electronics

MDPI

Systematic Review

Health Risk Assessment Using Machine Learning: Systematic Review

Stanley Ebhohimhen Abhadiomhen [1,2], Emmanuel Onyekachukwu Nzeakor [1] and Kiemute Oyibo [1,*]

[1] Department of Electrical Engineering and Computer Science, York University, 4700 Keele Street, Toronto, ON M3J 1P3, Canada; stanley.abhadiomhen@unn.edu.ng (S.E.A.); emmanuelnzeakor00@gmail.com (E.O.N.)
[2] Department of Computer Science, University of Nigeria, Nsukka 400241, Nigeria
* Correspondence: kiemute.oyibo@yorku.ca

Abstract: According to the World Health Organization, chronic illnesses account for over 70% of deaths globally, underscoring the need for effective health risk assessment (HRA). While machine learning (ML) has shown potential in enhancing HRA, no systematic review has explored its application in general health risk assessments. Existing reviews typically focus on specific conditions. This paper reviews published articles that utilize ML for HRA, and it aims to identify the model development methods. A systematic review following Tranfield et al.'s three-stage approach was conducted, and it adhered to the PRISMA protocol. The literature was sourced from five databases, including PubMed. Of the included articles, 42% (11/26) addressed general health risks. Secondary data sources were most common (14/26, 53.85%), while primary data were used in eleven studies, with nine (81.81%) using data from a specific population. Random forest was the most popular algorithm, which was used in nine studies (34.62%). Notably, twelve studies implemented multiple algorithms, while seven studies incorporated model interpretability techniques. Although these studies have shown promise in addressing digital health inequities, more research is needed to include diverse sample populations, particularly from underserved communities, to enhance the generalizability of existing models. Furthermore, model interpretability should be prioritized to ensure transparent, trustworthy, and broadly applicable healthcare solutions.

Keywords: health risk assessment; machine learning; artificial intelligence; chronic diseases; health risk prediction

Citation: Abhadiomhen, S.E.; Nzeakor, E.O.; Oyibo, K. Health Risk Assessment Using Machine Learning: Systematic Review. *Electronics* **2024**, *13*, 4405. https://doi.org/10.3390/electronics13224405

Academic Editor: Luca Mesin

Received: 17 October 2024
Revised: 6 November 2024
Accepted: 9 November 2024
Published: 11 November 2024

1. Introduction

In today's fast-paced and interconnected world, individuals are increasingly exposed to various health risks, with chronic diseases such as diabetes, hypertension, and obesity on the rise globally [1]. In fact, more than 70% of deaths worldwide are attributed to chronic illnesses [2], which are largely due to sedentary lifestyles and poor dietary habits [3]. According to the WHO global report [4], diabetes alone affects 422 million people globally, and its prevalence shows no signs of slowing. Similarly, hypertension is becoming more prevalent across populations, further contributing to the global health burden [5]. As indicated by Svendsen et al. [6], untimely and inadequate health risk assessment can lead to delayed diagnosis, ineffective treatment, and increased morbidity and mortality rates, which underscores the need for proactive strategies tailored to individual needs. Traditional methods often lack the sensitivity and specificity necessary for accurate risk identification [7]. Moreover, healthcare disparities and structural inequities can worsen access to healthcare resources, exacerbating health outcomes, particularly in underserved communities [8]. Therefore, automated and tailored approaches that address health inequities and improve the accuracy and timeliness of health risk assessments should be developed.

Machine learning (ML), in this regard, has demonstrated potential in enhancing health risk assessment (HRA) [9], thus enabling systems to learn from data and past experiences [10]. Consequently, numerous models and tools have emerged, which leverage widely used ML algorithms such as support vector machines (SVM) [11] and neural networks [12]. These methods generally follow a structured approach, with the core process being the training of algorithms on data to classify individuals or populations into different risk categories, as illustrated in Figure 1. Common data collection methods include secondary sources [13] and primary techniques such as survey data [14]. Given that these data sources can be noisy, the training process often incorporates several crucial steps: data preprocessing to refine and prepare the data, feature selection to identify the most relevant variables, and model tuning to enhance performance [15]. Additionally, techniques such as stratified sampling [16] and data augmentation [17] are commonly used to ensure that the dataset accurately represents diverse populations, which is crucial for addressing fairness and bias concerns. However, despite the growing application of machine learning in HRA, no scoping or systematic review has comprehensively explored its use across the broader field. Most previous reviews have typically focused on specific health risks or narrow aspects of HRA, without fully examining the general application of ML in health risk assessment (Refer to Table 1 for details of the related works). For example, Mishra et al. [18] conducted a systematic review, which focused exclusively on studies of pancreatic cancer, utilizing data from electronic health records. Usman et al. [19] carried out a systematic review that focused on machine learning-based models for predicting the progression of diabetic retinopathy. Likewise, Singh et al. [20] performed a scoping review centred on the application of AI for cardiovascular disease risk assessment within a personalized framework. Although Abdulazeemi et al. [21] conducted a systematic review of clinical health conditions predicted by machine learning models, their review predominantly focused on diagnostic models using real-world primary health care data. Therefore, the rationale for this current systematic review stems from the lack of existing reviews synthesizing studies on various health risk assessments using machine learning techniques.

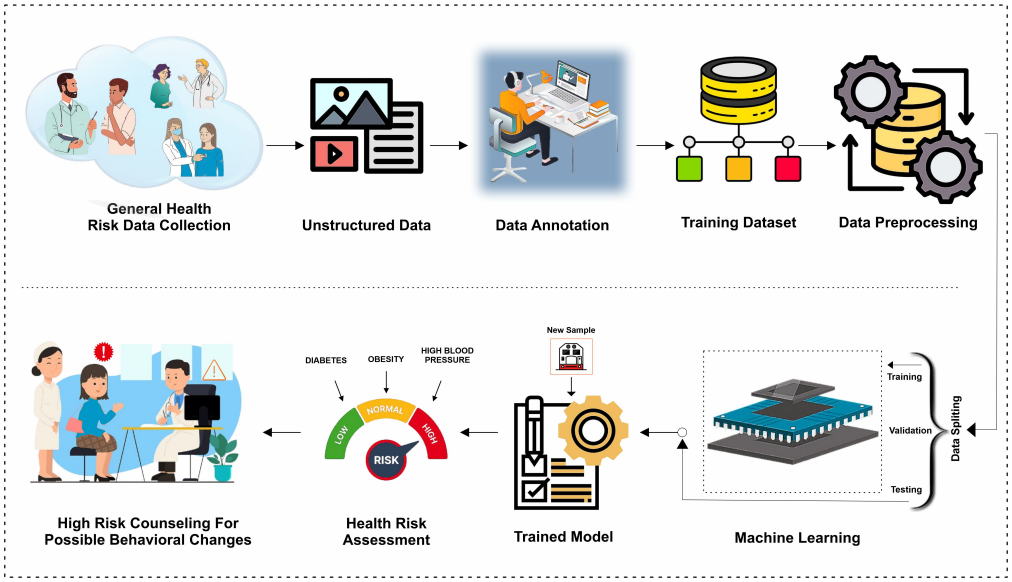

Figure 1. A typical machine learning workflow for health risk assessment, encompassing data collection, data annotation, data pre-processing to clean and prepare the data, and model training.

The main objective of this systematic review is to identify and synthesize the research findings from published articles on HRAs that utilize machine learning techniques and tools. Specifically, this review aims to explore the methods used for data collection, the machine learning algorithms used for predicting health risks, the validation methods, and the explainability of the models. Accordingly, the key research questions this review seeks to

answer are the following: (1) What data collection methods are employed in studies related to health risk assessment using machine learning? (2) What machine learning algorithms are used for health risk prediction, and how do the processes of model validation and explainability vary across different studies? (3) Are the data used for training the predictive models sufficient for and representative of diverse demographic groups, and are there potential bias or fairness concerns? (4) What unresolved issues and research gaps exist in this area?

Table 1. Summary of related works in health risk assessment using machine learning.

Reference	Objective	Key Conclusions	Deficiencies	Relevance to Current Work
Mishra et al. [18] (2022)	To systematically review studies on pancreatic cancer prediction models using electronic health records.	Demonstrated that electronic health records (EHR) provide valuable insights for pancreatic cancer prediction.	Focused only on pancreatic cancer, limiting generalizability to other diseases or risk assessments.	Highlights the use of EHR data for disease prediction, a data source relevant to the broader health risk assessment in this current review.
Usman et al. [19] (2023)	To review machine learning-based models for predicting diabetic retinopathy progression.	Showed that machine learning models can be effective in predicting diabetic retinopathy progression over time.	Limited to diabetic retinopathy, without examining applications for other chronic conditions or risk factors.	Underlines the potential of ML models for disease progression prediction, a concept valuable for general HRA models.
Singh et al. [20] (2024)	To explore AI applications in cardiovascular disease risk assessment within a personalized framework.	Found that AI can enable personalized risk assessments, improving cardiovascular disease prediction and prevention.	Narrowly focused on cardiovascular disease, lacking applications for other major health risks or generalized HRA.	Demonstrates the utility of AI in personalizing health risk assessments, supporting the application of ML to other health risks.
Abdulazeemi et al. [21] (2023)	To review ML models for predicting clinical health conditions, using primary health care data.	Showed that primary healthcare data can enhance the diagnostic accuracy of ML models in predicting various conditions.	Focused on diagnostic models rather than general health risk assessments; limited to primary care data.	Provides insights on the use of real-world data in diagnostic ML models, informing the synthesis of health risk models.
Fleuren et al. [22] (2020)	Systematic review and meta-analysis on ML models for early prediction of sepsis.	ML models showed strong potential for early sepsis prediction with AUROC ranging from 0.68 to 0.99 in ICU settings.	Study heterogeneity and varying sepsis definitions hindered pooling results. Clinical implementation was mixed.	Focuses on real-time prediction, while the current work aims to cover diverse health risk assessments using various ML models.
Xiong et al. [23] (2023)	Systematic review and meta-analysis on ML-based prediction models for asthma exacerbations.	11 studies identified with 23 prediction models. Logistic regression, boosting, and random forest were the most used ML methods. Overall AUROC of 0.80, boosting achieved the best performance (AUROC 0.84).	Limited to asthma exacerbations; does not address other health risks.	Offers important perspectives on the performance of ML models in predicting asthma risk.

2. Methodology

The systematic review was conducted following the three-step framework established by Tranfield et al. [24]: planning, searching and screening articles, and analyzing the results. In the first phase, the first (SE) and third (KO) authors defined the scope and objectives of the review and developed a comprehensive search protocol, including the search string and criteria for inclusion and exclusion, which guided the retrieval of articles from various databases. In the second phase, the search, screening, and analysis of articles were carried out. SE and the second (EO) author systematically searched six databases, including Google Scholar, using the predefined search string. Each author independently searched three databases according to the agreed protocol. The articles were then screened, selected, and analyzed based on key factors pertinent to the research focus. The final phase involved synthesizing and reporting the findings from the selected articles, identifying research gaps, and making recommendations for future research. The PRISMA 2020 checklist [25] was used to guide the writing of the review, providing essential guidelines for summarizing findings and highlighting gaps in the current literature.

2.1. Eligibility Criteria

The eligibility criteria for the systematic review were defined as follows: Articles were included if they were written in English, peer-reviewed, published in journals or conferences, and focused on health risk assessment (specifically for chronic disease conditions) using machine learning techniques. Articles were excluded if they did not meet any of these inclusion criteria. Review articles and theoretical studies without the practical application of machine learning models on actual data were also excluded. These criteria were aimed at ensuring the systematic review focused on high-quality and relevant research articles that specifically applied machine learning in health risk assessment.

Five major databases were searched to ensure comprehensive coverage of relevant articles: ACM Digital Library, PubMed, Scopus, IEEE Xplore, and WOS.These databases were selected for their comprehensive coverage of high-quality, peer-reviewed literature relevant to machine learning and health risk assessment, ensuring a thorough representation of current advancements in the field. However, we did not include preprint archives or grey literature in our search strategy because our primary focus was to synthesize findings from peer-reviewed studies that adhere to established scientific standards. The keywords and Boolean search strings used for querying the databases included the following: ("health risk assessment" OR "health risk prediction" OR "health risk management") AND ("machine learning" OR ML OR "AI" OR "artificial intelligence"). It is important to note that in addition to "health risk assessment", the terms "prediction" and "management" were included because they are often used interchangeably in the literature, ensuring that studies with different terminologies but similar concepts were captured. Similarly, "artificial intelligence" (AI) was incorporated alongside "machine learning" (ML) since both terms are also frequently used interchangeably in many contexts. Furthermore, hand searches of Google Scholar and cross-referencing of retrieved articles were conducted to identify more relevant articles that met the inclusion criteria.

2.2. Study Selection and Data Collection Process

The selection process began with a title and abstract screening (TA) of 570 unique articles retrieved from the database search. The first (SE) and second (EO) authors independently conducted this initial screening to evaluate whether the title and abstract included the keywords related to health risk assessment and machine learning, as outlined in Section 2.1. Following the TA screening, 89 articles advanced to the full-text review (FTR) stage. During this phase, SE and EO independently and simultaneously evaluated each article according to the inclusion criteria shown in Table 2. Articles were first evaluated based on general inclusion criteria: (1) Written in English; (2) Peer-reviewed; (3) Published in a journal or conference. The articles meeting these criteria were then further assessed to ensure they specifically addressed health risk assessment (HRA) using machine learning

techniques, which was the main inclusion criterion. Articles not directly relevant to the objectives of the systematic review were excluded, including those unrelated to chronic disease-based HRAs, even if they contain the search keywords. Ultimately, 26 articles were included in the final analysis. The first and second authors collaboratively reviewed the included articles, extracting relevant data into a Google Spreadsheet, where it was processed, charted, and tabulated. Refer to Table 3 for details of the key data items (e.g., author identification, publication year, study objectives, population, machine learning algorithms, and key findings) extracted and their definitions. Each data item was carefully selected by the authors to align with the objectives of the systematic review and best practices established in the literature [26].

