی سیستم‌ها متمرکز بود.

ترکیب مطالب و نتایج: یافته‌ها نشان داد که بیشترین کاربرد داده کاوی در پیش بینی و تشخیص بیماری‌های قلبی و عروقی، دیابت، بیماری‌های کلیوی، سرطان‌ها و سایر بیماری‌های مزمن بوده است. داده‌های مورد استفاده عمدتاً شامل داده‌های بالینی ساخت یافته و پرونده‌های الکترونیکی سلامت بودند. الگوریتم‌های پر کاربرد شامل درخت تصمیم، نیوی بیز، ماشین بردار پشتیبان، شبکه‌های عصبی و جنگل تصادفی بودند. مطالعات جدیدتر، علاوه بر دقت پیش بینی، بر تفسیرپذیری مدل‌ها، انتخاب ویژگی و کاربرد بالینی نتایج با استفاده از روش‌هایی مانند شاپ و لایم تمرکز داشته‌اند. همچنین هیچ الگوریتمی در تمامی مسائل برتری مطلق نشان نداد و عملکرد مدل‌ها به نوع بیماری، کیفیت داده‌ها و روش‌های پیش پردازش وابسته بود.

نتیجه گیری: نتایج این مرور نظام مند نشان داد که داده کاوی می‌تواند نقش مؤثری در تشخیص زودهنگام بیماری‌ها و پشتیبانی از تصمیم گیری بالینی ایفا کند. با این حال، چالش‌هایی مانند ناهمگونی داده‌ها، کیفیت اطلاعات و محدودیت‌های پیاده سازی بالینی همچنان وجود دارد. توسعه زیرساخت‌های داده، استاندارد سازی اطلاعات سلامت و توجه بیشتر به تفسیرپذیری مدل‌ها برای استفاده مؤثر از داده کاوی در نظام سلامت ضروری است.

کلیدواژه‌ها :

داده کاوی، تحلیل داده، سلامت، پیش بینی بیماری.

*نویسنده مسئول:

دکتر محمدحسین رونقی

نشانی: دانشگاه شیراز، شیراز، ایران.

پست الکترونیک: mh_ronghi@shirazu.ac.ir



Copyright © 2025 The Author[s].

This is an open access article distributed under the terms of the Creative Commons Attribution License [CC-BY-NC; <https://creativecommons.org/licenses/by-nc/4.0/legal-code.en>], which permits use, distribution, and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

مقدمه:

خطاها و تأثیرات ذهنی انسانی نیز می‌انجامد [۶].

با وجود افزایش مطالعات کاربردی در زمینه داده‌کاوی سلامت، مرورهای نظام‌مند پیشین عمدتاً بر یک بیماری خاص، یک الگوریتم مشخص، یا یک نوع داده محدود تمرکز داشته‌اند و کمتر به بررسی جامع کاربردهای داده‌کاوی در پیش‌بینی و تشخیص بیماری‌ها براساس داده‌های بالینی و پرونده‌های الکترونیکی سلامت پرداخته‌اند. تا جایی که نویسندگان این مطالعه آگاه هستند، مرور نظام‌مندی که به‌طور هم‌زمان دامنه کاربردها، انواع داده‌های بالینی و رویکردهای داده‌کاوی مورد استفاده در پیش‌بینی و تشخیص بیماری‌ها را بررسی کرده باشد، به‌صورت جامع گزارش نشده است. از این‌رو، انجام یک مرور نظام‌مند با این رویکرد می‌تواند به رفع این خلأ پژوهشی کمک کند.

در این مطالعه، منظور از داده‌کاوی در حوزه سلامت، به‌کارگیری تکنیک‌های تحلیلی و الگوریتم‌های یادگیری ماشین برای استخراج الگوها و پیش‌بینی نتایج بالینی از داده‌های ساخت‌یافته و نیمه ساخت‌یافته سلامت است. دامنه این مرور نظام‌مند عمدتاً شامل داده‌های برگرفته از پرونده‌های الکترونیکی سلامت، داده‌های بالینی بیماران، داده‌های جمعیت شناختی و ثبت بیماری‌ها است که باهدف پیش‌بینی و تشخیص بیماری‌ها و پشتیبانی از تصمیم‌گیری بالینی مورد استفاده قرار گرفته‌اند.

هدف این مطالعه ارائه یک مرور نظام‌مند از پژوهش‌های منتشرشده در زمینه کاربردهای داده‌کاوی در پیش‌بینی و تشخیص بیماری‌ها در حوزه سلامت است. این مرور، با تمرکز بر مطالعاتی که از داده‌های پرونده‌های الکترونیکی سلامت، داده‌های بالینی و داده‌های ثبتی بیماران استفاده کرده‌اند، به شناسایی حوزه‌های کاربرد و روندهای پژوهشی غالب در این زمینه می‌پردازد. این مطالعه با پیروی از دستورالعمل پریزما^۱ ۷ تلاش می‌کند تصویری منسجم از نقش تکنیک‌های داده‌کاوی در پشتیبانی از تصمیم‌گیری بالینی ارائه داده و خلأهای پژوهشی موجود را برای تحقیقات آینده برجسته سازد.

با گسترش استفاده از پرونده‌های الکترونیکی سلامت، میزان داده‌های تولیدشده در نظام‌های بهداشتی به‌طور قابل‌توجهی افزایش یافته است، این اطلاعات می‌توانند منبعی ارزشمند برای پزشکان باشند، اما به دلیل پیچیدگی ساختاری و تنوع بالای آن‌ها، استخراج دانش کاربردی نیازمند بهره‌گیری از روش‌های تحلیلی پیشرفته است [۱]. ابزارهای داده‌کاوی قادرند بخشی از محدودیت‌های انسانی، ازجمله خطاهای ناشی از پردازش ذهنی یا خستگی را جبران کرده و در بهبود فرآیند تصمیم‌گیری نقش مؤثری ایفا کنند [۲]. داده‌کاوی راهکاری کارآمد برای استخراج دانش از داده‌ها محسوب می‌شود، این فناوری قادر است به‌طور مؤثر در پیش‌بینی و تشخیص سریع و کم‌هزینه بیماری‌ها مورد استفاده قرار گیرد [۳].

داده‌کاوی به‌عنوان یکی از زیرشاخه‌های هوش مصنوعی، نقش مهمی در استخراج اطلاعات ارزشمند از داده‌های پیچیده و حجیم در حوزه سلامت ایفا می‌کند، این فناوری با استفاده از الگوهای پنهان و روابط ناشناخته، امکان پیش‌بینی بیماری‌ها، بهبود تشخیص‌های بالینی و مدیریت منابع بهداشتی را فراهم می‌آورد [۴]. با توجه به افزایش چشم‌گیر حجم داده‌های پزشکی، سازمان‌های سلامت نیازمند ابزارهایی هستند که بتوانند این انبوه اطلاعات را به دانش قابل‌استفاده تبدیل کنند، در این میان، داده‌کاوی با توانایی خود در پردازش خودکار و کشف روابط پیچیده میان داده‌ها، راهکاری برای غلبه بر محدودیت‌های روش‌های سنتی فراهم می‌کند؛ این فناوری با تحلیل دقیق داده‌ها، دستیابی به نتایجی را ممکن می‌سازد که با روش‌های معمول به‌دشواری حاصل می‌شوند و به همین دلیل، نقشی کلیدی در ارتقای دقت تصمیم‌گیری‌های بالینی ایفا می‌کند [۵]. مطالعات متعددی در حوزه سلامت نشان داده‌اند که تکنیک‌های داده‌کاوی می‌توانند به پیش‌بینی بروز بیماری‌ها و تشخیص برخی بیماری‌ها و مدیریت بالینی بیماران کمک کنند، بنابراین، استفاده از داده‌کاوی در سازمان‌های بهداشتی نه تنها به بهبود بهره‌وری در ارائه خدمات، بلکه به ارتقای کیفیت تصمیم‌گیری و کاهش

مواد و روش‌ها:

