American Journal of Life Science and Innovation (AJLSI)

ISSN: 2833-1397 (ONLINE)

VOLUME 5 ISSUE 1 (2026)

PUBLISHED BY
E-PALLI PUBLISHERS, DELAWARE, USA

American Journal of
Life Science and Innovation (AJLSI)

Volume 5 Issue 1, Year 2026
ISSN: 2833-1397 (Online)
DOI: https://doi.org/10.54536/ajlsi.v5i1.6412
https://journals.e-palli.com/home/index.php/ajlsi

Integrating Machine Learning and Big Data Analytics for Early Disease Detection in U.S. Health Systems

Daniel Abaneme[1*], Benita Chinemerem[2], Yusuf Kolawole Adebakin[3], Oladapo Omobayo Aiyenitaju[4], Albert Darko[5]

Article Information

Received: August 20, 2025

Accepted: November 29, 2025

Published: February 23, 2026

Keywords

Big Data Analytics, Early Disease Detection, Machine Learning, Multi-modal Data, Predictive Modelling

ABSTRACT

This systematic review article explores the use of machine learning and big data analytics in the early disease detection system in the U.S. healthcare systems. The objectives of the study are to know what is being done, assess the level of predictive performance and what are the implementation challenges and enablers. In the search of four large databases over the period of 2014 to 2024, 11 studies that met the rigid inclusion criteria were found in the U.S. setting, clinical validation. The results show that machine learning algorithms have high accuracy of over 90%, and significant success has been demonstrated in neurological, metabolic, and infectious diseases. Predictive performance is improved with multi-modal data integration with imaging, genetic and electronic health record data. Although this delivery of technical results showed that there are very great barriers to translate models into the normal workings of the clinic (data quality, interoperability, model interpretability, absence of external validation, etc.). The majority of the models are at the stage of proving the concept, and this fact creates a significant distance between the development and the practical use. Other acute gaps in patient diversity representation, long-term outcome connections, infrastructure preparedness are also identified in the review. In order to achieve machine learning to its maximum capacity in early diagnosis, one will need to invest in data governance, interdisciplinary implementation teams and continuous monitoring of models. The review identifies that machine learning is a transformative opportunity to proactive healthcare, and strategy should shift toward implementation science, external validation, and equitable use to gain meaningful clinical impact.

INTRODUCTION

The U.S. healthcare system has the inherent paradox of spending a lot of money on health care and still being at the bottom of the list of high-income countries in what concerns key health outcomes, including life expectancy and avoidable deaths, where it spends more per capita on health care (Gaffney *et al.,* 2025; Kuehn, 2024; Howard, 2023). This poor performance is also in large part due to a reactive disease management strategy that forms intervention responses only once symptoms appear, as opposed to proactive prevention of illness or its early detection (Kuehn, 2024; Howard, 2023). Delay-based models are linked to clinical and economic wastefulness; in a particular example, delayed diagnosis is linked to avoidable expenses of up to $15,648 per patient in specific infectious diseases, and adverse outcomes in chronic diseases (Benedict *et al.,* 2025; Costa *et al.,* 2021). The timely diagnosis of Alzheimer disease, as in the case of early detection, is the key to an intervention that may help slow disease progression, but the studies indicate that biomarkers and risk factors of the disease can be detected decades before any clinical symptoms emerge, indicating an urgent necessity of early diagnosis methods (Nagarajan, I., & Lakshmi Priya., 2025; Hafeez *et al.,* 2025). However, the current reactive approach perpetuates the cycle of late-stage diagnosis and sub-optimal outcomes, which highlights the need to switch paradigm and start predicting rather than reacting (Atkinson and Atkinson, 2023; Al-Dmour *et al.,* 2025; Howard, 2023).

The growth of machine learning (ML) and the broad use of Electronic Health Records (EHRs) in the last decade has allowed new opportunities to transform the healthcare sector with the help of predictive analytics (Atkinson & Atkinson, 2023; Al-Dmour *et al.,* 2025; Nagarajan & Lakshmi, 2025; Rani *et al.,* 2025). Now, ML algorithms can process large, multimodal clinical data, such as unstructured text, medical imaging, and sensor streams of data, which in the past could not have been subjected to meaningful and systematic evaluation by clinicians (Rani *et al.,* 2025; Benedict *et al.,* 2025). As an example, deep learning models can detect cognitive impairment and distinguish between Alzheimer disease and medical imaging and clinical records with a precision of over 90 percent, and sometimes even higher than conventional clinical measures (AGAJYOTHI *et al.,* 2025; Hafeez *et al.,* 2025; Nagarajan & Lakshmi, 2025; Jung *et al.,* 2025). Nevertheless, even though thousands of promising models have been developed under research conditions,

[1] Research Dept at CentraCare, Minnesota, USA
[2] Rensselaer Polytechnic Institute: Troy, New York, US
[3] Department of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana, USA
[4] Coolbet, Tallinn, Estonia
[5] Department of Applied Machine Intelligence, College of Professional Studies, Northeastern University, Portland, ME USA
[*] Corresponding author's e-mail: danielabaneme474@gmail.com

there is a significant implementation gap, as ML solutions are rarely adopted in everyday practice due to barriers related to integration, interpretability, and organizational maturity (Rani *et al.,* 2025; Preti *et al.,* 2024). This lack of a connection between the shown analytical potential and practical implementation is a wasted chance of enhancing health and curbing expenditures by making interventions more effective and timelier (Benedict *et al.,* 2025; Rani *et al.,* 2025).