Table 2. Inclusion and exclusion criteria for including and excluding articles in the systematic review.

Inclusion Criterion	Exclusion Criterion
Articles written in English.	Articles written in languages other than English.
Articles published in peer-reviewed journals or conference papers.	Articles that are book chapters, magazine articles, or webpages.
Articles that focus on health risk assessment and specific health risks using machine learning techniques. This includes articles that concentrate on specific health conditions related to health risks.	Articles that discuss theoretical frameworks without practical applications in health risk assessment or focus on unrelated health topics that do not involve machine learning applications.

Table 3. Data items and descriptions.

Data Item	Description
Identification (Author, First Author Country, Publication Year)	The authors' names, country of affiliation, and year of publication.
Study Focus	The health condition focus. Studies focusing on multiple disease types are classified as General Health Risks.
Data Source	The method used for collecting data, whether primary data sources or secondary.
Population	The demography and size of the population used in the study.
Predictors	The data features collected.
Machine Learning Algorithm(s)	The machine learning algorithms applied for predictive analysis.
Evaluation Metrics	The metrics used to evaluate model performance.
Model Validation Technique	The validation methods used to evaluate the model for generalizability.
Model Explainability	Whether the explainability of the model was considered or not.
Summary of Findings	The main outcomes of the study.

2.3. Data Synthesis Method

Through the collective efforts of the three authors, the retrieved articles were categorized using several key metrics to provide a nuanced understanding of study characteristics and methodological variations. First, the articles were grouped by geographical and chronological factors, such as region and year of publication. Further classification was based on study focus and data collection methods to capture trends in data sourcing techniques. For technical analysis, the articles were categorized according to the predictive model development process, including the machine learning algorithms used and the evaluation metrics reported. Additionally, articles were grouped based on model explainability considerations. We employed a narrative synthesis approach for data analysis, integrating and interpreting findings through both thematic and descriptive analysis rather than relying on quantitative aggregation. This method was informed by pattern recognition and thematic analysis theories [27], emphasizing the importance of capturing the context and nuances of each study. Descriptive synthesis was specifically used to outline general trends, such as geographical distribution, while thematic synthesis focused on more technical details.

The findings of this systematic review are presented through tables, summaries, and charts generated using MATLAB R2021a.

2.4. Risk of Bias and Certainty Assessment

Risk of Bias Assessment: The risk of bias (ROB) of the implemented prediction models in the included studies was assessed using the prediction model risk of bias assessment tool (PROBAST) [28]. The PROBAST evaluation model is based on four key domains: participants, predictors, outcome, and analysis. For each of these domains, judgments on the ROB were made by categorizing each study as having either a "low", "high", or "unclear" risk of bias. This assessment was performed independently and collaboratively by the first (SE) and second authors (EO). Discrepancies in judgments were resolved through discussion to ensure consistent evaluation across studies. The results of the risk of bias assessments are reported in a colour-coded Table.

Certainty Assessment: We employed the grading of recommendations, assessment, development and evaluations (GRADE) tool [29] to evaluate the certainty of evidence for each included study. The assessment considered the following factors: (1) Sample Size—the number of participants or samples analyzed; (2) Methodological Quality—included risk of bias (obtained via PROBAST), study limitations, and robustness of analysis; (3) Data Quality—encompassed completeness, accuracy, and reliability of data; (4) Analysis Methods—involved appropriateness and robustness of statistical techniques used; (5) Results Interpretation—focused on clarity and transparency of conclusions. Based on these factors, studies were classified into one of three categories: High Certainty for studies with well-designed methodologies, large sample sizes, robust analysis, and clear conclusions; Moderate Certainty for studies with some methodological limitations, moderate sample sizes, or less robust analysis but still provided valuable insights; and Low Certainty for studies with significant methodological flaws, small or unspecified sample sizes, or limited descriptions of the analytical process, making conclusions less reliable. Although the inclusion of a summary of findings (SoF) table typically enhances the comprehension and understanding of evidence [30], this systematic review predominantly included prediction modelling studies rather than traditional randomized controlled trials. Thus, an SoF table was not included in this review, as the adapted GRADE criteria were more suitable for evaluating the certainty of evidence in the context of machine learning-based health risk prediction studies [31].

3. Results

This section presents the results of the systematic review and analysis, including tables and charts that directly address the research questions.

3.1. Selection of Sources of Evidence

The screening and selection of eligible articles for the systematic review and analysis followed the PRISMA flowchart presented in Figure 2. Initially, 1015 articles were retrieved from five databases. After removing 445 duplicate records (43.84% of the retrieved articles), 570 unique articles were passed to the TA screening phase. The two researchers (SE and EO) who conducted the TA screening had six disagreements. Specifically, there were four instances where SE marked a paper as accepted while EO marked it as rejected and two instances where EO marked a paper as accepted while SE marked it as rejected. As shown in Table 4, there were 564 agreements, with both researchers accepting 83 papers and rejecting 481. As a result, the agreement rate was 98.95%, accompanied by a Cohen's Kappa score of 0.96. Following the resolution of all six disagreements, the papers were ultimately marked as accepted after the TA screening.

Table 4. Agreement between researchers during screening.

	Title/Abstract Screening	Full-Paper Screening
Agreements (Accepted)	83	21
Agreements (Rejected)	481	64
Disagreements (Only SE Accepted)	4	2
Disagreements (Only EO Accepted)	2	2
Disagreements Resolved (Accepted)	6	1
Disagreements Resolved (Rejected)	0	3
% Agreement	98.95	95.51
Cohen's Kappa	0.96	0.88

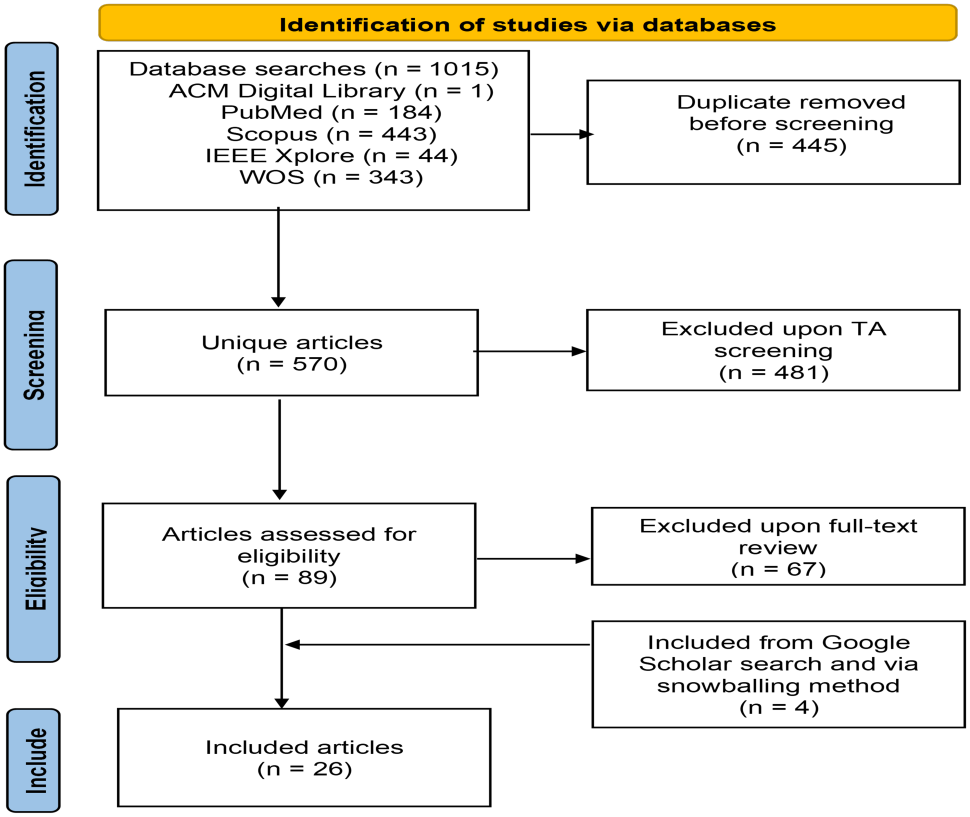

Figure 2. PRISMA flowchart for the screening and inclusion of articles in the systematic review. WOS: Web of Science. TA: Title, abstract screening.

Consequently, as detailed in Figure 2, a total of 481 (84.39%) out of 570 accessed papers were excluded as unrelated to the topic based on the TA screening, leaving 89 papers for full-text review. There were four disagreements between the two researchers (SE and EO) who conducted the full-text review, with two cases each where EO marked a paper as accepted while SE marked it as rejected and vice versa. As also shown in Table 4, there were 85 agreements, with both researchers accepting 21 papers and rejecting 64. This resulted in an agreement rate of 95.51% and a Cohen's Kappa score of 0.88. After deliberation, all four disagreements were resolved, with three articles being rejected by both researchers and each researcher accepting one paper. Hence, 67 articles (representing 75.28% of the 89 eligible articles for full-text screening) were excluded after the FTR for not meeting the inclusion criteria, leaving 22 articles for data analysis. Lastly, four articles were identified and added through a Google Scholar search and the snowballing method, bringing the total to 26 articles for the systematic review and final analysis.

3.2. Results of Synthesis

This section presents the summarized and synthesized results of the analysis of the data extracted from the included articles using charts and tables. The analysis is organized into the following subsections: geographical and chronological characteristics, study objectives, data collection methods, and the model development process, which includes machine learning algorithms, evaluation metrics, performance and model validation. Finally, the section on explainability considerations offers further insights into the included articles, as categorized in Section 2.3. Refer to Table S1 in the Supplementary Material for the broad characteristics of the included articles.

3.2.1. Geographical and Chronological Characteristics

Based on the country of the first author, the articles included in this systematic review span four continents—Asia, North America, Europe, and Australia—covering ten different countries: India, China, South Korea, Taiwan, Saudi Arabia, Bangladesh, USA, Greece, Russia, and Australia. As detailed in Table 5, a significant proportion of the included articles, seventeen (65.38%) out of the 26, are from Asia. These include seven articles (26.92%) from India [32–38], four articles (15.38%) from China [39–42], three articles (11.54%) from South Korea [43–45], while Taiwan [46], Saudi Arabia [47], and Bangladesh [48] each contribute one article (3.85%) respectively. North America contribute four articles (15.38%), all from the USA [49–52]. Australia accounts for a total of three (11.54%) articles [53–55]. Europe, represented by Greece [56] and Russia [57], each contributing one article (3.85%), makes up a total of 7.69% of the included articles. As shown in Figure 3 (which depicts the publication trend), the range of publication years spans from 2017 to 2024. Only one (3.85%) of the 26 included articles was published in 2017 [38], marking it as the earliest year in our dataset. The publication volume was evenly distributed between 2018 and 2019, with two articles (7.69%) published in 2018 [45,52] and another two published in 2019 [33,53]. Additionally, three (11.54%) of the included articles were published in 2020 [41,47,49] and one (3.85%) article in 2021 [39], placing 2021 and 2017 as the years with the lowest publication volume in our dataset. As can also be observed in the figure, the majority of the articles, 7/26 (26.92%), were published in 2023 [37,40,42,46,51,55,57]. This is followed by five articles (19.23%) published in 2022 [32,36,44,54,56] and another five articles (19.23%) published in 2024 [34,35,43,48,50].

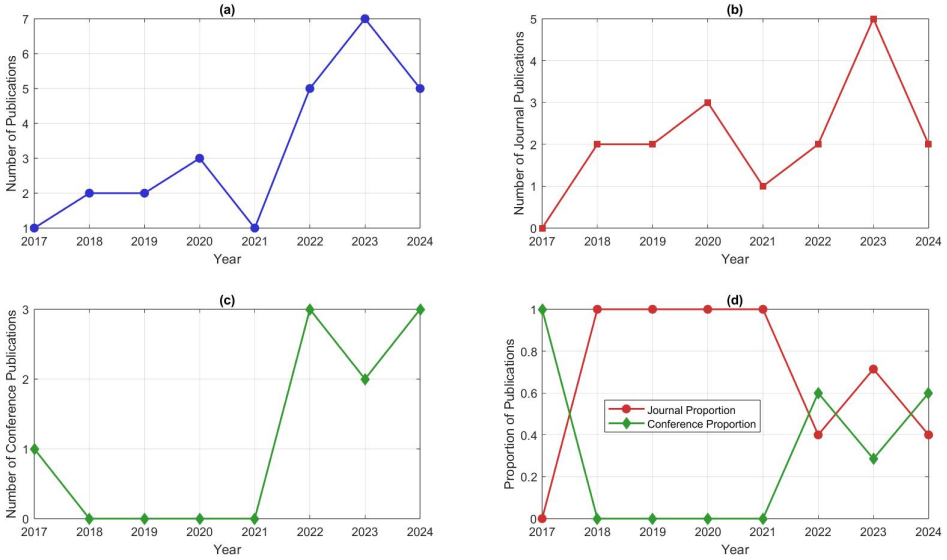

Figure 3. Publication distribution by year: (**a**) Total publications; (**b**) Journal publications; (**c**) Conference publications; (**d**) The proportion of journal and conference publications obtained by dividing the respective counts by the total counts.

Table 5. Geographical distribution of publications based on country and continent.