پژوهش این مطالعه یک مرور نظام‌مند بر روی کاربردهای داده‌کاوی در حوزه سلامت بود که براساس دستورالعمل‌های گزارش‌دهی ترجیحی برای مرورهای نظام‌مند و روش‌های متا‌پریزما ۷ انجام شد. برای انجام این مرور، یک جستجوی نظام‌مند در پایگاه‌های اطلاعاتی PubMed، Scopus و Science Direct و Magiran ۲۰ مارس ۲۰۲۵ انجام گرفته است. محدوده زمانی جستجو ۲۰۱۰ تا ۲۰۲۵ در نظر گرفته شد. استراتژی جستجو براساس ترکیبی از کلیدواژه‌ها و اصطلاحات کنترل‌شده طراحی شد و مفاهیم اصلی در سه خوشه «داده‌کاوی»، «پیش‌بینی و تشخیص بیماری» و «داده‌های بالینی سلامت» سازمان‌دهی گردید. در پایگاه پاب‌مد، علاوه بر کلیدواژه‌های آزاد، از اصطلاحات MeSH مرتبط نیز استفاده شد و واژگان با بهره‌گیری از عملگرهای منطقی AND و OR ترکیب شدند. استراتژی جستجوی پاب‌مد به صورت زیر بود:

"Data Mining"[MeSH] OR "Machine Learning"[Title/Abstract] OR "Data Mining"[Title/Abstract]] AND ["Disease [Prediction" [Title/Abstract] OR "Disease Diagnosis" [Title/Abstract]] AND ["Electronic Health Records" [MeSH] OR "Clinical Data" [Title/Abstract] OR "Patient Data" [Title/Abstract]]

در سایر پایگاه‌های انگلیسی‌زبان، استراتژی‌های جستجو با استفاده از معادل‌های اصطلاحی این کلیدواژه‌ها و با ساختار مشابه اعمال شد. برای جستجو در پایگاه‌های فارسی‌زبان از واژه‌های «داده‌کاوی»، «تحلیل داده»، «سلامت» و «پیش‌بینی بیماری» استفاده گردید.

علاوه بر جستجوی پایگاه‌های اطلاعاتی، به منظور کاهش سوگیری انتشار، جستجوی منابع خاکستری نیز انجام شد. در این راستا، مقالات کنفرانسی مرتبط، پایان‌نامه‌ها و گزارش‌های پژوهشی در دسترس از طریق پایگاه‌های علمی داخلی و بین‌المللی و موتور جستجوی گوگل اسکالر مورد بررسی قرار گرفتند. همچنین، فهرست منابع مقالات وارد شده به مرور به صورت دستی بررسی شد تا مطالعات

مرتبط احتمالی که در جستجوی اولیه شناسایی نشده بودند، شناسایی و در صورت احراز معیارهای ورود، به مرور افزوده شوند.

مطالعاتی وارد این مرور نظام‌مند شدند که واجد شرایط زیر بودند:

۱. مقالات پژوهشی اصیل منتشر شده در مجلات علمی که به بررسی کاربردهای داده‌کاوی در حوزه سلامت پرداخته بودند.

۲. مطالعاتی که از روش‌های داده‌کاوی، شامل الگوریتم‌های کلاسیک داده‌کاوی (مانند درخت تصمیم، خوشه‌بندی و قوانین انجمنی) و یا روش‌های یادگیری ماشین برای پیش‌بینی یا تشخیص بیماری‌ها براساس داده‌های سلامت استفاده کرده بودند.

۳. مطالعاتی که داده‌های مورد استفاده آن‌ها شامل داده‌های بالینی بیماران، پرونده‌های الکترونیکی سلامت، داده‌های جمعیت‌شناختی یا داده‌های ثبتی بیماری‌ها بود.

۴. مقالاتی که به طور مشخص بر پیش‌بینی یا تشخیص بیماری‌ها در محیط‌های بالینی تمرکز داشتند.

۵. مقالات منتشرشده به زبان فارسی و انگلیسی در بازه زمانی ۲۰۱۰ تا ۲۰۲۵ که متن کامل آن‌ها در دسترس بود.

مطالعات مروری پیشین، از جمله مرورهای نظام‌مند و مرورهای روایتی، به عنوان منبع داده اصلی وارد تحلیل نشدند و صرفاً برای زمینه‌سازی پژوهش و شناسایی شکاف‌های مطالعاتی مورد استفاده قرار گرفتند.

مطالعاتی که دارای یک یا چند مورد از شرایط زیر بودند، از این مرور نظام‌مند حذف شدند:

۱. مقالاتی که تمرکز اصلی آن‌ها خارج از حوزه سلامت یا فاقد ارتباط مستقیم با پیش‌بینی یا تشخیص بیماری‌ها بودند.

۲. مطالعاتی که صرفاً به تحلیل داده‌های تصویربرداری

ثبتی)، اندازه دیتاست، نوع الگوریتم داده کاوی یا یادگیری ماشین، هدف مدل، حوزه بیماری مورد مطالعه و معیارهای ارزیابی عملکرد مدل بود.

سنتز یافته‌ها به صورت کیفی و با استفاده از روش سنتز روایت انجام شد و نتایج مطالعات در قالب جدول یک خلاصه گردید. همچنین، مطالعات براساس نوع داده، نوع الگوریتم و هدف مدل گروه‌بندی شدند تا امکان مقایسه ساخت یافته یافته‌ها فراهم شود. به دلیل ناهمگونی در نوع داده‌ها، الگوریتم‌ها و معیارهای ارزیابی، انجام متاآنالیز امکان‌پذیر نبود.

برای ارزیابی خطر سوگیری در مطالعات وارد شده، با توجه به ماهیت روش شناختی آن‌ها، از چک‌لیست ارزیابی انتقادی موسسه جوآنا بری گز^۳ برای مطالعات تحلیلی مقطعی استفاده شد. از آنجا که اغلب مطالعات حوزه داده کاوی در سلامت مبتنی بر تحلیل ثانویه داده‌ها و توسعه یا ارزیابی مدل‌های پیش‌بینی هستند و از نظر طراحی تحلیلی ماهیتی مقطعی دارند، این ابزار به عنوان چارچوبی مناسب برای ارزیابی خطر سوگیری انتخاب شد.

چک‌لیست جوآنا بری گز شامل هشت معیار کلیدی است که جنبه‌هایی مانند وضوح معیارهای ورود، توصیف نمونه و منبع داده، اعتبار اندازه‌گیری متغیرها، تناسب روش‌های تحلیلی و مدیریت عوامل مخدوش‌کننده را پوشش می‌دهد. ارزیابی کیفیت به طور مستقل توسط دو پژوهشگر انجام شد. سطح توافق بین ارزیابان بزرگ‌تر از ۰/۷ بود (ضریب توافق کاپا = ۰/۸۱) در موارد عدم توافق، تصمیم نهایی از طریق بحث یا مشورت با پژوهشگر سوم اتخاذ شد.

نتایج ارزیابی کیفیت به صورت توصیفی و در قالب سطوح خطر سوگیری پایین، متوسط و بالا گزارش شد و هیچ مطالعه‌ای صرفاً براساس نتایج ارزیابی کیفیت حذف نشد. تمامی ۴۴ مطالعه وارد شده دارای خطر سوگیری پایین تا متوسط بودند. جزئیات ارزیابی کیفیت هر مطالعه در جدول ۱ ارائه شده است.

پزشکی، داده‌های ژنومیک، بیوانفورماتیک یا پردازش سیگنال‌های زیستی بدون تمرکز بر پیش‌بینی یا تشخیص بیماری‌های بالینی پرداخته بودند.

۳. مقالاتی که فاقد متن کامل یا دارای اطلاعات ناکافی برای استخراج داده‌های مورد نیاز بودند.

۴. مقالات تکراری یا مطالعاتی با نتایج یکسان که در بیش از یک منبع منتشر شده بودند.

۵. مقالاتی که به زبان‌هایی غیر از فارسی و انگلیسی منتشر شده بودند.

دامنه حوزه سلامت در این مطالعه شامل مطالعات مرتبط با مراقبت‌های بالینی، پیش‌بینی و تشخیص بیماری‌ها و تحلیل داده‌های بیماران در محیط‌های درمانی بوده و حوزه‌هایی نظیر سلامت عمومی کلان، دندان‌پزشکی و مطالعات غیر بالینی در دامنه این مرور قرار نگرفته‌اند.