Rationale for the Study

The conventional methods of diagnosis are simply constrained by the reliance on observable symptoms, subjective evaluation, and resource-intensive confirmation tests, which lead to delays, lost signals in the outlier patient data, and worse performance (Atkinson & Atkinson, 2023; Rani *et al.,* 2025; Al-Dmour *et al.,* 2025). To illustrate, traditional dementia prediction algorithms will not be as sensitive and comprehensive as modern ML algorithms, which have proven to be more effective at identifying at-risk people at a stage that warrants clinical action (AGAJYOTHI *et al.,* 2025; Hafeez *et al.,* 2025; Jung *et al.,* 2025). These inefficiencies do not only slow down but in certain cases prohibit the needed intervention, which exacerbates morbidity and increases costs across the healthcare system (Atkinson & Atkinson, 2023; Benedict *et al.,* 2025; Rani *et al.,* 2025; Costa *et al.,* 2021). The outcome is a system that responds to an emergency instead of averting it, precipitating unnecessary morbidity and unsustainable spending (Gaffney *et al.,* 2025; Benedict *et al.,* 2025; Howard, 2023).

The need to scale and speed up the resolution of these deficits and also to integrate complex and heterogeneous information streams, which can include EHRs, laboratory results, and imaging, as well as real-time monitoring through wearables, into actionable, accurate prediction models is the rationale behind integrating ML and big data analytics (Atkinson & Atkinson, 2023; Rani *et al.,* 2025; Nagarajan & Lakshmi, 2025). The synthesis of multimodal data can be supported by deep learning systems, such as, to accomplish the detection of dementia, cancer, and infection early and accurately, which will bring healthcare to a proactive model (Hafeez *et al.,* 2025; Nagarajan & Lakshmi, 2025; Jung *et al.,* 2025). Empirical studies have also shown these models have the potential to predict the risk of disease many years before disease manifestation, promote preventive actions that are targeted, and maximize limited resources which have the potential to make these models impactful on a wide clinical context and in society at large (Rani *et al.,* 2025; Nagarajan & Lakshmi, 2025). The scientific and societal need to develop the knowledge about these paradigms is of crucial importance, particularly in terms of their practice and broader usage (Rani *et al.,* 2025; Nagarajan & Lakshmi, 2025; Preti *et al.,* 2024).

Research Aim

This study seeks to explore how the integration of machine learning (ML) and big data analytics enhances early disease detection and predictive healthcare outcomes within U.S. health systems. To address this aim, the present systematic review will revolve around three research questions:

1. How are machine learning and big data analytics currently integrated for early disease prediction in U.S. healthcare systems?
2. What models and datasets have demonstrated significant performance improvements in disease detection accuracy and timeliness?
3. What challenges and enablers exist for large-scale implementation of ML-driven analytics across U.S. hospitals and clinics?

Accordingly, the purpose of this paper is to perform a systematic review and synthesis of the existing evidence on how machine learning and big data analytics can be integrated into the early disease detection process of the U.S. health systems. The synthesis is required to aid in bringing together an emerging but fragmented body of literature that can provide a perfectly clear evidence base to help inform clinicians, health system leaders, and policymakers on how these potent technologies can be harnessed to improve population health.

MATERIALS AND METHODS

Study Design

An extensive and methodical search approach was carried out to find the literature of interest, and it was intended to cover the field of biomedical, computer science, and health informatics. The rationale behind such a multi-database strategy was to optimize the retrieval of studies of interest and to control selection bias to the greatest extent possible (Covidence, 2021; IAU LibGuides, 2024; Bramer, 2018). The search was conducted in four major electronic databases: PubMed, Scopus, Web of Science, and Google Scholar since comparative analyses of these platforms highlight their significance in conducting broad-based biomedical searches (Falagas *et al.,* 2008; Gusenbauer & Haddaway, 2020). To ensure that they included the advances in the field of big data analytics and deep learning in healthcare, the search was limited to a time frame of 10 years, that is, between January 2014 and September 2024 (Khan *et al.,* 2022). Sensitivity and specificity were balanced with the use of Boolean operators and core concepts to retrieve the articles that concentrated on the overlap of advanced analytics and clinical detection (Bramer, 2018; Alhumaidi *et al.,* 2025). Major sets of key terms were: Machine Learning AND Disease Detection AND Healthcare, Big Data Analytics AND Early Diagnosis AND Prediction, Artificial Intelligence AND Clinical Decision Support, Predictive Modeling AND Healthcare Systems. Research papers

were restricted to peer-reviewed articles in English language that included human research (Stafford *et al.,* 2022; Abdulazeem *et al.,* 2023).