Continent	Country	Country (%)	Continent (%)
Asia	India	7 (26.92%)	17 (65.38%)
	China	4 (15.38%)	
	South Korea	3 (11.54%)	
	Taiwan	1 (3.85%)	
	Saudi Arabia	1 (3.85%)	
	Bangladesh	1 (3.85%)	
North America	USA	4 (15.38%)	4 (15.38%)
Australia	Australia	3 (11.54%)	3 (11.54%)
Europe	Greece	1 (3.85%)	2 (7.69%)
	Russia	1 (3.85%)	

Figure 3d displays the proportion of journal and conference publications. Notably, in 2022, 60% of the articles were published as conference papers and 40% as journal articles. In contrast, 2023 saw a moderate shift, with four (66.70%) of the six publications being journal articles and only two (33.33%) being conference papers. Table 6 shows the distribution of articles based on publication type. Of the 26 included papers, nine (34.62%) were published in conference venues [32,34–36,38,48,51,55,56], with the other 17 (65.38%) appearing in journals. Specifically, most of the conference publications, five out of nine (55.60%), are from India [32,34–36,38], with one each from Bangladesh [48], Greece [56], USA [51] and Australia [55]. As a result, six (66.70%) of the conference publications included in our systematic review originated from Asia.

Table 6. Distribution of studies in various categories based on publication type, data source, model validation and explainability considerations. An asterisk (*) after the primary data source indicates that the study relatively considered diverse populations.

	Publication Type		Data Source		Model Validation		Explainability	
Reference	**Conference**	**Journal**	**Primary**	**Secondary**	**Specified**	**Not Specified**	**Yes**	**No**
Yadav [32]	✓			✓		✓		✓
Singh [33]		✓	✓		✓			✓
Khan [53]		✓		✓	✓			✓
Ashwath [34]	✓			✓	✓			✓
Chen [39]		✓				✓		✓
Li [40]		✓	✓		✓		✓	
Thakkar [35]	✓			✓		✓		✓
Shifa [48]	✓		✓		✓		✓	
Park [43]		✓	✓*		✓		✓	
Shomorony [49]		✓	✓*		✓		✓	
Sahoo [50]		✓	✓		✓			✓
Diamantoulaki [56]	✓			✓		✓		✓
Rashid [44]		✓		✓	✓			✓
Parvataneni [51]	✓			✓	✓			✓
Chung [45]		✓	✓			✓		✓
Pawar [36]	✓			✓	✓			✓
Hossain [47]		✓	✓			✓		✓
Islam [54]		✓		✓		✓		✓
Bayati [52]		✓		✓	✓			✓
Kurysheva [57]		✓	✓		✓			✓
Salian [37]		✓		✓		✓		✓
GE [41]		✓		✓		✓		✓
Anbu[38]	✓			✓	✓		✓	
Lin [42]		✓	✓		✓		✓	
Yang [46]		✓	✓			✓		✓
Liu [55]	✓			✓	✓		✓	
Count	9	17	11	14	16	10	7	19

3.2.2. Study Objectives and Data Collection Methods

Study Objectives: Across the included studies, the focus was either on evaluating general health risks (encompassing multiple diseases) or on specific health issues. In total, nine different disease conditions were targeted in the studies, and they focused on specific health issues: asthma, cardiovascular health risk, diabetes, fetal health risk, glaucomatous optic neuropathy, human respiratory tract health, hypertension, mental health risk, and obesity. Table S1 in the Supplementary Material captures the study focus of the included articles. Twelve (46.15%) out of 26 included studies focused on general health risks [32,33,37,39,41,43–45,49,52,55,56], while the remaining fourteen (53.85%) focused on specific health conditions. Asthma, for instance, received particular attention in one study (3.85%) [40], which utilized a novel affinity graph-enhanced classifier based on data from patients at the Affiliated Shuguang Hospital of Shanghai Traditional Chinese Medicine University. Similarly, one study (3.85%) explored cardiovascular health risks [34], while two studies (7.69%) concentrated on diabetes health risks [47,53]. Fetal health risk ($n = 1$) [51] and maternal health risks ($n = 3$) [35,36,48] collectively accounted for approximately 15.38% of the study focus in the included articles. Other health conditions, such as glaucomatous optic neuropathy [57], mental health risks [50], human respiratory tract [54], hypertension [46], women's health risk [38] and obesity [42], were each explored in a single study, accounting for roughly 4% each of the total.

Data Collection Methods: The data collection methods varied among the studies. As shown in Table 6, most of the included studies, 14/26 (53.84%), relied on secondary data sources [32,34–38,41,44,51–56]. These include various public repositories like UCI [32,35–37], Kaggle [56], the National Heart, Lung, and Blood Institute [34], the University of California and the University of Porto [51], computational fluid dynamics data [54], Beth Israel deaconess medical centre [41] and publicly available databases of electronic health records collected from intensive care units [55], as well as private data from organizations such as the Melinda Gates Foundation [38], Healthone Labs [52], and private healthcare funds based in Australia [53]. In contrast, eleven of the studies (42.30%) used the primary data collection method [33,40,42,43,45–50,57]. The primary data sources mainly involved clinical data collected from hospitals, self-assessed health data, and genomic sequencing. In particular, three out of the eleven studies (27.30%) collected data via surveys [33,45,50], one study (9.10%) utilized genomic techniques such as whole genome sequencing and microbiome sequencing [49], while the rest (7/11, 63.64%) gathered clinical data directly from participants or health facilities. It is worth mentioning that only two studies (18.18%) from the above set ($n = 11$), based on primary data, relatively considered diverse populations [43,49], while the remaining nine (81.82%) focused on specific populations. A small portion of studies (1/26, 3.84%) [39], however, did not specify their data source. Furthermore, Table 7 shows the categorization of the included articles based on data sizes. As can be seen from the records, only eleven out of the 26 studies (42.31%) utilized datasets with sizes above 1000. In contrast, six of the studies (23.07%) employed datasets below 1000, and nine (34.62%) did not specify their dataset sizes.

Table 7. Sizes of sample employed in the included studies.

Sample Size	Papers	Count
Above 1000	[34,35,38,41–43,48,49,51,53,55]	11 (42.31%)
Below 1000	[33,40,46,47,50,57]	6 (23.07%)
Not Specified	[32,36,37,39,44,45,52,54,56]	9 (34.62%)

3.2.3. Machine Learning Algorithms, Evaluation Metrics, Performance and Model Validation

Machine Learning Algorithms: A total of nine different machine learning algorithms were employed to build models in the included studies: Support Vector Machine (SVM), Support Vector Regression, Random Forest (RF), K-Nearest Neighbour (KNN), Deep Neu-

ral Network (DNN), Neural Network (NN), Decision Trees (DT), Naive Bayes (NB), Linear Regression (LR), and Logistic Regression (LGR). For clarity, ensemble methods other than RF were categorized under "Ensemble" (EN). As detailed in Table 8, SVM, RF, DT, and LGR were the most frequently employed algorithms. SVM was utilized in five studies (19.23%) [33,40,42,44,48], RF appeared in nine articles (34.62%) [36,38,40,42,44,47,48,51,54], DT was featured in seven studies (26.92%) [32,36,37,39,42,47,48], while LGR appeared in eight articles (30.77%) [37,38,40,42–44,47,53], making RT the most common algorithms used in the included studies. KNN and DNN were also commonly used, with four studies (15.38%) employing each of KNN [37,42,44,54] and DNN [44,45,55,56] (with one study [44] implementing the two and five other algorithms. NN was utilized in three studies (11.54%) [34,38,42,44]. Additionally, ten studies (38.64%) [34–38,40,42,43,48,51], used ensemble learning techniques, incorporating multiple algorithms to enhance performance. Notably, twelve of the included studies (46.15%) employed more than one algorithm, whereas the other fourteen (53.85%) focused on a single algorithm. Moreover, seven studies (26.92%) developed novel machine (NM) learning algorithms [40,41,46,49,50,52,57]. For instance, Li et al. [40] introduced AGEC (an Affinity Graph Enhanced Classifier) to capture the inherent complexity of asthma-related patterns. Shomorony et al. [49] employed a combination of unsupervised machine learning techniques to identify multimodal biomarker signatures associated with health and disease risks. Likewise, Sahoo et al. [50] applied a model-agnostic advanced machine learning/artificial intelligence technique, glmboost, as an effective alternative to the traditional linear regression model.

Table 8. Popular machine learning algorithms employed in the included studies. NM: Novel Model, SVM: Support Vector Machine, SVR: Support Vector Regression, RF: Random Forest, KNN: K Nearest Neighbour, DNN: Deep Neural Network, NN: Neural Network, DT: Decision Trees, NB: Naive Bayes, EN: Ensemble, LR: Linear Regression, LGR: Logistics Regression.

Reference	NM	SVM	SVR	RF	KNN	DNN	NN	DT	NB	EN	LR	LGR	Total
Yadav [32]								✓					1
Singh [33]		✓											1
Khan [53]												✓	1
Ashwath [34]										✓			1
Chen [39]								✓					1
Li [40]	✓	✓	✓	✓						✓		✓	6
Thakkar [35]										✓			1
Shifa [48]		✓		✓				✓	✓	✓			5
Park [43]										✓		✓	2
Shomorony [49]	✓												1
Sahoo [50]	✓										✓		2
Diamantoulaki [56]						✓							1
Rashid [44]		✓		✓	✓	✓	✓		✓			✓	7
Parvataneni [51]				✓						✓			2
Chung [45]						✓							1
Pawar [36]				✓				✓	✓	✓			4
Hossain [47]				✓				✓	✓			✓	4
Islam [54]				✓	✓								2
Bayati [52]	✓												1
Kurysheva [57]	✓												1
Salian [37]					✓			✓		✓		✓	4
GE [41]	✓												1
Anbu [38]				✓			✓			✓		✓	4
Lin [42]		✓		✓	✓		✓	✓		✓		✓	7
Yang [46]	✓												1
Liu [55]						✓							1
Count	7	5	1	9	4	4	3	7	4	10	1	8	

Evaluation Metrics: The common evaluation metrics reported across the 26 studies include accuracy, precision, recall, RMSE (root mean squared error), and AUC (area under the curve). Among these, accuracy emerged as the most frequently used measure, appearing in 18 of the 26 included articles (69.23%), with the exception of eight studies [41,42,45,49,50,52,54,55]. Precision, recall, and an F1-score that provides a more subtle evaluation of model effectiveness (addressing both false positives and false negatives), were also notably utilized; six articles employed them collectively [37,41,47,51,55,56]. The AUC was used in seven studies (26.92%) [33,36,40,42,55–57] based on its relevance for evaluating binary classification tasks and the trade-off between true and false positives. RMSE was reported in three articles (11.54%) [45,50,54], with one of the studies [54] utilizing RMSE along with mean percentage error and mean absolute percentage error. Specialized metrics, such as hamming loss, ranking loss, and coverage, were utilized in one study [41] that focused on developing a multi-label neural network model, as these metrics are commonly used to evaluate the performance of multi-label models. Additionally, eleven out of the 26 included articles (42.31%) employed multiple metrics for evaluation [33,36,37,40,41,43,47,51,54–56]. Common combinations include accuracy with precision, recall, and F1-score [37,47,51], and the integration of AUC with accuracy [33,36,40,43].

Model Performance: Since the studies were conducted in different contexts, with varying standards and datasets, a direct comparison may not be appropriate, as such differences could introduce bias and potentially lead to inaccurate conclusions. Therefore, in presenting a comparative analysis of the studies based on implemented machine learning algorithms and their performance metrics, we focused on those studies ($n = 12$, 46.15%) that implemented multiple algorithms [36–38,40,42–44,47,48,50,51,54] and extracted the best-performing model, as reported in Table 9. Based on the analysis of the studies, RF consistently emerged as the best-performing algorithm, being identified as the top choice in five (41.67%) of the studies [36,47,48,51,54]. Other ensemble methods, such as XGBoost, were the second most frequently reported best-performing algorithms, accounting for 25% ($n = 3$) of the studies [38,42,48]. NM, LGR, and DT had fewer occurrences as top performers. Overall, the highest accuracy in the included studies reached approximately 100%. AUC also exhibited significant variability, ranging from 55.85% to 99.80%. Model Validation: Several validation strategies were reported across the included studies. *k*-fold cross-validation appeared as the most common technique, with several studies specifying different values for *k*. Specifically, 10-fold cross-validation was used in four studies (15.38%) [43,44,50,53], 5-fold cross-validation in two studies (7.69%) [42,52], and 2-fold in one study (3.85%) [40]. Although three more articles reported the utilization of *k*-fold cross-validation [32,36,38], two [32,38] did not specify the exact value for *k*, with one study [36] experimenting with values ranging from 0 to 10. Holdout validation was used in four studies (15.38%) [33,48,51,55]. Procrustes cross-validation (commonly used in the context of statistical shape analysis) [57] and external validation [49] were reported in one article (3.85%), respectively. The external validation involved testing on an independent dataset. Nonetheless, ten articles (38.46%) [32,35,37,39,41,45–47,54,56] did not specify their model validation technique, as shown in Table 6.

Table 9. Performance comparison of different machine learning algorithms across twelve studies with multiple implementations. NM: Novel Model, SVM: Support Vector Machine, SVR: Support Vector Regression, RF: Random Forest, KNN: K Nearest Neighbour, DNN: Deep Neural Network, NN: Neural Network, DT: Decision Trees, NB: Naive Bayes, EN: Ensemble, LR: Linear Regression, LGR: Logistics Regression, AUC: Area Under the Curve, RMSE: Root Mean Squared Error.