در مرحله نخست، مطالعات از پایگاه‌های اطلاعاتی مورد نظر با استفاده از استراتژی جستجوی تعریف شده استخراج شدند. پس از حذف مقالات تکراری، عناوین و چکیده‌ها توسط دو پژوهشگر به طور مستقل و براساس معیارهای ورود و خروج بررسی گردید. در مرحله غربالگری عنوان و چکیده و همچنین بررسی متن کامل مقالات، دو نویسنده به طور مستقل مطالعات را ارزیابی کردند. برای ارزیابی میزان توافق بین دو پژوهشگر در مرحله غربالگری عنوان و چکیده، ضریب کاپای کوهن^۲ محاسبه شد که نشان‌دهنده سطح توافق قابل قبول بین ارزیابان بود. در موارد اندک عدم توافق، اختلافات از طریق بحث میان دو ارزیاب برطرف شد و در صورت لزوم، نویسنده سوم به عنوان داور نهایی تصمیم‌گیری کرد.

برای استخراج داده‌ها، از یک چک‌لیست استاندارد شده که متناسب با ماهیت مطالعات داده کاوی طراحی گردیده بود استفاده شد. آیتم‌های استخراج شده شامل: سال انتشار، کشور مطالعه، نوع داده مورد استفاده (مانند پرونده‌های الکترونیکی سلامت، داده‌های بالینی یا داده‌های

جدول ۱: کاربردهای داده کاوی در منابع پیشین

منبع	سال	کشور	الگوریتم‌های داده کاوی	نوع داده	اندازه نمونه	معیار عملکرد	طراحی مطالعه	نوع مسئله	هدف پژوهش
[۷]	۲۰۱۱	هند	Decision Tree, Naive Bayes	داده‌های بالینی / سوابق الکترونیکی سلامت	۹۰۹ رکورد (۴۵۵ برای آموزش، ۴۵۴ برای تست)	دقت	مطالعه گذشته‌نگر مبتنی بر تحلیل ثانویه داده‌ها با مقایسه الگوریتم‌های یادگیری ماشین	طبقه‌بندی	پیش‌بینی بیماری قلبی
[۸]	۲۰۱۲	هند	Genetic Algorithm, Apriori	داده‌های بالینی قلبی	گزارش نشده	دقت	تجربی	استخراج قوانین انجمنی	پیش‌بینی بیماری قلبی با الگوریتم ژنتیک برای استخراج قوانین کشف قوانین پیش‌بینی دقیق‌تر در جمعیت Andhra Pradesh
[۹]	۲۰۱۳	ایران	CHAID, Neural Network, SVM, Decision Tree	داده‌های بالینی	۳۵۰ رکورد (۸۰ درصد برای آموزش، ۲۰ درصد برای تست)	دقت	تحلیلی	طبقه‌بندی	طراحی مدل پیش‌بینی انفارکتوس میوکارد با استخراج الگوهای پنهان از پرونده‌ها
[۱۰]	۲۰۱۴	هند	HITS Algorithm, Weighted Associative Classifiers	داده‌های بالینی	۲۹۶ رکورد	درجه احتمال بیماری، دقت	بررسی مقطعی	طبقه‌بندی	ارائه معیار آماری DDP برای پیش‌بینی بیماری‌ها با داده کاوی
[۱۱]	۲۰۱۴	هند	Decision Tree, Naive Bayes, k-Nearest Neighbour	داده‌های بالینی	۳۰۳ نمونه	دقت، مقایسه مدل‌ها	تجربی	طبقه‌بندی	پیش‌بینی بیماری قلبی با تکنیک‌های طبقه‌بندی داده کاوی
[۱۲]	۲۰۱۵	پاکستان	Decision Tree, Neural Network, Naive Bayes	داده‌های بالینی	۵۹۷ رکورد	TP, FP rate, rate, Precision, F-measure, ROC curve	تجربی	طبقه‌بندی	برای یافتن الگوهای جالب از داده‌های بیماران قلبی برای بهبود تشخیص و درمان
[۱۳]	۲۰۱۵	آمریکا	CARE - [Collaborative - Distributed Computing [Filtering]	سوابق الکترونیکی سلامت	بیش از ۱۳ میلیون بیمار منحصربه‌فرد (۴ میلیون با بیش از ۵۰۰/۱۰ ویزیت؛ ۰۰۰/۱۰ برای آموزش و ۰۰۰/۱۰ برای ارزیابی)	بیش از ۱۳ میلیون بیمار منحصربه‌فرد (۴ میلیون با بیش از ۵۰۰/۱۰ ویزیت؛ ۰۰۰/۱۰ برای آموزش و ۰۰۰/۱۰ برای ارزیابی)	مطالعه موردی [تحلیل گذشته‌نگر روی داده‌های موجود؛ مقیاس‌پذیری با محاسبات توزیع شده و تحلیل زمینه‌ای]	طبقه‌بندی	بررسی چالش‌های یکپارچه‌سازی الگوریتم‌های پیش‌بینی بیماری در سیستم‌های بالینی
[۱۴]	۲۰۱۵	ایتالیا	SVM, Decision Tree, Naive Bayesian classifier	داده‌های بالینی	۱۱۰۰۶ رکورد	دقت، تحلیل هزینه خطا	گذشته‌نگر	طبقه‌بندی	غربالگری سلامت خون اهداکنندگان
[۱۵]	۲۰۱۵	کره جنوبی	-	داده‌های بالینی	۸۸۸۹ بیمار	دقت، AUC	تجربی	طبقه‌بندی	توسعه مدل پیش‌بینی ریسک CHD با fuzzy logic و decision tree برای جمعیت کره‌ای

جدول ۱: کاربردهای داده‌کاوی در منابع پیشین

منبع	سال	کشور	الگوریتم‌های داده‌کاوی	نوع داده	اندازه نمونه	معیار عملکرد	طراحی مطالعه	نوع مسئله	هدف پژوهش
[۱۶]	۲۰۱۵	هند	Decision Tree, Naive Bayes, SVM, k-NN	داده‌های بالینی	۷۶۸ نمونه	دقت پیش‌بینی، مقایسه مدل‌ها	تجربی	طبقه‌بندی	تشخیص دیابت با تکنیک‌های داده‌کاوی برای کاهش تست‌ها و افزایش دقت تشخیص
[۱۷]	۲۰۱۶	هند	KNN, ID3, Decision Tree, Neural Network	داده‌های بالینی	گزارش نشده	دقت	تجربی	طبقه‌بندی	پیش‌بینی سطح خطر بیماری قلبی با استفاده از تکنیک‌های داده‌کاوی
[۱۸]	۲۰۱۷	هند	Naïve Bayes, SMO, REP Tree, J48, MLP	داده‌های بالینی	۱۴۶۷ نمونه	دقت	تجربی	طبقه‌بندی	تحلیل و پیش‌بینی سرطان سینه و دیابت با تکنیک‌های طبقه‌بندی در WEKA و مقایسه عملکرد الگوریتم‌ها
[۱۹]	۲۰۱۷	هند	Generate class-based association, classification, Keyword-Based Clustering	داده‌های بالینی	گزارش نشده	دقت، زمان محاسباتی، کارایی	تجربی	استخراج قوانین انجمنی و خوشه‌بندی	پیش‌بینی بیماری با داده‌کاوی
[۲۰]	۲۰۱۷	ایران	Apriori, ID3	داده‌های بالینی	۴۹۰ رکورد (۲۲۰ مبتلا به سرطان معده، ۲۷۰ سالم)	دقت پیش‌بینی، قوانین انجمنی	مقطعی	استخراج قوانین انجمنی و طبقه‌بندی	تعیین عوامل مؤثر در بروز سرطان معده
[۲۱]	۲۰۱۸	ایران	SVM, ANN	داده‌های بالینی	۱۳۲۴ رکورد	خطای مطلق پایین‌تر، آماره هاسمر-لمشو بالاتر، حساسیت، ویژگی، مساحت زیر منحنی	طبقه‌بندی	طبقه‌بندی	پیش‌بینی بیماری عروق کرونر قلب
[۲۲]	۲۰۱۸	کره جنوبی	Decision Tree, Neural Network	داده‌های بالینی و نظرسنجی سلامت	گزارش نشده	متوسط خطای مربعات، نرخ طبقه‌بندی اشتباه، ضریب جینی	مقطعی	طبقه‌بندی	شناسایی عوامل خطر بیماری پریدنتال
[۲۳]	۲۰۱۹	ایران	Regression, Neural Network, AdaBoost, Artificial Neural	داده‌های بالینی و سلامت عمومی	۱۰۰۰۰ نفر	حساسیت	مقطعی	طبقه‌بندی	پیش‌بینی بیماری‌های عروق کرونر با استفاده از داده‌های سلامت جمعیتی
[۲۴]	۲۰۱۹	هند	C4.5, PRISM, Hybrid	داده‌های بالینی	گزارش نشده	دقت، نرخ مثبت واقعی، نرخ مثبت کاذب	داده ثانویه	طبقه‌بندی	طراحی یک سیستم پیش‌بینی بیماری‌های عمومی با استفاده از تکنیک‌های داده‌کاوی ترکیبی
[۲۵]	۲۰۲۰	عراق	Naïve Bayes, J48, SMO, MLP	داده‌های بالینی	گزارش نشده	دقت، منحنی ROC، مقدار AUC	تجربی	طبقه‌بندی	مقایسه الگوریتم‌های پیش‌بینی بیماری قلبی در محیط WEKA