Reference	Algorithms Reported	Metric	Performance by Metric	Best-Performing Algorithm
Li [40]	NM, SVM, SVR, RF, EN, LGR	Accuracy	**NM (72.50%)** SVM (69.80%) SVR (64.01%) RF (54.21%) BestEN (68.40%) LGR (59.24%)	NM
Shifa [48]	SVM, RF, DT, NB, EN	Accuracy	SVM (96.40%) RF (**97.30%**) DT (94.60%) NB (96.10%) BestEN (97.30%)	RF, EN (XGBoost)
Park [43]	EN, LGR	AUC	BestEN (88.50%) BestLGR (**89.60%**)	LGR
Sahoo [50]	NM, LR	RMSE	-	NM
Rashid [44]	SVM, RF, KNN, DNN, NN, NB, LGR	Accuracy	**DNN (99%)**	DNN
Parvataneni [51]	RF, EN	Accuracy	**RF (93.30%)** **RF (70.22%)**	RF
Pawar [36]	RF, DT, NB, EN	Accuracy	DT (53.25%) NB (56.23%) BestEN (67.65%)	RF
Hossain [47]	RF, DT, NB, LGR	Accuracy	**RF (97.40%)** DT (96.50%) NB (87.40%) LGR (92.41%)	RF
Islam [54]	RF, KNN	RMSE	**RF (5.477)** KNN (7.754)	RF
Salian [37]	KNN, DT, EN, LGR	Accuracy	KNN (79.90%) **DT (94.08%)** LGR (60.52%) EN (90.78%)	DT
Anbu [38]	RF, NN, EN, LGR	Accuracy	RF (81.14%) NN (79.26%) **EN (85.68%)** LGR (76.30%)	EN
Lin [42]	SVM, RF, KNN, NN, DT, EN, LGR	Accuracy	SVM (81.00%) RF (80.00%) KNN (78.00%) NN (81.00%) DT (79.00%) EN (81.00%) LGR (81.00%)	SVM, NN, EN, LGR

3.2.4. Explainability Considerations

In this section, we reveal the extent to which model explainability was considered across the various articles. As can be seen in Table 6, only a minority of studies employed specific techniques to enhance model transparency. Specifically, seven out of 26 studies (26.92%) incorporated some form of model interpretability technique [38,40,42,43,48,49,55], while the remaining nineteen studies (73.08%) did not report any details regarding model interpretability.

Among the studies that included interpretability techniques, various approaches were used. Three articles reported some aspects of model interpretability by using correlation heatmaps [40,48,49] to provide visual representations of feature interrelationships, with colour-coded matrices showing the strength and direction of correlations. Li et al. [40], for instance, utilized a heatmap to examine the correlations between features for asthma prediction based on routine blood markers. Their analysis revealed that platelet distribution width and mean platelet volume, among other blood indicators, had a significant impact on the final prediction results. Three other studies [38,43,55] used feature importance algorithms to evaluate the contribution of each feature to the model's predictions. For example, in one of these studies [43], age consistently emerged as the most influential predictor across various health risks, with other variables such as systolic blood pressure and fasting

blood sugar proving significant for conditions like hypertension and diabetes. Additionally, body measurements like waist circumference (WC) and body mass index were highlighted as key predictors for heart disease and stroke. Furthermore, one study [42] employed the shapley additive explanations (SHAP) algorithm to interpret feature contributions within a health risk model for overweight individuals. WC and hip circumference (HC) were identified as the most significant factors influencing predictions. Smaller WC values were associated with reduced risk, while HC showed a more mixed impact. Other features like sex and osteoporosis had minimal influence, with SHAP values clustering around zero, indicating they played a less prominent role in the risk model.

3.3. Risk of Bias Assessment and Certainty of Evidence

Risk of Bias Assessment: The PROBAST technique was used to assess the risk of bias of the predictive models in the included studies. As shown in Table 10, and Figure 4, fourteen of the studies (53.85%) were rated as low risk for participant selection, while nine (34.62%) were rated as high risk. For predictors, seventeen of the studies (65.38%) were classified as low risk. In the outcome assessment category, sixteen (61.54%) of the studies were considered low risk, fifteen of the studies (57.69%) were judged as low risk in the analysis domain, and eight (30.77%) were high risk. Overall, eight (30.77%) of the studies were rated as low risk [42,43,48–51,53,55], while the remaining eighteen (69.20%) were categorized as high risk.

Table 10. Risk of bias assessment based on PROBAST (judgement: **+ Low** , **− High** , **? Unclear**).

Author	Participants	Predictors	Outcome	Analysis	Overall
Yadav [32]	−	?	−	−	−
Singh [33]	?	+	?	−	−
Khan [53]	+	+	+	+	+
Ashwath [34]	−	+	−	+	−
Chen [39]	−	?	+	−	−
Li [40]	−	+	+	+	−
Thakkar [35]	−	+	+	+	−
Shifa [48]	+	+	+	+	+
Park [43]	+	+	+	+	+
Shomorony [49]	+	+	+	+	+
Sahoo [50]	+	+	+	+	+
Diamantoulaki [56]	?	+	+	−	−
Rashid [44]	+	−	−	+	−
Parvataneni [51]	+	+	+	+	+
Chung [45]	−	+	+	?	−
Pawar [36]	−	+	+	+	−
Hossain [47]	+	+	−	?	−
Islam [54]	−	−	+	−	−
Bayati [52]	−	−	+	?	−
Kurysheva [57]	+	+	−	−	−
Salian [37]	?	?	−	−	−
GE [41]	+	−	−	+	−
Anbu [38]	+	−	−	+	−
Lin [42]	+	+	+	+	+
Yang [46]	+	?	−	−	−
Liu [55]	+	+	+	+	+

Certainty of Evidence: Using the GRADE tool, we evaluated the certainty of evidence for the included studies. Table S1 in the Supplementary Material, Tables 6 and 7 displays the results used to determine the final certainty of evidence. Four out of 26 studies (15.38%) [42,43,53,55] were assessed as having high certainty of evidence. Fourteen studies (53.85%) [32,34,36–39,44–47,52,54,56,57] has low certainty of evidence. The remaining

eight studies (30.77%) [33,35,40,41,48–51] were classified as having moderate certainty of evidence.

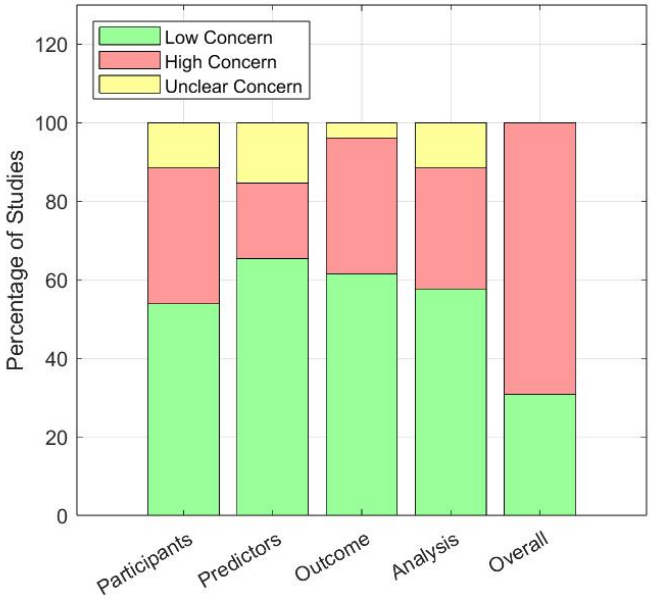

Figure 4. Stacked bar chart illustrating the distribution of bias concerns across various risk domains.

4. Discussion

This section discusses the key findings of the systematic review, addressing the four research questions based on the presented results.

4.1. General Characteristics

Geographical and Chronological Characteristics: The geographical distribution of the included studies reveals that they were spread across four continents: Asia, North America, Europe, and Australia, with the majority, 65.38%, originating from Asia, particularly from countries like India ($n = 7$), China ($n = 4$), and South Korea ($n = 3$). This strong representation of Asian countries is consistent with literature [58], which highlights the rapid adoption of artificial intelligence and ML technologies in the region. This trend can be attributed to increasing investments in digital health infrastructure [59] and a growing focus on addressing healthcare challenges through technology [60]. The COVID-19 pandemic, in particular, exposed critical gaps in healthcare systems, such as delayed detection of outbreaks and overwhelmed hospitals, especially in densely populated Asian countries. In response, countries such as China and India have accelerated the adoption of advanced technologies [61] to tackle healthcare challenges and ensure more resilient systems. South Korea has also made significant strides in AI-based healthcare systems [62]. These efforts may have contributed to the substantial number of publications from the region. Additionally, the rapid development of mobile technologies and electronic health records in Asia [63] provides abundant datasets, which further support the application and acceleration of ML technologies in healthcare. On the other hand, North America, represented by four studies from the USA, remains a key player in ML-driven healthcare innovations [64]. While no study was included from South America, Africa, in particular, showed no representation, highlighting a significant gap in the application of ML in healthcare in the region. This gap may stem from the limited access to advanced technologies and lower investments in health informatics [65]. According to Oladipo et al. [66], regions with weaker digital infrastructure and fewer research funding opportunities may struggle to keep pace with rapid advancements in AI and ML applications in healthcare. Bridging this gap will require dedicated efforts to enhance access to ML technologies and improve research capacity in underrepresented regions. Moreover, the temporal distribution of studies shows a notable

increase in publications over the past five years, with the majority (26.92%) published in 2023, signalling a growing interest in applying ML to health risk assessment.

Study Focus: The studies included in our systematic review exhibited a clear divide between those addressing general health risks and those focusing on specific health conditions. The focus on general health risks in twelve out of 26 studies (46.15%) reveals a significant level of interest in identifying multiple health conditions via ML, and this is crucial for population health management and preventive healthcare approaches. Specifically, by addressing multiple diseases, these studies aimed to provide insights into overall health risk factors rather than concentrating on a single condition, thus offering a more holistic perspective on health risk management. In contrast, the studies on specific health conditions ($n = 14$), such as asthma, diabetes, and cardiovascular health risks, focused on a deeper exploration of individual disease mechanisms and more targeted interventions. As demonstrated in the study by Cusi et al. [67], specific health risk assessments may be particularly useful in clinical settings, where accurate predictions for individual conditions can improve treatment planning and resource allocation. Overall, the diversity in the focus across the included studies reflects a balance between population-level health risk management and precision healthcare.

Data Collection Methods: The data collection methods varied substantially among the included studies. The majority (n = 14, 53.84%) relied on secondary data sources, highlighting the wider availability and utilization of public repositories such as UCI and Kaggle, which collectively served as the main data sources in seven studies. Although these repositories offer access to large, curated datasets crucial for training and validating machine learning models, there are significant drawbacks to relying predominantly on secondary data. For instance, as deduced by Ong et al. [68], the limited demographic diversity in many public datasets can result in models that fail to generalize well across diverse populations, potentially exacerbating health disparities. This problem stems from the fact that secondary datasets, especially publicly available ones, frequently lack detailed demographic, socioeconomic, and geographic data, which omits important variables that impact health outcomes [69]. As for primary data, our review included a total of eleven studies (42.30%) that are based on primary data, which is typically collected directly from participants or health facilities. While primary data allow for more tailored and context-specific analyses, only a small portion ($n = 2$, 18.18%) of these studies considered diverse populations, raising concerns about the generalizability of the models. This lack of diversity could lead to biased outcomes, as models trained on homogeneous populations may not perform well across different demographic groups [70]. Studies, such as reference [71], have also shown direct links between the risks of overfitting models and narrow datasets, which could lead to inaccuracies when the model is applied to more diverse or heterogeneous populations. Moreover, when evaluating the sufficiency of the datasets used to train the models, our review revealed that only eleven of the studies (42.31%) utilized datasets with sizes above 1000. Although this can be considered a reasonable threshold for developing reliable machine learning models, larger datasets often provide more variability [72], allowing models to capture complex patterns and improve prediction accuracy. However, six of the studies (23.07%) that employed datasets with fewer than 1000 samples present a limitation. Small sample sizes increase the risk of overfitting and reduce the model's ability to generalize to unseen data [73]. In addition, there was concern about the lack of transparency in nine studies (34.62%) that did not disclose their dataset sizes, which may obscure potential biases or limitations in the data used to train the models. Without knowing the size or scope of the datasets, evaluating the robustness of the machine learning models becomes challenging, and this can compromise the reliability of the study outcomes.

4.2. Model Development Process

Machine Learning Algorithms: In total, nine different machine learning algorithms were employed across the studies, with RF being the most frequently used ($n = 9$, 34.62%), followed by LGR ($n = 8$, 30.77%), DT ($n = 7$, 26.92%), and SVM ($n = 5$, 19.23%). The

dominance of RF in the included studies may be attributed to its ability to handle complex feature interactions and its robustness against overfitting due to the ensemble learning approach [74]. Specifically, RF constructs multiple decision trees and combines their predictions, which enhances both predictive accuracy and generalizability, making RF particularly suitable for HRAs. This was demonstrated in several articles, where RF consistently emerged as the best-performing algorithm, being identified as the top choice in 41.67% of the studies (5/12) that implemented more than one ML algorithm. Other ensemble methods, such as XGBoost, also showed notable performance, being identified as the best algorithm in three studies, indicating that the performance of RF as an ensemble learning model does not happen by chance. In contrast, single learning models, such as SVM and LGR, while also robust, are often less adaptable in managing variability and complexity without significant data preprocessing [15]. As supported by Ahsan et al. [75], data preprocessing techniques—such as feature scaling, feature selection, and handling data complexities—are essential for enhancing machine learning model performance. Additionally, substantial evidence in the literature [76,77] supports that kernel methods can significantly enhance SVM's ability to capture non-linear relationships. However, the limited consideration of these robustness factors in implementing the machine learning algorithms, as observed in several studies, may have contributed to reduced accuracy in some applications. For example, the influence of data complexity or health context on model performance (including the ensemble model) is evident in Table 9, where the accuracy of specific models varies significantly across different health scenarios.