جدول ۱: کاربردهای داده‌کاوی در منابع پیشین

منبع	سال	کشور	الگوریتم‌های داده‌کاوی	نوع داده	اندازه نمونه	معیار عملکرد طراحی مطالعه	نوع مسئله	هدف پژوهش
[۲۶]	۲۰۲۰	هند	Decision Tree, Particle Swarm Optimization	داده‌های بالینی	۳۰۳ نمونه	دقت	طبقه‌بندی	پیش‌بینی و طبقه‌بندی بیماری قلبی با مدل هیبریدی J48-PSO برای بهبود دقت تشخیص
[۲۷]	۲۰۲۰	هند	Random Forest, SVM, Naive Bayes	داده‌های بالینی	گزارش نشده	دقت، مقایسه الگوریتم‌ها و انتخاب بهترین برای پیش‌بینی	طبقه‌بندی	پیش‌بینی بیماری عروق کرونر با داده‌کاوی
[۲۸]	۲۰۲۱	بنگلادش	Naive Bayes, SVM, KNN, Decision Tree, Neural Network, Logistic Regression, Random Forest	داده‌های بالینی	دیتاست استاندارد UCI Cleveland Heart Disease	دقت	طبقه‌بندی	پیش‌بینی بیماری قلبی با انتخاب ویژگی‌ها
[۲۹]	۲۰۲۱	هند	Random Forest, SVM, KNN, Naive Bayes, Regression for Logistic	داده‌های بالینی و سنسور محور	گزارش نشده	دقت	طبقه‌بندی	مقایسه الگوریتم‌های داده‌کاوی برای پیش‌بینی بیماری‌ها
[۳۰]	۲۰۲۱	هند	Naive Bayes, Decision Tree	داده‌های بالینی	۳۰۳ نمونه	دقت	طبقه‌بندی	پیش‌بینی بیماری قلبی با تکنیک‌های داده‌کاوی برای کاهش تست‌ها و بهبود دقت
[۳۱]	۲۰۲۳	ایران	Support Vector Machine [SVM], Random Forest [RF], Naive Bayes, KNN	داده‌های بالینی و دموگرافیک	۴۱۴۵ نمونه	دقت	مقطعی	پیش‌بینی ابتلا به بیماری مزمن کلیوی با استخراج قواعد انجمنی طبقه‌بندی
[۳۲]	۲۰۲۱	عراق	Naive Bayes, Random Forest, multilayer perceptron, sequential minimal optimization	داده‌های بالینی	گزارش نشده	دقت، منحنی ROC، مقدار PRC	مقطعی	تحلیل الگوریتم‌های داده‌کاوی برای پیش‌بینی بیماری قلبی
[۳۳]	۲۰۲۰	ایران	Decision Tree [[C4.5	داده‌های بالینی	گزارش نشده	دقت، ویژگی، حساسیت	مطالعه موردی	پیش‌بینی وضعیت بیماران نادر
[۳۴]	۲۰۲۲	بنگلادش	Random Forest, Decision Tree, KNN, Naive Bayes, SVM	داده‌های بالینی	گزارش نشده	دقت، F1-score، مقایسه الگوریتم‌ها	طبقه‌بندی	پیش‌بینی بیماری قلبی با الگوریتم‌های مختلف

جدول ۱: کاربردهای داده‌کاوی در منابع پیشین

منبع	سال	کشور	الگوریتم‌های داده‌کاوی	نوع داده	اندازه نمونه	معیار عملکرد	طراحی مطالعه	نوع مسئله	هدف پژوهش
[۳۵]	۲۰۲۲	هند	Random Forest, Decision Tree, Naïve Bayes, SVM, Logistic Regression	داده‌های بالینی	گزارش نشده	دقت، recall، F1-score مقایسه الگوریتم‌ها	تجربی	طبقه‌بندی	پیش‌بینی بیماری‌های مختلف با استفاده از داده‌های پزشکی
[۳۶]	۲۰۲۳	ایران	Random Forest	داده‌های بالینی	۱۵۰ نفر (زنان) دانشجوی دانشگاه آزاد شیراز؛ انتخاب تصادفی ساده	دقت، ویژگی	مقطعی	طبقه‌بندی	پیش‌بینی اختلال اضطراب فراگیر در زنان دانشجو
[۳۷]	۲۰۲۳	ایران	Neural Network, Simple Bayes, Bayesian Network, Logistic Regression, J48, ID3	داده‌های بالینی	۴۱۱۸ بیمار	دقت، ویژگی	گذشته‌نگر	رگرسیون	طراحی سیستم پشتیبان برای پیش‌بینی بقای بیماران مبتلا به ملانوما با دقت بالا
[۳۸]	۲۰۲۴	هند	Random Forest, mRMR	داده‌های بالینی	گزارش نشده	دقت، precision، recall، F1-score، log loss	تجربی	طبقه‌بندی	تشخیص خودکار دیابت با یادگیری ماشین
[۳۹]	۲۰۲۵	ایران	Regression, Random Forest, SVM, Gradient Boosting, Neural Networks	داده‌های عمومی و بالینی بیماران COVID-۱۹	۱/۰۴۸,۵۷۶ بیمار	دقت، حساسیت، Specificity, F1 score	گذشته‌نگر	طبقه‌بندی	پیش‌بینی شدت بیماری و مرگ‌ومیر با استفاده از یادگیری ماشین برای تخصیص منابع پزشکی
[۴۰]	۲۰۲۵	آمریکا	Gradient-Boosted Trees [GBT], LightGBM, Random Forest [RF], XGBoost, Logistic Regression [LR], AdaBoost	سوابق الکترونیکی سلامت	۰۱۲/۴ مورد ADRD و ۷۲۳/۱۱۹ کنترل	AUC-ROC، دقت، حساسیت، specificity, F1 score; SHAP	گذشته‌نگر	طبقه‌بندی	پیش‌بینی زودرس ADRD با استفاده از داده‌های EHR برای بهبود تشخیص و مدیریت منابع
[۴۱]	۲۰۲۵	قطر / بنگلادش	Tab Transformer، Random Forest، ExtraTree Classifier، SVM	داده‌های بالینی	۱۱۹۰ نمونه	AUC، دقت	تجربی	طبقه‌بندی	پیش‌بینی بیماری قلبی با مدل هیبریدی Tab ML و Transformer برای تشخیص زودرس و مدیریت درمان

جدول ۱: کاربردهای داده کاوی در منابع پیشین

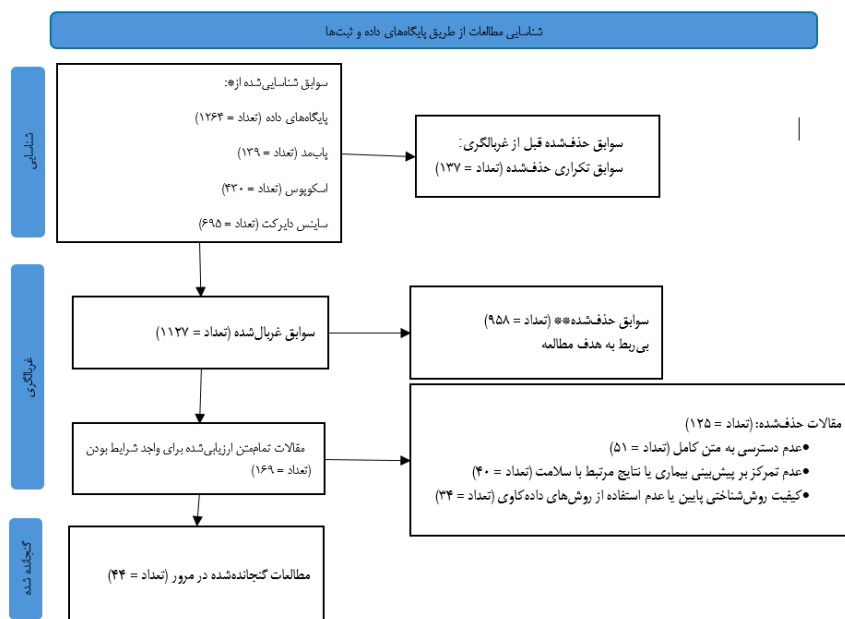
منبع	سال	کشور	الگوریتم های داده کاوی	نوع داده	اندازه نمونه	معیار عملکرد طراحی مطالعه	نوع مسئله	هدف پژوهش
[۴۲]	۲۰۲۵	اردن/ عربستان/ امارات/ چک/ آفریقای جنوبی	Gradient Boosting Machines [GBM], Deep Neural Networks [DNN], Logistic Regression, Random Forest, SVM, Neural Networks	داده های بالینی	تعداد دقیق گزارش نشده	دقت، Precision, Recall, F1-Score, AUROC	طبقه بندی	پیشنهاد چارچوب ML برای پیش بینی نتایج بیماری و بهینه سازی مراقبت بیمار محور
[۴۳]	۲۰۲۵	امارات متحده عربی	Naive Bayes, domain expert, SHAP, LIME	سوابق الکترونیکی سلامت	۱۱۵ بیمار	دقت، AUC, F1-score, MCC	طبقه بندی	مقایسه مدل های ML برای تشخیص زودرس MPS با EHR و شناسایی ویژگی های کلیدی با SHAP و LIME
[۴۴]	۲۰۲۵	پاکستان/ عربستان سعودی/ کره جنوبی	Mutual Information, SMOTE, PSO-ANN, Logistic Regression, Random Forest	داده های بررسی تغذیه و سلامت ملی	تعداد دقیق گزارش نشده	دقت	طبقه بندی	پیش بینی CHD با مدل های ML پیشرفته برای بهبود ارزیابی ریسک قلبی عروقی
[۴۵]	۲۰۲۴	آمریکا	Generalized Linear Model, L1-regularized Logistic Regression, Random Forest, XGBoost, J48 Decision Tree	سوابق الکترونیکی سلامت	۳۹/۷۴۳ بیمار	دقت، AUC	طبقه بندی	شناسایی و ارجاع بیماران در معرض خطر از PCP به متخصص چشم با ML براساس EHR
[۴۶]	۲۰۲۵	پاکستان/ کره جنوبی	TabNet, XGBoost, Logistic Regression [LR], SVM	داده های بالینی	اندازه دقیق گزارش نشده	دقت، F1-score, Precision, Recall, ROC-AUC, PR-AUC, MCC	طبقه بندی	پیشنهاد چارچوب هیبریدی TabNet با stacked ensemble برای پیش بینی بیماری قلبی و بهبود دقت و تفسیرپذیری
[۴۷]	۲۰۲۵	چین	XGBoost, Decision Tree, Multilayer Perceptron, SHAP	داده های بالینی	۱۱۴۳۶ نمونه	AUC, F1-score	طبقه بندی	پیش بینی زودرس CKD با مدل ML تفسیرپذیر [بدون دیابت/ فشارخون/ بیماری قلبی] برای تشخیص زودرس و تأخیر پیشرفت بیماری
[۴۸]	۲۰۲۵	آمریکا	ML, knowledge graphs	سوابق الکترونیکی سلامت	۷۴۷ مورد	F-score	طبقه بندی	کاهش تأخیر تشخیص AHP با مدل های ML براساس EHR برای پیش بینی ارجاع و تشخیص مثبت

جدول ۱: کاربردهای داده کاوی در منابع پیشین							
منبع	سال	کشور	الگوریتم‌های داده کاوی	نوع داده	اندازه نمونه	معیار عملکرد طراحی مطالعه	هدف پژوهش
[۴۹]	۲۰۲۵	آمریکا	XGBoost ^۴ , Random Forest [RF] ^۵ , Neural Network	سوابق الکترونیکی سلامت	۱۰۴۵۶۵ بیمار	AUROC	پیش‌بینی AKI و CKD در دوره پساکوئید با ML براساس EHR ملی، ارزیابی اهمیت تاریخی کوئید و توسعه اپلیکیشن وب برای استفاده بالینی
[۵۰]	۲۰۲۵	چین	Light GBM ^۶ , SHAP	داده‌های بالینی	۳۳۵۸ نمونه	AUC	توسعه مدل ML تفسیرپذیر برای پیش‌بینی ریسک CVD ۳ ساله در T2D جدید با ترکیب TyG index و ویژگی‌های اولتراسوند و ابزار وب برای کاربرد بالینی

یافته‌ها:

مطالعه که با اهداف پژوهش هم‌راستا نبودند، حذف شدند و ۱۶۹ مقاله مرتبط برای بررسی متن کامل انتخاب شدند. از این میان، ۱۲۵ مقاله به دلایل مختلف حذف شدند که شامل نبود متن کامل، عدم تمرکز بر پیش‌بینی بیماری، یا ضعف روش‌شناسی بود. در نهایت، ۴۴ مطالعه که دارای معیارهای ورود بودند، در این مرور نظام‌مند گنجانده شدند.

شاده‌های مطابق با (شکل ۱)، در ابتدا ۱۲۶۴ مطالعه با جستجو در پایگاه‌های داده معتبر شناسایی و استخراج شدند. پس از حذف ۱۳۷ مورد تکراری، ۱۱۲۷ مطالعه برای بررسی عنوان و چکیده باقی ماند. در این مرحله، ۹۵۸



شکل ۱: فلوچارت انتخاب منابع

بحث و نتیجه گیری:

نتایج حاصل از مرور ۴۴ مطالعه منتخب نشان می‌دهد که کاربرد داده‌کاوی و یادگیری ماشین در حوزه سلامت عمدتاً باهدف پیش‌بینی بیماری‌ها، تشخیص زودهنگام، پشتیبانی از تصمیم‌گیری بالینی و مدیریت منابع درمانی انجام شده است. براساس یافته‌ها، تمرکز اصلی مطالعات بر بیماری‌های قلبی و عروقی بوده و پس‌از آن، بیماری‌هایی نظیر دیابت، بیماری‌های کلیوی، سرطان‌ها و برخی اختلالات مزمن موردتوجه قرار گرفته‌اند [۲۳،۲۷،۳۱،۳۲، ۳۳،۳۷،۴۱،۴۴،۴۶،۴۷،۵۰]. این گرایش تمرکز یافته بر بیماری‌های قلبی و متابولیک را می‌توان با دو عامل اصلی توجیه کرد. نخست، بار بالای بیماری و هزینه‌های سرسام‌آور درمانی این دسته از بیماری‌ها در نظام سلامت است که محققان را بر آن داشته تا با استفاده از داده‌کاوی به دنبال راهکارهایی برای کاهش هزینه‌ها و مدیریت بهتر بیماران باشند. به‌طور مشابه رونقی و حسینی نیز نشان دادند استفاده از فناوری‌های دیجیتالی می‌تواند باعث تسهیل فرایندهای سلامت شود [۵۱]. دوم، ماهیت داده‌های بالینی این بیماری‌ها است؛ داده‌های مربوط به دیابت و بیماری‌های قلبی اغلب ساخت یافته، کمی و متوالی هستند که برای الگوریتم‌های یادگیری ماشین بسیار مناسب می‌باشند. اثرات این تمرکز در مطالعات موردبررسی، بهبود چشمگیر دقت تشخیص زودهنگام و کاهش نرخ مرگ‌ومیر از طریق پیش‌بینی دقیق بحران‌های قلبی و کنترل سطح قند خون بوده است. بااین‌حال، تمرکز کمتر بر بیماری‌های نادر یا دارای داده‌های غیر ساختاریافته نشان‌دهنده چالش‌های موجود در اعمال داده‌کاوی در این حوزه‌هاست.