It is crucial, therefore, to explore adaptive algorithms that can better accommodate data diversity to improve accuracy and reliability in diverse healthcare applications. Additionally, the substantial reliance on established ML algorithms in the included studies suggests a potential "reinvention of the wheel". In some cases, such as references [36,37], contributions mainly centred around data collection and preprocessing techniques, with limited technical novelty in the models themselves. In addition, some studies did not benchmark their approaches against other algorithms (particularly those that implemented a single algorithm, as can be observed in Table 8), raising questions about their efficacy. On the other hand, a small subset of the studies (seven, or 26.92%) introduced novel machine (NM) learning models to address specific limitations of classical approaches. For example, motivated by the observation that traditional models often fail to capture the correlation between data samples, which limits their accuracy, Li et al. [40]. proposed an affinity graph-enhanced classifier (AGEC), arguably the first attempt to directly exploit an affinity graph for classification. Sahoo et al. [50] presented a model-agnostic technique, glmboost, as an alternative to linear regression for fitting non-linear prediction models and performing variable selection in real-world mental health data. Although effective, this contribution appears somewhat incremental, as similar approaches to capturing non-linear relationships through non-linear regression have already been reported widely in the literature [78–80]. Similarly, Bayati et al. [52] combined well-known concepts from multi-task learning and group dimensionality reduction to develop a method that minimizes the number of biomarkers in HRA. GE [41] introduced a multi-label neural network method to predict chronic diseases, combining neural networks with multi-label learning using a cross-entropy loss function and backpropagation. Despite this effort, existing multi-label learning and neural network models, such as references [81,82], may have already demonstrated effectiveness in similar tasks, implying that the proposed approach may not significantly advance current capabilities.

Furthermore, ensuring the reliability and generalizability of machine learning models, particularly in health risk assessments, is also crucial. Model validation plays a critical role in this process, as it helps to confirm that the models perform well across different datasets and scenarios, reducing the risk of overfitting and enhancing their real-world applicability [83]. Reflecting this, various validation strategies were reported in the included studies. k-fold cross-validation, with values of $k = 10$, $k = 5$, and $k = 2$, was the most commonly used technique. This involves splitting the dataset into k equally sized subsets

or folds. While the *k*-fold cross-validation is generally effective, external validation can be more robust [84], as it involves testing the model on an entirely independent dataset, which better mimics real-world conditions. However, only one study (3.84%) employed external validation. To add to this, ten of the studies (38.46%) did not even specify their validation technique, which also raises concerns about the reliability of the findings reported in the studies, particularly in the context of healthcare, where the stakes are high.

Explainability Considerations: In the application of machine learning in healthcare, explainability is not merely a technical advantage but a necessity for ensuring that machine learning models can be trusted and seamlessly integrated into clinical workflows [85]. The reason is not farfetched because, for models used in diagnostics or HRAs to be truly effective, they must be interpretable to healthcare professionals [86]. However, our review shows that only seven (26.92%) of the studies incorporated explainability methods, leaving a substantial portion, nineteen out of 26 studies (73.08%), without any form of interpretability. Among those that did, three out of seven studies relied predominantly on heatmaps as their primary interpretability technique. Although heatmaps offer a visual representation of correlations between features, they fall short of providing a detailed explanation of how features contribute to individual predictions [87]. Heatmaps can show the relationships between variables, but they do not clarify the decision-making process of the model in specific cases, making them less effective in explaining why a particular prediction was made [88]. True explainability requires methods that provide granular insights into feature contributions, interactions, and the reasoning behind predictions [89]. Among the most promising techniques is SHAP [90], designed to offer more actionable explanations by assigning specific contribution values to each feature in a model. This helps clarify how individual features influence specific predictions, ensuring greater transparency in the decision-making process. Despite its clear advantages, SHAP was utilized in only one (3.85%) of the studies [42], highlighting an underutilization of advanced interpretability techniques in healthcare machine learning studies, which aligns with the broader conclusion drawn by Amann et al. [91].

Risk of Bias Assessment: The PROBAST technique was applied to assess the risk of bias in the predictive models used in the included studies. A total of 18 studies (69.23%) were categorized as high risk due to significant methodological concerns, including unspecified sample sizes, lack of validation, and unclear reporting of data handling processes. For example, Yadav et al. [32] demonstrated these limitations in their analysis, as many important details were missing, including unspecified population size, unspecified validation techniques, no details regarding how missing data were handled, and a lack of interpretation of feature importance. In contrast, eight studies (30.77%) were classified as low risk due to considerations such as relatively large sample populations, clear definitions of predictive outcomes, comprehensive descriptions of data handling methods, appropriate validation techniques, and the use of robust statistical analysis. Illustratively, the study by Lin et al. [42] was classified as low risk for several reasons, including reliance on a large retrospective cohort to enhance generalizability, robust statistical analysis for handling missing data, the use of SHAP beeswarm plots for transparent feature interpretation, and exhaustive cross-validation to enhance model reliability and minimize bias. Similarly, Liu et al. [55] exhibited comparable methodological rigour, further supporting their classification as low risk. It is important to note that although some studies, such as [50], classified as low risk could be considered to fall under moderate risk—primarily because they did not account for a large sample population—they were categorized as low risk because PROBAST does not offer a judgment category for moderate risk of bias.

In addition, we used the GRADE tool for the certainty of evidence assessment. Four out of 26 studies (15.38%) were categorized as having high certainty of evidence, attributed to large sample sizes, robust analyses, and clear conclusions. For instance, Park et al. [43] included 425,148 participants and employed a rigorous 10-fold cross-validation approach. In contrast, fourteen studies (53.85%) were classified as having low certainty of evidence due to methodological limitations, small or unspecified sample sizes, and limited descriptions

of the analytical process. For example, Yang et al. [46] reported a small sample size with notable methodological flaws. Eight studies (30.77%) were rated as having moderate certainty of evidence, balancing strengths such as robust analysis with limitations like moderate sample sizes or minor methodological issues. Studies like Singh et al. [33] and Shomorony et al. [49], in this case, showed moderate certainty due to their thorough analyses, despite minor constraints such as the risk of overfitting.

4.3. Emerging Techniques in HRA

Even though our systematic review has provided a relatively comprehensive coverage of existing ML applications for general health risk assessment, it is important to acknowledge the rapid advancements in the field over the last year or two, which our search protocol did not capture. In this section, we discuss emerging approaches such as federated learning, explainable AI (XAI), and the multimodal data integration that holds promise in HRA.

Federated Learning for Privacy-Preserving HRA: One major limitation of traditional machine learning models concerns data from different sources being pooled in a centralized location for training, which can lead to breaches in data privacy [92]. Federated learning (FL) has gained popularity as a solution for preserving data privacy in ML by allowing the training of models at decentralized locations close to the data-generating sources/devices. This approach enables collaborative model training across decentralized datasets without sharing raw data, preserving privacy while leveraging the power of collective data [93]. In healthcare, where data sensitivity is paramount, federated learning is particularly useful. As a result, FL has been adopted in several recent studies on HRA using machine learning, such as references [56,94–97]. Notably, Diamantoulaki et al. [56] proposed a deep learning model for assessing the occurrence of different diseases or complications using a federated learning approach. Brisimi et al. [94] developed a decentralized optimization framework for predicting hospitalizations due to cardiac events using a sparse SVM classifier. This approach enables multi-institutional collaborations without the need for data exchange. However, it remains unclear how to handle imbalances in practical applications where some institutions may possess more data or specific types of data, potentially leading to skewed model performance. Adnan et al. [95] demonstrated the application of a differentially private federated learning framework for analyzing histopathology images, with a focus on cancer detection. Similarly, Selvakanmani et al. [96] proposed a federated learning approach integrated with transfer learning for breast cancer classification, leveraging pre-trained ResNet and domain adversarial training to address data privacy and domain shift challenges across multiple medical centres. Huma et al. [97] introduced a federated learning-based framework for diabetes detection using a combined machine learning model. While the results presented in these studies support the potential of federated learning as an effective health risk assessment tool, as was also observed in most of the studies included in our systematic review, they did not fully address the challenges related to the generalization and fairness of federated learning in real-world healthcare settings, where data distribution may be highly heterogeneous and certain regions may be underrepresented.

Explainable AI for Transparency and Trust in ML Models: Apart from the XAI techniques applied in the studies included in our systematic review, such as SHAP, recent advancements in XAI techniques include local interpretable model-agnostic explanations (LIME) [98] and integrated gradients [99]. Specifically, LIME is tailored to provide local interpretability by approximating a complex model with a simpler one to understand individual predictions, whereas integrated gradients help in attributing the output of a deep learning model to its input features by considering the path taken by the model from a baseline to the actual input. Based on their apparent effectiveness, there are several studies in the literature that have demonstrated the effectiveness of these techniques in enhancing the explainability of HRA models, helping to clarify the factors influencing model predictions. For example, Curia [100] applied the LIME XAI technique to enhance the interpretability of predictions, offering a clearer understanding of the factors influencing

the risk of developing type 1 diabetes in a population. Ref. [101] utilized LIME to improve the transparency of the deep learning model by revealing the factors that influenced the classification of COVID-19, pneumonia, and tuberculosis from chest X-ray images. De et al. [102], used integrated gradients together with saliency, guided backpropagation, input × gradients, and DeepLIFT interpretation techniques to analyzed four convolutional neural network models: AlexNet, SqueezeNet, ResNet50, and VGG16 to provide actionable insights in the evaluation of cancer in Barrett's esophagus.

Multimodal Learning: Another advancement in the field of health risk assessment centres around using multimodal learning (which aims to integrate heterogeneous data sources) to potentially enhance HRA predictive performance and provide a comprehensive view of patient health [103]. For instance, modern healthcare data are diverse and often originate from various sources, such as clinical records, genomic data and imaging modalities. Thus, the idea behind multimodal learning is that these heterogeneous data sources can be integrated to enhance predictive performance. Moreover, with the proliferation of IoT devices and wearable sensors, real-time health monitoring has become increasingly feasible [104]. According to [105], continuous data collection from wearable devices, such as heart rate, glucose levels, and physical activity, can provide additional sources for multimodal fusion for timely health risk prediction and management. However, integrating these data into HRA models can present challenges [106], including handling vast, real-time data streams and ensuring data accuracy. Although Sangeetha et al. [107] have demonstrated that multimodal fusion deep learning neural networks can enhance lung cancer classification, future studies should explore ML frameworks that incorporate real-time data from IoT and wearable devices to enable dynamic HRA and proactive health management, potentially reducing the burden of chronic diseases and enhancing patient quality of life. Furthermore, recent deep learning architectures, such as transformers [108] and graph neural networks (GNNs) [109], also hold promise for HRA by enabling sophisticated data processing and complex pattern recognition, especially in structured and unstructured healthcare data. Transformers, originally developed for natural language processing, have demonstrated versatility in handling sequential and temporal data, making them suitable for longitudinal patient data [110]. For example, in HRA, transformers have been used to enhance heart disease prediction by analyzing temporal sequences of clinical data [111]. GNNs, on the other hand, could be used to model relational data, such as interactions between genetic markers or patient similarities based on demographic and clinical profiles [112]. Recent studies have applied GNNs in HRA to uncover hidden relationships between various health factors and improve risk prediction models [113]. In addition, reinforcement learning provides the potential for personalized health interventions, adapting recommendations based on individual patient responses over time [114]. In the context of HRA, reinforcement learning has been utilized to optimize treatment plans by dynamically adjusting interventions based on patient feedback and disease progression [114]. Incorporating/advancing these techniques in future HRA research could push the boundaries of precision medicine, tailoring risk assessments and interventions to individual patient needs.

4.4. Ethical Implications of ML in HRA

The application of ML in HRA raises significant ethical and bias-related concerns, particularly regarding the equitable treatment of diverse patient populations and the responsible handling of sensitive health data. Here, we discuss some of the primary ethical considerations, including bias mitigation, fairness, and privacy, which are essential for ensuring that ML models in HRA are both effective and ethically sound.

Bias in ML Models: ML models are often trained on data that may not fully represent the diversity of patient populations, which can lead to biased predictions. When models are disproportionately trained on data from certain demographic groups, such as specific age ranges, genders, or ethnic backgrounds, they may perform less accurately for underserved populations. This lack of representativeness can exacerbate healthcare disparities, poten-

tially leading to misdiagnosis or suboptimal care for individuals from marginalized groups. To mitigate these, techniques, such as re-sampling [115], re-weighting [116], and adversarial debiasing [117] have shown promise in effectively enhancing fairness and accuracy in ML model predictions across diverse groups. Moreover, ensuring fairness in ML-driven HRA systems means actively striving to provide equal treatment and accurate risk predictions for all patients. Techniques such as fairness-aware ML [118], which incorporates fairness constraints during model training, can also help achieve more equitable outcomes, as biased risk predictions could lead to unequal healthcare interventions, further entrenching existing health disparities.

Privacy and Data Security in HRA Models: Given the sensitive nature of healthcare data, maintaining privacy and data security is a fundamental ethical requirement in ML applications [119]. Privacy-preserving ML techniques, such as differential privacy and federated learning, offer potential solutions by allowing model training on distributed data sources without directly accessing sensitive patient information. While differential privacy techniques add noise to data to prevent individual patient identification, federated learning enables collaborative training across institutions without centralizing the data. As noted in the previous section, these approaches are particularly relevant in HRA settings, where privacy regulations such as the Health Insurance Portability and Accountability Act [120] in the United States and the General Data Protection Regulation [121] in Europe impose strict data protection requirements. Therefore, the need for incorporation of privacy-preserving techniques in HRA studies cannot be overemphasized to ensure compliance with ethical standards and legal regulations. Moreover, deploying ML models for HRA across varied healthcare environments, such as urban hospitals, rural clinics, and low-resource settings—also presents unique ethical challenges [122]. Each setting may differ significantly in patient demographics, data access, and regulatory requirements, which highlights the need for adaptable models sensitive to these distinctions. For instance, a model trained in a high-resource hospital may not generalize well to a rural clinic with limited resources or to populations with different health profiles. Although cross-validation on diverse datasets and sensitivity analyses can help ensure that these models perform equitably across different healthcare environments, developing adaptable and context-aware ML models for HRA is essential to further mitigate ethical challenges and promote inclusive healthcare solutions. Additionally, mechanisms for accountability, such as model audits and validation by independent third parties, are essential to ensure that models behave as intended and are regularly updated to reflect new medical knowledge and data sources.

4.5. Limitations of ML-Based HRA Models and Recommendations for Future Research

Despite the significant progress in applying machine learning to health risk assessments, several challenges remain prevalent across existing studies. First, the substantial reliance on traditional machine learning models in existing studies suggested a tendency to "reinvent the wheel". While some novel methods have been proposed to address limitations in traditional approaches, most remain incremental advancements over the current state of the art, offering limited contributions to significantly enhancing capabilities. Secondly, the lack of high-quality, well-annotated datasets continues to be a major challenge, as healthcare data tend to be noisy, incomplete, or biased and often contains complex patterns that limit the models' ability to capture intricate relationships needed for accurate predictions in real-world clinical settings. This was evident in Table 9, which shows the effectiveness of existing machine learning methods varies across different diseases, illustrating the relationship between complexity and performance. Thirdly, many studies lack representation of diverse populations, as the datasets often fail to cover a broad demographic range. In this context, the importance of diverse datasets in machine learning, particularly in healthcare applications, cannot be overstated. Diverse datasets ensure that models are trained on a representative sample of the population, which is crucial for producing accurate and generalizable results. When datasets lack demographic diversity—whether in terms of age,

gender, ethnicity, or socioeconomic status—there is a significant risk of developing models that do not perform well for all segments of the population. This underrepresentation can lead to biases in healthcare outcomes, ultimately compromising the effectiveness and equity of healthcare solutions. For instance, several studies, including references [123,124], have shown that algorithms trained on homogeneous datasets can perpetuate or exacerbate existing health disparities. A notable example is the work by Obermeyer et al. [125], which demonstrated how an algorithm designed to predict health needs in a large patient population exhibited bias against patients from certain ethnicities due to a lack of representation in the training data. Such biases can lead to misallocation of healthcare resources and inadequate treatment recommendations for underrepresented groups. The fourth limitation in existing studies is the insufficient emphasis on explainability, which remains a critical issue in the application of the machine learning model. Many studies either overlooked the importance of interpretability or relied on basic techniques such as heatmaps, which fail to provide clear explanations of how features influence predictions. In healthcare, where transparency and trust are essential for clinical adoption, the obvious absence of advanced explainability techniques can pose a significant barrier to integrating ML models into decision-making processes. Furthermore, the limited use of external validation across the studies may raise questions about the generalizability of reported findings, as models may perform well during internal validation but fail in real-world scenarios.

The above limitations highlight the need for continued research in the application of ML in HRA. To tackle the challenges, several solutions are recommended. First, the development of more advanced machine learning techniques (such as adaptive algorithms capable of accommodating data diversity) is essential. Such techniques should be able to learn complex patterns and relationships even in the presence of noisy or incomplete data, which would mitigate the limitations caused by these data quality issues and improve accuracy and reliability in varied healthcare applications. Second, future research should focus on creating frameworks for accessing more diverse, high-quality datasets. This could involve collaboration with healthcare institutions for accessing reliable data, community engagement initiatives, targeted recruitment of underrepresented populations, combination of primary and secondary data sources, and adopting techniques such as data augmentation or synthetic data generation to overcome limited dataset sizes and ensure that findings are generalizable across broader populations. Third, continuous emphasis should be placed on prioritizing explainability in healthcare machine learning models to promote fairness and the development of reliable, transparent, and trustworthy solutions. In fact, making model interpretability a compulsory consideration before accepting solutions for publication or implementation is essential. Finally, the implementation of uniform standards for reporting data processing and validation methods is essential to further enhance the robustness and transparency of ML models. Moreover, combining external validation and internal validation should be deemed necessary to improve the generalizability of these models across diverse populations. To enhance the reporting of validation techniques, a standardized template should clearly include the cross-validation approach applied, comparisons between internal and external validation, and details on training-validation splits and any hyperparameter tuning conducted.

4.6. Limitations of Our Study

This systematic review offers several advantages of HRA. Most importantly, it provides a comprehensive overview of the current landscape of ML applications in HRA, identifying the most commonly used algorithms, data collection methods, and evaluation techniques. In reviewing 26 studies across various regions and healthcare settings, notable trends were identified, such as the growing reliance on ensemble methods like RF and the preference for secondary data sources in model development. Additionally, the review effectively pinpointed critical gaps, including the lack of explainability considerations in the majority of studies and the frequent omission of model validation specifications in many other studies. Despite these achievements, the study has some drawbacks.

The main limitation is the specific terms used in the Boolean search string, such as "health risk assessment". These terms may have unintentionally excluded relevant studies, as many papers focusing on health conditions might not explicitly use the phrase "health risk assessment". To address this limitation, future reviews should incorporate synonyms and broader terms in their search strategies to capture a wider range of relevant literature. Furthermore, restricting the review to English-language articles may have excluded important studies published in other languages, especially from non-English-speaking regions with significant healthcare research output. In future work, including studies published in other languages could provide a more global perspective, with assistance from translators to ensure accurate interpretation. Moreover, recent advancements in HRA, such as federated learning for privacy preservation and multimodal learning, were conspicuously absent in the included articles, perhaps due to the lack of extensive specific keywords in our search protocol, as previously mentioned. While we incorporated recent developments in our discussion through targeted Google Scholar searches, this gap highlights an opportunity for future reviews focused on emerging areas, including wearable sensor data and real-time monitoring systems.

Another limitation is the selection of databases used for the search. Although this review utilized five major databases, there is a possibility that many important studies were missed, particularly those published in niche or regional journals that may not be indexed in the chosen databases. Expanding the range of databases in future reviews, including those that specialize in medical or regional research, could enhance the comprehensiveness of the review. Moreover, hand-searching relevant conference proceedings, grey literature, and reference lists could help in identifying missed articles that may have been overlooked by traditional database searches. The final limitation of this systematic review relates to the use of the PROBAST tool. Firstly, missing details in several studies, such as unreported sample characteristics and unspecified validation methods, significantly affected the classification of risk of bias, leading to the potential classification of some studies as "high" risk due to these omissions. Secondly, while PROBAST is a widely accepted tool for assessing the risk of bias in prediction models, it does not provide an option for "moderate" risk of bias. In this review, studies that might have fallen into the "moderate" category (between low and high risk) were classified as "low" risk, potentially leading to reporting bias. This may have impacted the overall risk assessment by oversimplifying distinctions between studies with minor methodological flaws and those with more robust methodologies. Future research could benefit from incorporating complementary tools that allow for a more nuanced classification of risk.

5. Conclusions

This paper has presented a systematic review of ML techniques, including data collection and validation methods utilized in various studies on health risk assessment using machine learning. The systematic review synthesized findings from 26 articles, with eleven studies (42.31%) focused on general health risks. Notably, the majority of the articles were published within the last five years, reflecting a significant increase in research output. Asia emerged as the most represented region, contributing the largest share of study publications. Established machine learning algorithms, such as SVM, RF, DT, and LGR, were frequently employed. There is a growing interest in ensemble learning methods. Accuracy, precision, recall, and AUC emerged as the most common evaluation metrics. While k-fold cross-validation remained the most widely used validation technique, a significant portion of the studies ($n = 10$) lacked explicit details on validation strategies. Moreover, explainability considerations, though gaining traction, were only addressed in seven (26.92%) of the included studies, indicating a need for greater emphasis on model transparency in future research.

Furthermore, most of the studies (nine out of eleven) that utilized primary data collection methods focused on a specific population, which limits the broader applicability of their outcomes. This specificity raises concerns about potential bias and generalization

issues, which are compounded by the lack of detailed data processing techniques in most studies. These limitations, including poor reporting on data handling, validation strategies, and model transparency, explain why relatively few studies were rated as low risk. Addressing these gaps will be crucial for advancing the role of machine learning in health risk assessment toward more reliable, generalizable, and transparent applications.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/electronics13224405/s1, Table S1: Characteristics of the included studies.

Author Contributions: Conceptualization, K.O.; methodology, S.E.A.; Screening, S.E.A. and E.O.N.; data curation, S.E.A., E.O.N. and K.O.; writing—original draft preparation, S.E.A.; writing—review and editing, S.E.A. and K.O.; funding acquisition, K.O. All authors have read and agreed to the published version of the manuscript.

Funding: This work was undertaken thanks in part to funding from the Connected Minds Program, supported by Canada First Research Excellence Fund, Grant No. CFREF-2022-00010.

Data Availability Statement: No new data were created in this study.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

Acronyms	Meaning
SVM	Support Vector Machine
RF	Random Forest
KNN	K Nearest Neighbour
DNN	Deep Neural Network
NN	Neural Network
DT	Decision Trees
NB	Naive Bayes
EN	Ensemble
LR	Linear Regression
LGR	Logistics Regression
NM	Novel Model
ACC	Accuracy
AUC	Area Under the Curve
RMSE	Root Mean Squared Error
SVR	Support Vector Regression
SHAP	Shapley Additive Explanations
HRA	Health Risk Assessment
WHO	World Health Organization
ACM	Association for Computing Machinery
WOS	Web of Science
PRISMA	Preferred Reporting Items for Systematic Review and Meta-analysis

References

1. Betancourt, M.; Roberts, K.; Bennett, T.L.; Driscoll, E.; Jayaraman, G.; Pelletier, L. Monitoring chronic diseases in Canada: The chronic disease indicator framework. *Chronic Dis. Inj. Can.* **2014**, *34*, 1. [CrossRef] [PubMed]
2. Schmidt, H.; Mah, C.L.; Cook, B.; Hoang, S.; Taylor, E.; Blacksher, E.; Goldberg, D.S.; Novick, L.; Aspradaki, A.A.; Tzoutzas, I.; et al. Chronic disease prevention and health promotion. In *Public Health Ethics: Cases Spanning the Globe*; OAPEN Library: The Hague, The Netherlands, 2016; pp. 137–176.
3. González, K.; Fuentes, J.; Márquez, J.L. Physical inactivity, sedentary behavior and chronic diseases. *Korean J. Fam. Med.* **2017**, *38*, 111. [CrossRef] [PubMed]
4. World Health Organization. *Global Report on Diabetes*; World Health Organization: Geneva, Switzerland, 2016.
5. Meador, M.; Lewis, J.H.; Bay, R.C.; Wall, H.K.; Jackson, C. Who are the undiagnosed? Disparities in hypertension diagnoses in vulnerable populations. *Fam. Community Health* **2020**, *43*, 35–45. [CrossRef] [PubMed]

6. Svendsen, M.T.; Bak, C.K.; Sørensen, K.; Pelikan, J.; Riddersholm, S.J.; Skals, R.K.; Mortensen, R.N.; Maindal, H.T.; Bøggild, H.; Nielsen, G.; et al. Associations of health literacy with socioeconomic position, health risk behavior, and health status: A large national population-based survey among Danish adults. *Bmc Public Health* **2020**, *20*, 1–12. [CrossRef]
7. Shinde, S.A.; Rajeswari, P.R. Intelligent health risk prediction systems using machine learning: A review. *Int. J. Eng. Technol.* **2018**, *7*, 1019–1023. [CrossRef]
8. Allery, F.; Pineda-Moncusí, M.; Tomlinson, C.; Pontikos, N.; Thygesen, J.H.; Khalid, S.; Consortium, C.C.U.I. Towards mitigating health inequity via machine learning: A nationwide cohort study to develop and validate ethnicity-specific models for prediction of cardiovascular disease risk in COVID-19 patients. *medRxiv* **2023**. [CrossRef]
9. Chiu, Y.L.; Jhou, M.J.; Lee, T.S.; Lu, C.J.; Chen, M.S. Health data-driven machine learning algorithms applied to risk indicators assessment for chronic kidney disease. *Risk Manag. Healthc. Policy* **2021**, *14*, 4401–4412. [CrossRef]
10. Chen, Z.; Liu, B. *Lifelong Machine Learning*; Springer Nature: Berlin/Heidelberg, Germany, 2022.
11. Joachims, T. Estimating the generalization performance of a SVM efficiently. In *Technical Report*; Universität Dortmund: Dortmund, Germany, 2001.
12. Abdi, H.; Valentin, D.; Edelman, B. *Neural Networks*; Sage: Thousand Oaks, CA, USA, 1999; p. 124.
13. Vartanian, T.P. *Secondary Data Analysis*; Oxford University Press: Oxford, UK, 2010.
14. Johnston, M.P. Secondary data analysis: A method of which the time has come. *Qual. Quant. Methods Libr.* **2014**, *3*, 619–626.
15. Maharana, K.; Mondal, S.; Nemade, B. A review: Data pre-processing and data augmentation techniques. *Glob. Transitions Proc.* **2022**, *3*, 91–99. [CrossRef]
16. Liberty, E.; Lang, K.; Shmakov, K. Stratified sampling meets machine learning. In Proceedings of the International Conference on Machine Learning. PMLR, New York, NY, USA, 19–24 June 2016; pp. 2320–2329.
17. Mumuni, A.; Mumuni, F. Data augmentation: A comprehensive survey of modern approaches. *Array* **2022**, *16*, 100258. [CrossRef]
18. Mishra, A.K.; Chong, B.; Arunachalam, S.P.; Oberg, A.L.; Majumder, S. Machine Learning Models for Pancreatic Cancer Risk Prediction Using Electronic Health Record Data-A Systematic Review and Assessment. *Off. J. Am. Coll. Gastroenterol. Acg* **2022**, *119*, 1466–1482. [CrossRef]
19. Usman, T.M.; Saheed, Y.K.; Nsang, A.; Ajibesin, A.; Rakshit, S. A systematic literature review of machine learning based risk prediction models for diabetic retinopathy progression. *Artif. Intell. Med.* **2023**, *143*, 102617. [CrossRef]
20. Singh, M.; Kumar, A.; Khanna, N.N.; Laird, J.R.; Nicolaides, A.; Faa, G.; Johri, A.M.; Mantella, L.E.; Fernandes, J.F.E.; Teji, J.S.; et al. Artificial intelligence for cardiovascular disease risk assessment in personalised framework: A scoping review. *Eclinical Med.* **2024**, *73*. [CrossRef]
21. Abdulazeem, H.; Whitelaw, S.; Schauberger, G.; Klug, S.J. A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data. *PLoS ONE* **2023**, *18*, e0274276. [CrossRef]
22. Fleuren, L.M.; Klausch, T.L.; Zwager, C.L.; Schoonmade, L.J.; Guo, T.; Roggeveen, L.F.; Swart, E.L.; Girbes, A.R.; Thoral, P.; Ercole, A.; et al. Machine learning for the prediction of sepsis: A systematic review and meta-analysis of diagnostic test accuracy. *Intensive Care Med.* **2020**, *46*, 383–400. [CrossRef]
23. Xiong, S.; Chen, W.; Jia, X.; Jia, Y.; Liu, C. Machine learning for prediction of asthma exacerbations among asthmatic patients: A systematic review and meta-analysis. *BMC Pulm. Med.* **2023**, *23*, 278. [CrossRef]
24. Tranfield, D.; Denyer, D.; Smart, P. Towards a methodology for developing evidence-informed management knowledge by means of systematic review. *Br. J. Manag.* **2003**, *14*, 207–222. [CrossRef]
25. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, 102660.
26. Ali, O.; Abdelbaki, W.; Shrestha, A.; Elbasi, E.; Alryalat, M.A.A.; Dwivedi, Y.K. A systematic literature review of artificial intelligence in the healthcare sector: Benefits, challenges, methodologies, and functionalities. *J. Innov. Knowl.* **2023**, *8*, 100333. [CrossRef]
27. Braun, V.; Clarke, V. Using thematic analysis in psychology. *Qual. Res. Psychol.* **2006**, *3*, 77–101. [CrossRef]
28. Wolff, R.F.; Moons, K.G.; Riley, R.D.; Whiting, P.F.; Westwood, M.; Collins, G.S.; Reitsma, J.B.; Kleijnen, J.; Mallett, S.; PROBAST Group. PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann. Intern. Med.* **2019**, *170*, 51–58. [CrossRef] [PubMed]
29. Ryan, R.; Santesso, N.; Hill, S. *Preparing Summary of Findings (SoF) Tables*; Cochrane Consumers and Communication Group: Melbourne, Australia, 2016.
30. Oyibo, K.; Toyonaga, S. Conceptual Frameworks for Designing and Evaluating Persuasive Messages aimed at Changing Behavior: Systematic Review. *Comput. Hum. Behav. Rep.* **2024**, *15*, 100448. [CrossRef]
31. Zhou, L.; Pan, S.; Wang, J.; Vasilakos, A.V. Machine learning on big data: Opportunities and challenges. *Neurocomputing* **2017**, *237*, 350–361. [CrossRef]
32. Yadav, A.; Mittal, A.K. An ensemble machine learning based approach for health risk prediciton. In Proceedings of the 2021 4th International Conference on Recent Trends in Computer Science and Technology (ICRTCST), Jamshedpur, India, 11–12 February 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 302–308.
33. Singh, A.; Ramkumar, K. Evaluation of SVM Kernels for Health Risks Assessment. *Helix* **2019**, *9*, 5009–5023. [CrossRef]