بررسی نوع الگوریتم‌های به‌کاررفته نشان می‌دهد که در اغلب مطالعات، الگوریتم‌های یادگیری ماشین کلاسیک نظیر درخت تصمیم، بیز ساده^۴، ماشین بردار پشتیبان، شبکه‌های عصبی چندلایه، جنگل تصادفی و روش‌های تقویتی بیشترین کاربرد را داشته‌اند [۲۱،۲۵،۲۷،۲۸،۳۲،۳۴،۳۵]. این یافته با ماهیت داده‌های استفاده شده در مطالعات که عمدتاً شامل داده‌های بالینی ساخت یافته و سوابق

4 Naïve Bayes

الکترونیکی سلامت بوده‌اند، هم‌خوانی دارد. مطالعات اولیه؛ ۲۰۲۰-۲۰۱۰ عمدتاً بر مقایسه عملکرد الگوریتم‌ها و انتخاب مدل مناسب برای پیش‌بینی بیماری‌ها تمرکز داشته‌اند [۷،۱۲]. در مقابل، مطالعات جدیدتر، به‌ویژه از سال ۲۰۲۱ به بعد، علاوه بر بهبود دقت پیش‌بینی، بر تفسیرپذیری مدل‌ها، انتخاب ویژگی و کاربرد بالینی نتایج تأکید بیشتری داشته‌اند [۳۷،۴۰،۴۳،۴۷،۵۰]. در مطالعات منتشرشده در سال‌های ۲۰۲۴ و ۲۰۲۵، استفاده از مدل‌های پیشرفته یادگیری ماشین و روش‌های هیبریدی مانند تقویت گرادیانی^۵، اکس‌تریم گرادیان بوستینگ^۶، لایت‌جی‌بی‌ام^۷ و معماری‌های مبتنی بر ترنسفورمر^۸ برای داده‌های جدولی گزارش شده است [۴۰-۴۲،۴۶،۴۷]. این مطالعات اغلب از داده‌های بزرگ سوابق الکترونیکی سلامت بهره برده و علاوه بر پیش‌بینی، به شناسایی عوامل مؤثر و توضیح‌پذیری تصمیمات مدل با استفاده از روش‌هایی مانند توضیحات افزایشی شاپلی^۹ و توضیح‌پذیری موضعی مستقل از مدل^{۱۰} پرداخته‌اند [۴۰،۴۳،۴۷،۵۰].

ازنظر عملکرد الگوریتم‌ها، نتایج مطالعات نشان می‌دهد که هیچ الگوریتم برتری مطلق در تمامی مسائل ندارد و عملکرد مدل‌ها به نوع بیماری، کیفیت داده و ویژگی‌های ورودی وابسته است. برای مثال، در پیش‌بینی بیماری قلبی، جنگل تصادفی و شبکه بیزی^{۱۱} عملکرد بهتری نسبت به سایر الگوریتم‌ها گزارش شده است [۳۲،۳۴،۴۴]، درحالی‌که در برخی مطالعات، ماشین بردار پشتیبان^{۱۲} نسبت به شبکه‌های عصبی عملکرد مطلوب‌تری داشته است [۴]. چالش مشترک مطرح‌شده در بسیاری از مطالعات، کیفیت داده‌های سلامت بوده است. داده‌های بالینی اغلب ناقص، ناهمگون و دارای نویز هستند و این مسئله می‌تواند دقت مدل‌های پیش‌بینی را کاهش دهد [۲۰]. بنابراین، مراحل پیش‌پردازش شامل پاک‌سازی داده‌ها، مدیریت داده‌های

5 Gradient Boosting

6 XGBoost

7 LightGBM

8 Transformer

9 SHAP: Shapley Additive Explanations

10 LIME: Local Interpretable Model-agnostic Explanations

11 Bayes Net

12 Support Vector Machine

ماشین کلاسیک همچنان ستون اصلی مطالعات این حوزه را تشکیل می‌دهند؛ با این حال، در سال‌های اخیر گرایش به سمت مدل‌های پیشرفته‌تر، روش‌های هیبریدی و مدل‌های تفسیرپذیر به‌ویژه در مطالعات مبتنی بر سوابق الکترونیکی سلامت افزایش یافته است. این روند نشان‌دهنده حرکت از تمرکز صرف بر دقت پیش‌بینی به سمت کاربردپذیری بالینی و شفافیت تصمیمات مدل‌ها است. با وجود پتانسیل بالای داده‌کاوی در بهبود نظام سلامت، چالش‌هایی مانند کیفیت داده، ناهمگونی منابع اطلاعاتی و محدودیت‌های پیاده‌سازی بالینی همچنان باقی است.

پیشنهاد می‌شود توسعه زیرساخت‌های داده، استانداردسازی اطلاعات سلامت و تقویت همکاری میان متخصصان بالینی و علم داده برای بهره‌برداری مؤثر از این فناوری موردتوجه قرار گیرد. استفاده از روش کیفی در بازه زمانی محدود و پایگاه‌های مشخص از جمله محدودیت‌های این مطالعه بوده است. لذا در مطالعات آتی استفاده از روش‌های تجربی و کمی جهت پیاده‌سازی مدل‌های داده‌کاوی در حوزه بهداشت و درمان موردتوجه قرار گیرد.

ملاحظات اخلاقی:

پیروی از اصول اخلاق در پژوهش

این مطالعه دارای تأییدیه اخلاقی به شماره IR.US.PSYEDU.REC.1404.007 از دانشگاه شیراز است.

حامی مالی

این مطالعه تحت حمایت مالی هیچ سازمان یا مؤسسه‌ای نبوده است.

مشارکت نویسندگان

محمدحسین رونقی و احمدرضا ملک پور مفهوم‌سازی، روش‌شناسی، تحقیق و بررسی، تحلیل، ویراستاری و نهایی سازی مقاله را انجام داده‌اند.

تعارض منافع

هیچ گونه تضاد منافعی از سوی پژوهشگران گزارش نشده است.

گم‌شده و انتخاب ویژگی نقش کلیدی در موفقیت مدل‌های داده‌کاوی دارند [۲۱،۳۷].

در مجموع، یافته‌های این مرور نشان می‌دهد که گرچه حرکت به سمت مدل‌های پیشرفته‌تر در سال‌های اخیر آغاز شده است، اما اکثر مطالعات مرور شده همچنان بر داده‌های بالینی و سوابق الکترونیکی سلامت و الگوریتم‌های یادگیری ماشین غیر تصویری و غیر ژنومی متکی هستند. می‌توان گفت؛ داده‌کاوی و یادگیری ماشین نقش مؤثری در پیش‌بینی بیماری‌ها، تشخیص زودهنگام و پشتیبانی از تصمیم‌گیری بالینی دارند و بیشترین کاربرد آن‌ها در بیماری‌های قلبی و عروقی و اختلالات مزمن مشاهده شده است. این یافته‌ها با مطالعات پیشین همخوانی دارد. همانند آنچه رونقی و حسینی نشان دادند، فناوری‌های دیجیتالی و داده‌کاوی می‌توانند فرایندهای سلامت را تسهیل کنند. با این حال، مطالعات جدیدتر نسبت به گذشته بر جنبه‌های تفسیرپذیری و کاربرد بالینی تأکید بیشتری دارند [۵۱].

بررسی سیستماتیک و به‌روز بودن مطالعات از نقاط قوت این مطالعه است. با این حال، از نقاط ضعف آن می‌توان به استفاده از رویکرد کیفی و محدودیت در پایگاه‌ها و بازه زمانی اشاره کرد.

تداوم الگوریتم‌های کلاسیک یادگیری ماشین حتی در سال‌های ۲۰۲۴ و ۲۰۲۵ علی‌رغم پیشرفت مدل‌های پیشرفته یکی از یافته‌های غیرمنتظره پژوهش است. یافته‌های این مرور می‌تواند راهنمایی برای محققان، پزشکان و سیاست‌گذاران باشد تا بر بهبود کیفیت داده‌ها و توسعه مدل‌های قابل تفسیر سرمایه‌گذاری بیشتری انجام دهند و در نهایت کیفیت مراقبت‌های بالینی را ارتقا بخشند.