34. Ashwath, S.; Gurukishore, G.; Maheedhar, A.; Elizabeth, N.E. Enhanced Cardiovascular Risk Prediction Using ML Powered Web Application. In Proceedings of the 2024 11th International Conference on Signal Processing and Integrated Networks (SPIN), Noida, India, 21–22 March 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 199–204.
35. Thakkar, D.; Gandhi, V.C.; Trivedi, D. Forecasting Maternal Women's Health Risks using Random Forest Classifier. In Proceedings of the 2024 International Conference on Inventive Computation Technologies (ICICT), Lalitpur, Nepal, 24–26 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 961–965.
36. Pawar, L.; Malhotra, J.; Sharma, A.; Arora, D.; Vaidya, D. A robust machine learning predictive model for maternal health risk. In Proceedings of the 2022 3rd International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 17–19 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 882–888.
37. Salian, P.; Puneeth, B.; Nargis, T.; Salian, S.; Vanishree, B. User Input Based Health Risk Assessment to Predict Diabetes, Obesity and Heart Risk factors. In Proceedings of the 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 5–7 April 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–7.
38. Anbu, S.; Sarmah, B. Machine learning approach for predicting womens health risk. In Proceedings of the 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 6–7 January 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–4.
39. Chen, L.; Zhu, H. Analysis of physical health risk dynamic evaluation system based on sports network technology. *Comput. Commun.* **2022**, *181*, 257–266. [CrossRef]
40. Li, D.; Abhadiomhen, S.E.; Zhou, D.; Shen, X.J.; Shi, L.; Cui, Y. Asthma prediction via affinity graph enhanced classifier: A machine learning approach based on routine blood biomarkers. *J. Transl. Med.* **2024**, *22*, 100. [CrossRef]
41. Ge, R.; Zhang, R.; Wang, P. Prediction of chronic diseases with multi-label neural network. *IEEE Access* **2020**, *8*, 138210–138216. [CrossRef]
42. Lin, W.; Shi, S.; Lan, H.; Wang, N.; Huang, H.; Wen, J.; Chen, G. Identification of influence factors in overweight population through an interpretable risk model based on machine learning: A large retrospective cohort. *Endocrine* **2024**, *83*, 604–614. [CrossRef]
43. Park, H.; Jung, S.Y.; Han, M.K.; Jang, Y.; Moon, Y.R.; Kim, T.; Shin, S.Y.; Hwang, H. Lowering Barriers to Health Risk Assessments in Promoting Personalized Health Management. *J. Pers. Med.* **2024**, *14*, 316. [CrossRef]
44. Rashid, J.; Batool, S.; Kim, J.; Wasif Nisar, M.; Hussain, A.; Juneja, S.; Kushwaha, R. An augmented artificial intelligence approach for chronic diseases prediction. *Front. Public Health* **2022**, *10*, 860396. [CrossRef]
45. Chung, K.; Yoo, H.; Choe, D.E. Ambient context-based modeling for health risk assessment using deep neural network. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *11*, 1387–1395. [CrossRef]
46. Yang, W.F.; Liu, H.-H.; Ting, C.T. The Feasibility of Applying Artificial Intelligence Detection Technology in Predicting the Risk of Hypertension. *Adv. Artif. Intell. Mach. Learn.* **2023**, *3*, 2019. [CrossRef]
47. Hossain, M.A.; Ferdousi, R.; Alhamid, M.F. Knowledge-driven machine learning based framework for early-stage disease risk prediction in edge environment. *J. Parallel Distrib. Comput.* **2020**, *146*, 25–34. [CrossRef]
48. Shifa, H.A.; Mojumdar, M.U.; Rahman, M.M.; Chakraborty, N.R.; Gupta, V. Machine learning models for maternal health risk prediction based on clinical data. In Proceedings of the 2024 11th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 28 February–1 March 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1312–1318.
49. Shomorony, I.; Cirulli, E.; Huang, L.; Napier, L.; Heister, R.; Hicks, M.; Cohen, I.; Yu, H.; Swisher, C.; Schenker-Ahmed, N.; et al. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. *Genome Med.* **2020**, *12*, 7. [CrossRef]
50. Sahoo, I.; Amona, E.; Kuttikat, M.; Chan, D. Enhancing Mental Health Predictions: A Gradient Boosted Model for Sri Lankan Camp Refugees. *Soc. Sci.* **2024**, *13*, 255. [CrossRef]
51. Parvataneni, K.; Zaidi, S.H.; Kazmi, F.; Kazmi, S.H. AI Framework for Fetal Health Risk Prediction. In Proceedings of the 2023 5th International Conference on Bio-engineering for Smart Technologies (BioSMART), Paris, France, 7–9 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
52. Bayati, M.; Bhaskar, S.; Montanari, A. Statistical analysis of a low cost method for multiple disease prediction. *Stat. Methods Med. Res.* **2018**, *27*, 2312–2328. [CrossRef]
53. Khan, A.; Uddin, S.; Srinivasan, U. Chronic disease prediction using administrative data and graph theory: The case of type 2 diabetes. *Expert Syst. Appl.* **2019**, *136*, 230–241. [CrossRef]
54. Islam, M.S.; Husain, S.; Mustafa, J.; Gu, Y. A novel machine learning prediction model for aerosol transport in upper 17-generations of the human respiratory tract. *Future Internet* **2022**, *14*, 247. [CrossRef]
55. Liu, Y.; Zhang, Z.; Thompson, C.; Leibbrandt, R.; Qin, S.; Yepes, A.J. Stacked Attention-based Networks for Accurate and Interpretable Health Risk Prediction. In Proceedings of the 2023 International Joint Conference on Neural Networks (IJCNN), Gold Coast, Australia, 18–23 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–8.
56. Diamantoulaki, I.; Diamantoulakis, P.D.; Bouzinis, P.S.; Sarigiannidis, P.; Karagiannidis, G.K. Health risk assessment with federated learning. In Proceedings of the 2022 International Balkan Conference on Communications and Networking (BalkanCom), Sarajevo, Bosnia and Herzegovina, 22–24 August 2022; IEEE: Piscataway, NJ, USA, 2022; pp. 57–61.

57. Kurysheva, N.I.; Rodionova, O.Y.; Pomerantsev, A.L.; Sharova, G.A.; Golubnitschaja, O. Machine learning–couched treatment algorithms tailored to individualized profile of patients with primary anterior chamber angle closure predisposed to the glaucomatous optic neuropathy. *EPMA J.* **2023**, *14*, 527–538. [CrossRef]
58. Wong, B.K.M.; Vengusamy, S.; Bastrygina, T. Healthcare digital transformation through the adoption of artificial intelligence. In *Artificial Intelligence, Big Data, Blockchain and 5G for the Digital Transformation of the Healthcare Industry*; Elsevier: Amsterdam, The Netherlands, 2024; pp. 87–110.
59. Li, X.; Krumholz, H.M.; Yip, W.; Cheng, K.K.; De Maeseneer, J.; Meng, Q.; Mossialos, E.; Li, C.; Lu, J.; Su, M.; et al. Quality of primary health care in China: Challenges and recommendations. *Lancet* **2020**, *395*, 1802–1812. [CrossRef]
60. Sun, S.; Xie, Z.; Yu, K.; Jiang, B.; Zheng, S.; Pan, X. COVID-19 and healthcare system in China: Challenges and progression for a sustainable future. *Glob. Health* **2021**, *17*, 1–8. [CrossRef] [PubMed]
61. Bhatia, R. Telehealth and COVID-19: Using technology to accelerate the curve on access and quality healthcare for citizens in India. *Technol. Soc.* **2021**, *64*, 101465. [CrossRef] [PubMed]
62. Lim, K.; Heo, T.Y.; Yun, J. Trends in the approval and quality management of artificial intelligence medical devices in the Republic of Korea. *Diagnostics* **2022**, *12*, 355. [CrossRef] [PubMed]
63. Shilpa, D.; Naik, P.R.; Shewade, H.D.; Sudarshan, H. Assessing the implementation of a mobile App-based electronic health record: A mixed-method study from South India. *J. Educ. Health Promot.* **2020**, *9*, 102. [PubMed]
64. Habehh, H.; Gohel, S. Machine learning in healthcare. *Curr. Genom.* **2021**, *22*, 291. [CrossRef]
65. Akingbola, A.; Adegbesan, A.; Ojo, O.; Otumara, J.U.; Alao, U.H. Artificial intelligence and cancer care in Africa. *J. Med. Surg. Public Health* **2024**, *3*, 100132. [CrossRef]
66. Oladipo, E.K.; Adeyemo, S.F.; Oluwasanya, G.J.; Oyinloye, O.R.; Oyeyiola, O.H.; Akinrinmade, I.D.; Elutade, O.A.; Areo, D.O.; Hamzat, I.O.; Olakanmi, O.D.; et al. Impact and challenges of artificial intelligence integration in the African health sector: A review. *Trends Med. Res.* **2024**, *19*, 220–235. [CrossRef]
67. Cusi, K.; Isaacs, S.; Barb, D.; Basu, R.; Caprio, S.; Garvey, W.T.; Kashyap, S.; Mechanick, J.I.; Mouzaki, M.; Nadolsky, K.; et al. American Association of Clinical Endocrinology clinical practice guideline for the diagnosis and management of nonalcoholic fatty liver disease in primary care and endocrinology clinical settings: Co-sponsored by the American Association for the Study of Liver Diseases (AASLD). *Endocr. Pract.* **2022**, *28*, 528–562.
68. Ong Ly, C.; Unnikrishnan, B.; Tadic, T.; Patel, T.; Duhamel, J.; Kandel, S.; Moayedi, Y.; Brudno, M.; Hope, A.; Ross, H.; et al. Shortcut learning in medical AI hinders generalization: Method for estimating AI model generalization without external data. *NPJ Digit. Med.* **2024**, *7*, 124. [CrossRef]
69. Frederick, K.; Barnard-Brak, L.; Sulak, T. Under-representation in nationally representative secondary data. *Int. J. Res. Method Educ.* **2012**, *35*, 31–40. [CrossRef]
70. Chin, M.H.; Afsar-Manesh, N.; Bierman, A.S.; Chang, C.; Colón-Rodríguez, C.J.; Dullabh, P.; Duran, D.G.; Fair, M.; Hernandez-Boussard, T.; Hightower, M.; et al. Guiding principles to address the impact of algorithm bias on racial and ethnic disparities in health and health care. *JAMA Netw. Open* **2023**, *6*, e2345050. [CrossRef] [PubMed]
71. Cunningham, P.; Delany, S.J. Underestimation bias and underfitting in machine learning. In Proceedings of the Trustworthy AI-Integrating Learning, Optimization and Reasoning: First International Workshop, TAILOR 2020, Virtual Event, 4–5 September 2020; Revised Selected Papers 1; Springer: Berlin/Heidelberg, Germany, 2021; pp. 20–31.
72. Althnian, A.; AlSaeed, D.; Al-Baity, H.; Samha, A.; Dris, A.B.; Alzakari, N.; Abou Elwafa, A.; Kurdi, H. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Appl. Sci.* **2021**, *11*, 796. [CrossRef]
73. Rácz, A.; Bajusz, D.; Héberger, K. Effect of dataset size and train/test split ratios in QSAR/QSPR multiclass classification. *Molecules* **2021**, *26*, 1111. [CrossRef] [PubMed]
74. Rhodes, J.S.; Cutler, A.; Moon, K.R. Geometry-and accuracy-preserving random forest proximities. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 10947–10959. [CrossRef]
75. Ahsan, M.M.; Mahmud, M.P.; Saha, P.K.; Gupta, K.D.; Siddique, Z. Effect of data scaling methods on machine learning algorithms and model performance. *Technologies* **2021**, *9*, 52. [CrossRef]
76. Roman, I.; Santana, R.; Mendiburu, A.; Lozano, J.A. In-depth analysis of SVM kernel learning and its components. *Neural Comput. Appl.* **2021**, *33*, 6575–6594. [CrossRef]
77. Wu, Y.; Shen, X.J.; Abhadiomhen, S.E.; Yang, Y.; Gu, J.N. Kernel ensemble support vector machine with integrated loss in shared parameters space. *Multimed. Tools Appl.* **2023**, *82*, 18077–18096. [CrossRef]
78. Abhishek, K.; Datta, S.; Mahapatra, S.S. Multi-objective optimization in drilling of CFRP (polyester) composites: Application of a fuzzy embedded harmony search (HS) algorithm. *Measurement* **2016**, *77*, 222–239. [CrossRef]
79. Kumar, J.A.; Krithiga, T.; Anand, K.V.; Sathish, S.; Namasivayam, S.K.R.; Renita, A.; Hosseini-Bandegharaei, A.; Praveenkumar, T.; Rajasimman, M.; Bhat, N.; et al. Kinetics and regression analysis of phenanthrene adsorption on the nanocomposite of CaO and activated carbon: Characterization, regeneration, and mechanistic approach. *J. Mol. Liq.* **2021**, *334*, 116080. [CrossRef]
80. Raza, M.B.; Datta, S.P.; Golui, D.; Barman, M.; Das, T.K.; Sahoo, R.N.; Upadhyay, D.; Rahman, M.M.; Behera, B.; Naveenkumar, A. Synthesis and performance evaluation of novel bentonite-supported nanoscale zero valent iron for remediation of arsenic contaminated water and soil. *Molecules* **2023**, *28*, 2168. [CrossRef]
81. Cerri, R.; Barros, R.C.; De Carvalho, A.C. Hierarchical multi-label classification using local neural networks. *J. Comput. Syst. Sci.* **2014**, *80*, 39–56. [CrossRef]

82. Kurata, G.; Xiang, B.; Zhou, B. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 521–526.
83. Cabitza, F.; Campagner, A.; Soares, F.; de Guadiana-Romualdo, L.G.; Challa, F.; Sulejmani, A.; Seghezzi, M.; Carobene, A. The importance of being external. methodological insights for the external validation of machine learning models in medicine. *Comput. Methods Programs Biomed.* **2021**, *208*, 106288. [CrossRef]
84. Ramspek, C.L.; Jager, K.J.; Dekker, F.W.; Zoccali, C.; van Diepen, M. External validation of prognostic models: What, why, how, when and where? *Clin. Kidney J.* **2021**, *14*, 49–58. [CrossRef]
85. Rasheed, K.; Qayyum, A.; Ghaly, M.; Al-Fuqaha, A.; Razi, A.; Qadir, J. Explainable, trustworthy, and ethical machine learning for healthcare: A survey. *Comput. Biol. Med.* **2022**, *149*, 106043. [CrossRef]
86. Allgaier, J.; Mulansky, L.; Draelos, R.L.; Pryss, R. How does the model make predictions? A systematic literature review on the explainability power of machine learning in healthcare. *Artif. Intell. Med.* **2023**, *143*, 102616. [CrossRef]
87. Fuxin, L.; Qi, Z.; Khorram, S.; Shitole, V.; Tadepalli, P.; Kahng, M.; Fern, A. From heatmaps to structured explanations of image classifiers. *Appl. Lett.* **2021**, *2*, e46. [CrossRef]
88. Zhao, S.; Guo, Y.; Sheng, Q.; Shyr, Y. Advanced heat map and clustering analysis using heatmap3. *Biomed Res. Int.* **2014**, *2014*, 986048. [CrossRef]
89. Sapkota, S.C.; Yadav, A.; Khatri, A.; Singh, T.; Dahal, D. Explainable hybridized ensemble machine learning for the prognosis of the compressive strength of recycled plastic-based sustainable concrete with experimental validation. *Multiscale Multidiscip. Model. Exp. Des.* **2024**, *7*, 6073–6096. [CrossRef]
90. Van den Broeck, G.; Lykov, A.; Schleich, M.; Suciu, D. On the tractability of SHAP explanations. *J. Artif. Intell. Res.* **2022**, *74*, 851–886. [CrossRef]
91. Amann, J.; Blasimme, A.; Vayena, E.; Frey, D.; Madai, V.I.; Consortium, P. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 310. [CrossRef]
92. El Mestari, S.Z.; Lenzini, G.; Demirci, H. Preserving data privacy in machine learning systems. *Comput. Secur.* **2024**, *137*, 103605. [CrossRef]
93. Neo, E.X.; Hasikin, K.; Mokhtar, M.I.; Lai, K.W.; Azizan, M.M.; Razak, S.A.; Hizaddin, H.F. Towards integrated air pollution monitoring and health impact assessment using federated learning: A systematic review. *Front. Public Health* **2022**, *10*, 851553. [CrossRef]
94. Brisimi, T.S.; Chen, R.; Mela, T.; Olshevsky, A.; Paschalidis, I.C.; Shi, W. Federated learning of predictive models from federated electronic health records. *Int. J. Med. Inform.* **2018**, *112*, 59–67. [CrossRef]
95. Adnan, M.; Kalra, S.; Cresswell, J.C.; Taylor, G.W.; Tizhoosh, H.R. Federated learning and differential privacy for medical image analysis. *Sci. Rep.* **2022**, *12*, 1953. [CrossRef]
96. Selvakanmani, S.; Devi, G.D.; Rekha, V.; Jeyalakshmi, J. Privacy-Preserving Breast Cancer Classification: A Federated Transfer Learning Approach. *J. Imaging Inform. Med.* **2024**, *37*, 1488.
97. Huma, Z.E.; Tariq, N.; Zaidi, S. Predictive Machine Learning Models for Early Diabetes Diagnosis: Enhancing Accuracy and Privacy with Federated Learning. *J. Comput. Biomed. Inform.* **2024**, *8*. [CrossRef]
98. Singla, K.; Biswas, S. Machine learning explanability method for the multi-label classification model. In Proceedings of the 2021 IEEE 15th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 27–29 January 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 337–340.
99. Poulton, A.; Eliens, S. Explaining transformer-based models for automatic short answer grading. In Proceedings of the 5th International Conference on Digital Technology in Education, Busan, Republic of Korea, 23–25 July 2021; pp. 110–116.
100. Curia, F. Explainable and transparency machine learning approach to predict diabetes develop. *Health Technol.* **2023**, *13*, 769–780. [CrossRef]
101. Krishnamoorthy, T.V.; Venkataiah, C.; Rao, Y.M.; Prasad, D.R.; Chowdary, K.U.; Jayamma, M.; Sireesha, R. A novel NASNet model with LIME explanability for lung disease classification. *Biomed. Signal Process. Control* **2024**, *93*, 106114. [CrossRef]
102. de Souza, L.A., Jr.; Mendel, R.; Strasser, S.; Ebigbo, A.; Probst, A.; Messmann, H.; Papa, J.P.; Palm, C. Convolutional Neural Networks for the evaluation of cancer in Barrett's esophagus: Explainable AI to lighten up the black-box. *Comput. Biol. Med.* **2021**, *135*, 104578. [CrossRef]
103. Zhou, H.Y.; Yu, Y.; Wang, C.; Zhang, S.; Gao, Y.; Pan, J.; Shao, J.; Lu, G.; Zhang, K.; Li, W. A transformer-based representation-learning model with unified processing of multimodal input for clinical diagnostics. *Nat. Biomed. Eng.* **2023**, *7*, 743–755. [CrossRef]
104. Almalawi, A.; Khan, A.I.; Alsolami, F.; Abushark, Y.B.; Alfakeeh, A.S. Managing security of healthcare data for a modern healthcare system. *Sensors* **2023**, *23*, 3612. [CrossRef]
105. Kazanskiy, N.L.; Khonina, S.N.; Butt, M.A. A review on flexible wearables-Recent developments in non-invasive continuous health monitoring. *Sens. Actuators Phys.* **2024**, *366*, 114993. [CrossRef]
106. Duan, J.; Xiong, J.; Li, Y.; Ding, W. Deep learning based multimodal biomedical data fusion: An overview and comparative review. *Inf. Fusion* **2024**, *112*, 102536. [CrossRef]
107. Sangeetha, S.; Mathivanan, S.K.; Karthikeyan, P.; Rajadurai, H.; Shivahare, B.D.; Mallik, S.; Qin, H. An enhanced multimodal fusion deep learning neural network for lung cancer classification. *Syst. Soft Comput.* **2024**, *6*, 200068.

108. Islam, S.; Elmekki, H.; Elsebai, A.; Bentahar, J.; Drawel, N.; Rjoub, G.; Pedrycz, W. A comprehensive survey on applications of transformers for deep learning tasks. *Expert Syst. Appl.* **2023**, *241*, 122666. [CrossRef]
109. Veličković, P. Everything is connected: Graph neural networks. *Curr. Opin. Struct. Biol.* **2023**, *79*, 102538. [CrossRef]
110. Siebra, C.A.; Kurpicz-Briki, M.; Wac, K. Transformers in health: A systematic review on architectures for longitudinal data analysis. *Artif. Intell. Rev.* **2024**, *57*, 32. [CrossRef]
111. Rahman, A.U.; Alsenani, Y.; Zafar, A.; Ullah, K.; Rabie, K.; Shongwe, T. Enhancing heart disease prediction using a self-attention-based transformer model. *Sci. Rep.* **2024**, *14*, 514. [CrossRef]
112. Von Arnim, G.P. Personalized Drug Recommendation with Pretrained GNNs on a Large Biomedical Knowledge Graph. Ph.D. Thesis, Freie Universität Berlin, Berlin, Germany, 2024.
113. Wang, G. A methodology for long-term offshore structural health monitoring using stand-alone GNSS: Case study in the Gulf of Mexico. *Struct. Health Monit.* **2024**, *23*, 463–478. [CrossRef]
114. Deliu, N.; Williams, J.J.; Chakraborty, B. Reinforcement learning in modern biostatistics: Constructing optimal adaptive interventions. In *International Statistical Review*; Wiley Online Library: Hoboken, NJ, USA, 2024.
115. Maragkoudakis, E.; Papadopoulos, S.; Varlamis, I.; Diou, C. Sampling Strategies for Mitigating Bias in Face Synthesis Methods. *arXiv* **2024**, arXiv:2405.11320.
116. Gupta, S.; Kaur, K.; Jain, S. Eqbal-rs: Mitigating popularity bias in recommender systems. *J. Intell. Inf. Syst.* **2024**, *62*, 509–534. [CrossRef]
117. Lim, J.; Kim, Y.; Kim, B.; Ahn, C.; Shin, J.; Yang, E.; Han, S. Biasadv: Bias-adversarial augmentation for model debiasing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 3832–3841.
118. Weerts, H.; Pfisterer, F.; Feurer, M.; Eggensperger, K.; Bergman, E.; Awad, N.; Vanschoren, J.; Pechenizkiy, M.; Bischl, B.; Hutter, F. Can fairness be automated? Guidelines and opportunities for fairness-aware AutoML. *J. Artif. Intell. Res.* **2024**, *79*, 639–677. [CrossRef]
119. Dhirani, L.L.; Mukhtiar, N.; Chowdhry, B.S.; Newe, T. Ethical dilemmas and privacy issues in emerging technologies: A review. *Sensors* **2023**, *23*, 1151. [CrossRef]
120. Marks, M.; Haupt, C.E. AI chatbots, health privacy, and challenges to HIPAA compliance. *JAMA* **2023**, *330*, 309–310. [CrossRef]
121. Franke, L.; Liang, H.; Farzanehpour, S.; Brantly, A.; Davis, J.C.; Brown, C. An Exploratory Mixed-Methods Study on General Data Protection Regulation (GDPR) Compliance in Open-Source Software. *arXiv* **2024**, arXiv:2406.14724.
122. Gonzalez-Colom, R.; Monterde, D.; Papa, R.; Kull, M.; Anier, A.; Balducci, F.; Cano, I.; Coca, M.; De Marco, M.; Franceschini, G.; et al. Toward Adoption of Health Risk Assessment in Population-Based and Clinical Scenarios: Lessons From JADECARE. *Int. J. Integr. Care* **2024**, *24*, 23. [CrossRef]
123. Alberto, I.R.I.; Alberto, N.R.I.; Ghosh, A.K.; Jain, B.; Jayakumar, S.; Martinez-Martin, N.; McCague, N.; Moukheiber, D.; Moukheiber, L.; Moukheiber, M.; et al. The impact of commercial health datasets on medical research and health-care algorithms. *Lancet Digit. Health* **2023**, *5*, e288–e294. [CrossRef]
124. Fu, L.; Fang, Y.; Dong, Y. The healthcare inequality among middle-aged and older adults in China: A comparative analysis between the full samples and the homogeneous population. *Health Econ. Rev.* **2022**, *12*, 34. [CrossRef]
125. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]