یافته‌های این مرور نظام‌مند نشان داد که داده‌کاوی و یادگیری ماشین نقش مؤثری در پیش‌بینی و تشخیص بیماری‌ها و پشتیبانی از تصمیم‌گیری بالینی ایفا می‌کنند. بیشترین کاربرد این روش‌ها در حوزه بیماری‌های قلبی و عروقی و بیماری‌های مزمن مشاهده شد و داده‌های مورد استفاده عمدتاً شامل داده‌های بالینی ساخت یافته و سوابق الکترونیکی سلامت بودند. الگوریتم‌های یادگیری



References

1. Ahmed N, Ameer R, Khan A, SN N, Gull S. Data mining in healthcare: An overview of applications, techniques, and challenges. *IJAIMS*. 2024; 3(1):1-10. DOI: [10.58921/ijai-ms.v3i1.102](https://doi.org/10.58921/ijai-ms.v3i1.102)
2. Soleimanfar E, Navabakhsh M, Sebt MV. Identification of patterns and factors affecting the health of employees based on datamining of occupational examinations with the purpose of promoting occupational health. *Iran J Health Educ Health Promot*. 2019; 7(3):295-305. [In Persian] DOI: [10.29252/ijhehp.7.3.295](https://doi.org/10.29252/ijhehp.7.3.295)
3. Dekamini F, Ehsanifar M. Comparison of the efficiency of data mining algorithms in predicting the diagnosis of diabetes. *IJDM*. 2021; 21(4):264-75. [In Persian] [Link](#)
4. Yadav S, Singh MK, Kumar P. Data mining applications for enhancing healthcare services: A comprehensive review. *Int J Eng Technol Manag Sci*. 2023; 7(5):325-33. DOI: [10.46647/ijetms.2023.v07i05.038](https://doi.org/10.46647/ijetms.2023.v07i05.038)
5. Mathias JS, Agrawal A, Feinglass J, Cooper AJ, Baker DW, Choudhary A. Development of a 5 year life expectancy index in older adults using predictive mining of electronic health record data. *J Am Med Inform Assoc*. 2013; 20(e1):e118-24. DOI: [10.1136/amiajnl-2012-001360](https://doi.org/10.1136/amiajnl-2012-001360) PMID: [23538722](https://pubmed.ncbi.nlm.nih.gov/23538722/)
6. Kolling ML, Furstenuau LB, Sott MK, Rabaioli B, Ulmi PH, Bragazzi NL, et al. Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development. *Int J Environ Res Public Health*. 2021; 18(6):3099. DOI: [10.3390/ijerph18063099](https://doi.org/10.3390/ijerph18063099) PMID: [33802880](https://pubmed.ncbi.nlm.nih.gov/33802880/)
7. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. *Int J Comput Appl*. 2011; 17:43-48. DOI: [10.5120/2237-2860](https://doi.org/10.5120/2237-2860)
8. Jabbar MA, Deekshatulu BL, Chandra P. An evolutionary algorithm for heart disease prediction. In *International conference on information processing 20120* (pp. 78-389). Berlin, Heidelberg: Springer Berlin Heidelberg. DOI: [10.1007/978-3-642-31686-9_44](https://doi.org/10.1007/978-3-642-31686-9_44)
9. Safdari R, GhaziSaeedi M, Arji G, Gharooni M, Soraki M, Nasiri M. A model for predicting myocardial infarction using data mining techniques. *IJMI*. 2013; 2(4):1-6. [Link](#)
10. Nagavelli R, Rao CG. Degree of Disease possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining. *Proceeding of the International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2014)*. IEEE; 2014. p 1-6. DOI: [10.1109/ICRAIE.2014.6909265](https://doi.org/10.1109/ICRAIE.2014.6909265)
11. Joshi S, Nair MK. Prediction of heart disease using classification-based data mining techniques. In *computational intelligence in data mining-volume 2: Proceedings of the International Conference on CIDM, New Delhi: Springer India*; 2014. p 503-511. DOI: [10.1007/978-81-322-2208-8_46](https://doi.org/10.1007/978-81-322-2208-8_46)
12. Shafique U, Majeed F, Qaiser H, Mustafa IU. Data mining in healthcare for heart diseases. *Int J Innov Appl Stud*. 2015; 10(4):1312-20. [Link](#)
13. Feldman K, Davis D, Chawla NV. Scaling and contextualizing personalized healthcare: A case study of disease prediction algorithm integration. *J Biomed Inform*. 2015; 57:377-85. DOI: [10.1016/j.jbi.2015.07.017](https://doi.org/10.1016/j.jbi.2015.07.017) PMID: [26254848](https://pubmed.ncbi.nlm.nih.gov/26254848/)
14. Khalilinezhad M, Minaei B, Vernazza G, Dellepiane S. Prediction of healthy blood with data mining classification by using Decision Tree, Naive Bayesian and SVM approaches. *Proceeding of the Sixth International Conference on Graphic and Image Processing (ICGIP 2014)*. SPIE; 2015. 9443: 544-53. DOI: [10.1117/12.2179871](https://doi.org/10.1117/12.2179871)
15. Kim J, Lee J, Lee Y. Data-mining-based coronary heart disease risk prediction model using fuzzy logic and decision Tree. *Healthc Inform Res*. 2015; 21(3):167-74. DOI: [10.4258/hir.2015.21.3.167](https://doi.org/10.4258/hir.2015.21.3.167) PMID: [26279953](https://pubmed.ncbi.nlm.nih.gov/26279953/)
16. Thirumal PC, Nagarajan N. Utilization of data mining techniques for diagnosis of diabetes mellitus-a case study. *ARPN J Eng Appl Sci*. 2015; 10(1):8-13. [Link](#)
17. Thomas J, Princy RT. Human heart disease prediction system using data mining techniques. *Proceeding of the 2016 International Conference on Circuit, Power and Computing Technologies [ICCPCT]*. IEEE; 2016. p 1-5. DOI: [10.1109/ICCPCT.2016.7530265](https://doi.org/10.1109/ICCPCT.2016.7530265)
18. Verma D, Mishra N. Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. IEEE; 2017. p 533-38. DOI: [10.1109/ISSI.2017.8389229](https://doi.org/10.1109/ISSI.2017.8389229)
19. Ramasamy S, Nirmala K. Disease prediction in data mining using association rule mining and keyword-based clustering algorithms. *Int J Comput Appl*. 2017; 42:1-8. DOI: [10.1080/1206212X.2017.1396415](https://doi.org/10.1080/1206212X.2017.1396415)
20. Mahmoodi SA, Mirzaie K, Mahmoodi SM. Determining the effective factors in the incidence of gastric cancer by using data mining approach. *Payavard*. 2017; 11(3):332-41. [In Persian] DOI: [10.31557/APJCP.2019.20.9.2607](https://doi.org/10.31557/APJCP.2019.20.9.2607)
21. Ayatollahi H, Gholamhosseini L, Salehi M. Comparing the performance of data mining algorithms in predicting coronary artery diseases. *JHBMI*. 2018; 5(2):252-64. [In Persian] [Link](#)

22. Lee JH, Jeong SN, Choi SH. Predictive data mining for diagnosing periodontal disease: The Korea National Health and Nutrition Examination Surveys (KNHANES V and VI) from 2010 to 2015. *J Public Health Dent*. 2019; 79(1):44-52. DOI: [10.1111/jphd.12293](https://doi.org/10.1111/jphd.12293) PMID: [30468241](https://pubmed.ncbi.nlm.nih.gov/30468241/)
23. Tabatabaei SM, Saadatjoo F, Mirzaei M. The prediction model for cardiovascular disease using Yazd's health study data (YaHS) ta (YaHS). *JSSU*. 2019; 27(3):1346-60. [In Persian] DOI: [10.18502/ssu.v27i3.1188](https://doi.org/10.18502/ssu.v27i3.1188)
24. Sarkar BK, Sana SS. An e-healthcare system for disease prediction using hybrid data mining technique. *JM2*. 2019; 14(3):628-61. DOI: [10.1108/JM2-05-2018-0069](https://doi.org/10.1108/JM2-05-2018-0069)
25. Saleh B, Saedi A, Al-Aqbi A, Salman L. Analysis of weka data mining techniques for heart disease prediction system. *Int J Med Rev*. 2020; 7(1):15-24. DOI: [10.30491/ijmr.2020.221474.1078](https://doi.org/10.30491/ijmr.2020.221474.1078)
26. Deepika P, Sasikala S. Enhanced model for prediction and classification of cardiovascular disease using decision tree with particle swarm optimization. In2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE; 2020. p. 1068-72. DOI: [10.1109/ICECA49313.2020.9297398](https://doi.org/10.1109/ICECA49313.2020.9297398)
27. KL SK, Madhurya JA. Retracted: Coronary artery disease prediction using data mining techniques. In2020 3rd International Conference on Intelligent Sustainable Systems (ICISS). IEEE; 2020. p. 693-97. DOI: [10.1109/ICISS49785.2020.9316014](https://doi.org/10.1109/ICISS49785.2020.9316014)
28. Tasnim F, Habiba SU. A comparative study on heart disease prediction using data mining techniques and feature selection. In2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE; 2021. p. 338-41. DOI: [10.1109/ICREST51555.2021.9331158](https://doi.org/10.1109/ICREST51555.2021.9331158)
29. Mehra N, Mittal D. Healthcare data analysis using data mining techniques for disease prediction. *IJCSE*. 2021; 12(5):1224-37. DOI: [10.21817/INDJCSE/2021/v12i5/211205047](https://doi.org/10.21817/INDJCSE/2021/v12i5/211205047)
30. Fadnavis R, Dhore K, Gupta D, Waghmare J, Kosankar D. Heart disease prediction using data mining. *J Phys Conf Ser*. 2021; 1913(1):012099. DOI: [10.1088/1742-6596/1913/1/012099](https://doi.org/10.1088/1742-6596/1913/1/012099)
31. 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. [In Persian] [Link](#)
32. Al-Taie RR, Saleh BJ, Saedi AF, Salman LA. Analysis of WEKA data mining algorithms Bayes net, random forest, MLP and SMO for heart disease prediction system: A case study in Iraq. *Int J Electr Comput Eng*. 2021; 11(6):5229-39. DOI: [10.11591/ijece.v11i6.pp5229-5239](https://doi.org/10.11591/ijece.v11i6.pp5229-5239)
33. Sojoudi M, Ebrahimbabaie F. Decision tree-based use of data mining techniques to determine and predict the rare patients status: the case study: Rare Patients Registry System (SABNA). *J New Res Approaches Manag Account*. 2020; 4(33):53-65. [In Persian] [Link](#)
34. Deb A, Koli MS, Akter SB, Chowdhury AA. An outcome based analysis on heart disease prediction using machine learning algorithms and data mining approaches. In2022 IEEE World AI IoT Congress (AIIoT). IEEE; 2022. p. 1-7. DOI: [10.1109/AIIoT54504.2022.9817194](https://doi.org/10.1109/AIIoT54504.2022.9817194)
35. Bhatnagar AK, Madan P, Rana A, Sharma S, Sonawane S, Josphine CV. An efficient techniques for disease prediction from medical data using data mining and machine learning. In2022 5th International Conference on Contemporary Computing and Informatics (IC3I). IEEE; 2022. p. 839-44. DOI: [10.1109/IC3I56241.2022.10073286](https://doi.org/10.1109/IC3I56241.2022.10073286)
36. Gholami Z, Zare H. Predicting generalized anxiety disorder among female students using random forest approach. *Res Womens*. 2023; 14(1):35. [In Persian] [Link](#)
37. Sanaei F, Mousavi SA, Toloie Eshlaghy A, Rajabzadeh-Ghatari A. Designing a support system for predicting the survival of patients with melanoma based on data mining algorithms. *J Health Syst Res*. 2023; 19(2):160-5. [In Persian] DOI: [10.48305/jhsr.v19i2.1499](https://doi.org/10.48305/jhsr.v19i2.1499)
38. Nissar I, Mir WA, Shaikh TA, Areen T, Kashif M, Khiani S, et al. An intelligent healthcare system for automated diabetes diagnosis and prediction using machine learning. *Procedia Comput Sci*. 2024; 235:2476-85. DOI: [10.1016/j.procs.2024.04.233](https://doi.org/10.1016/j.procs.2024.04.233)
39. Hamidi HH, Moradi M. Analysis of disease severity and mortality prediction using machine learning during COVID-19. *Acta Psychologica*. 2025; 258:105136. DOI: [10.1016/j.actpsy.2025.105136](https://doi.org/10.1016/j.actpsy.2025.105136)
40. Akter S, Liu Z, Simoes EJ, Rao P. Using machine learning and electronic health record (EHR) data for the early prediction of Alzheimer's Disease and Related Dementias. *J Prev Alzheimers Dis*. 2025; 12(7):100169. DOI: [10.1016/j.tjpad.2025.100169](https://doi.org/10.1016/j.tjpad.2025.100169) PMID: [40246680](https://pubmed.ncbi.nlm.nih.gov/40246680/)
41. Sumon MSI, Islam MSB, Rahman MS, Hossain MSA, Khandakar A, Hasan A, et al. CardioTabNet: A novel hybrid transformer model for heart disease prediction using tabular medical data. *Health Inf Sci Syst*. 2025; 13(1):44. DOI: [10.1007/s13755-025-00361-7](https://doi.org/10.1007/s13755-025-00361-7) PMID: [40703220](https://pubmed.ncbi.nlm.nih.gov/40703220/)
42. Abualigah L, Alomari SA, Almomani MH, Zitar RA, Saleem K, Migdady H, et al. Artificial intelligence-driven translational medicine: A machine learning framework for



predicting disease outcomes and optimizing patient-centric care. *J Transl Med.* 2025; 23(1):302. DOI: [10.1186/s12967-025-06308-6](https://doi.org/10.1186/s12967-025-06308-6)

of things services in healthcare sector. *J Health Adm.* 2018; 21(73):106-17. [In Persian] [Link](#)

43. AlShehhi A, Alblooshi H, Fadul R, Tumzghi N, Tenaiji AA, Harbi MA, et al. Comparison of machine learning models for mucopolysaccharidosis early diagnosis using UAE medical records. *Sci Rep.* 2025; 15(1):28813. DOI: [10.1038/s41598-025-13879-3](https://doi.org/10.1038/s41598-025-13879-3) PMID: [40770294](https://pubmed.ncbi.nlm.nih.gov/40770294/)
44. Rehman MU, Naseem S, Butt AU, Mahmood T, Khan AR, Khan I, et al. Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Sci Rep.* 2025; 15(1):13361. DOI: [10.1038/s41598-025-96437-1](https://doi.org/10.1038/s41598-025-96437-1)
45. Young JA, Chang CW, Scales CW, Menon SV, Holy CE, Blackie CA. Machine learning methods using artificial intelligence deployed on electronic health record data for identification and referral of at-risk patients from primary care physicians to eye care specialists: Retrospective, case-controlled study. *JMIR AI.* 2024; 3(1):e48295. DOI: [10.2196/48295](https://doi.org/10.2196/48295) PMID: [38875582](https://pubmed.ncbi.nlm.nih.gov/38875582/)
46. Yasmeen R, Khan L, Choi A. Heart disease prediction using hybrid TabNet architecture with stacked ensemble learning. *Front Physiol.* 2025; 16:1665128. DOI: [10.3389/fphys.2025.1665128](https://doi.org/10.3389/fphys.2025.1665128) PMID: [41267789](https://pubmed.ncbi.nlm.nih.gov/41267789/)
47. He J, Wang X, Zhu P, Wang X, Zhang Y, Zhao J, et al. Identification and validation of an explainable early-stage chronic kidney disease prediction model: A multicenter retrospective study. *EClinicalMedicine.* 2025; 84:103286. DOI: [10.1016/j.eclinm.2025.103286](https://doi.org/10.1016/j.eclinm.2025.103286) PMID: [40567347](https://pubmed.ncbi.nlm.nih.gov/40567347/)
48. Bhasuran B, Schmolly K, Kapoor Y, Jayakumar NL, Doan R, Amin J, et al. Reducing diagnostic delays in Acute Hepatic Porphyrin using electronic health records data and machine learning: A multicenter development and validation study. *medRxiv.* 2023: 2023.08.30.23293130. DOI: [10.1101/2023.08.30.23293130](https://doi.org/10.1101/2023.08.30.23293130) PMID: [37693437](https://pubmed.ncbi.nlm.nih.gov/37693437/)
49. Zhang Y, Ghahramani N, Li R, Chinchilli VM, Ba DM. Prediction of acute and chronic kidney diseases during the post-covid-19 pandemic with machine learning models: Utilizing national electronic health records in the US. *EBioMedicine.* 2025; 115:105726. DOI: [10.1016/j.ebiom.2025.105726](https://doi.org/10.1016/j.ebiom.2025.105726) PMID: [40288236](https://pubmed.ncbi.nlm.nih.gov/40288236/)
50. Jiang ZZ, Jiang YF, Wang Y, Zhou Y, Peng RL, Ma CY, et al. Development of an explainable machine learning model for 3-year cardiovascular risk prediction in new-onset type 2 diabetes using the TyG index and ultrasound features. *BMC Med Inform Decis Mak.* 2025; 25(1):409. DOI: [10.1186/s12911-025-03247-6](https://doi.org/10.1186/s12911-025-03247-6) PMID: [41188894](https://pubmed.ncbi.nlm.nih.gov/41188894/)
51. Ronaghi MH, Hosseini F. Identifying and ranking internet