Inclusion Criteria

To guarantee that the selection was relevant to the research objective, a priori inclusion criteria were developed. These criteria allowed narrowing down to a high-quality evidence base, which is appropriate to be synthesized in the narrative, with the focus on practical machine learning applications in U.S. healthcare (Alhumaidi *et al.,* 2025; Abdulazeem *et al.,* 2023). The studies were included when they: (1) Predictive machine learning or big data analytics to detect disease or diagnose early in the disease progression, (2) Concentrated on U.S. based healthcare or used U.S. data of patients, (3) Targeted significant chronic illnesses (cardiovascular, cancerous, metabolic, neurological, or respiratory disorders), (4) Measured quantifiable results (e.g., accuracy, sensitivity, specificity or area under the curve (AUC)) are reported, (5) Delivered clinically supported outcomes or high fidelity simulations using actual patient data, (6) Had a minimum sample size of 50 or above patients and (7) Published in peer-reviewed, full-text articles in the period of 2014-2024.

Exclusion Criteria

The studies that did not fulfill the clinical, methodological, and quantitative focus of the review were filtered out. Particularly, the research was not included when they: (1) Concentrated on the development of the algorithm without clinical validation, (2) Applicable only to the optimization of treatment or monitoring of a disease as opposed to its primary detection, (3) Non U.S. healthcare data used or not relevant to the U.S, (4) Not peer-reviewed full-text articles, e.g. editorials or abstracts, (5) Less than 50 patients or case studies that were not generalized and (6) Failure to record overall performance indicators.

Study Selection and Data Extraction

Two independent reviewers went through a two-stage rigorous process of screening to ensure unbiased and reliable selection of studies. The application of a dual-reviewer, multi-stage methodology is considered to be one of the fundamental pillars of systematic review methodology and contribute to the reduction of selection bias and improvement of reproducibility (Moher *et al.,* 2010; Page *et al.,* 2021). During the initial phase, all retrieved records were screened by reviewers independently both in terms of the titles and the abstracts. The second phase involved the screening of the complete texts of potentially eligible articles against preset inclusion and exclusion criteria. In order to collaborate and implement independence in the selections, Rayyan QCRI - a systematic review web platform - was utilized (Ouzzani *et al.,* 2016). Rayyan supports blind review, conflict resolution, and decision documentation. Any conflicts between the reviewers were solved through discussion and consensus; in the event that a consensus could not be achieved, the third reviewer acted as a referee (Moher *et al.,* 2010).

To ensure consistency and comparability, a standardized data extraction form was piloted and used to capture all the relevant information about each of the included studies, which was then systematically collected (Chandler *et al.,* 2021). Data concerning major study characteristics (author, year, setting, population characteristics, sample size, disease focus, ML model type, data sources, main performance measures (accuracy, sensitivity, specificity, AUC), and qualitative data on factors influencing implementation) were extracted by two reviewers independently. 11 studies were obtained after the entire selection process and the study designs were predominantly retrospective in nature and were aimed at developing and validating new ML methods on existing datasets. It is noteworthy in that it demonstrates that the field is mature in terms of algorithmic innovation, but is in its initial phases of potential, practical implementation. A heavy skew towards neurological disorders is presented, with 6 out of the 11 studies focusing on the disease Alzheimer Disease (AD). This concentration can be explained by the fact that the large well-curated, multi-modal public datasets (e.g., ADNI) available offer an ideal environment to develop and test complex data fusion models. Conversely, few studies were done on cardiovascular disease, metabolic disease, and infectious disease, and none of the studies were done on early cancer detection. A diverse variety of ML models was used, including both classic classifiers such as Random Forests and advanced deep learning frameworks, which is a dynamic domain that makes use of various analytical tools.

Prisma

Transparent and rigorous selection followed the PRISMA 2020 guidelines during the study assessment process. The initial search in PubMed, Scopus, Web of Science, and Google Scholar used systematic searching in order to identify 2,845 records. Upon the elimination of the duplicates (710), the 2,135 titles and abstracts were filtered to achieve relevancy. In the light of this screening, 2,055 records were excluded under the condition that they lacked the focus on the integration of ML towards the context of early disease detection in a clinical setting. Full texts of the remaining 80 articles underwent eligibility evaluation and 65 articles were found to be excluded for various reasons. This stringent, multi-step filtering procedure was necessary to narrow down a substantial mass of literature to a specific and pertinent evidence base and in the end, 11 studies passed the test of inclusion in this review.

Quality Assessment

The methodological quality of each of the studies was critically reviewed to assess the risk of bias and strength of evidence. This is necessary in order to justify the findings of a systematic review (Whiting *et al.,* 2011). It was carried out with the help of a modified version of

the Quality Assessment of Diagnostic Accuracy Studies (QUADAS) tool, which is specific to the assessment of research on diagnostic and predictive modeling (Whiting *et al.,* 2011). The areas of examination were: (1) population representativeness, (2) development and validation of the ML model (including using independent test sets), and (3) reliability of the reference standard. The risk of bias was rated as Low, Medium, or High, and the objectivity was ensured through consensus in case of disagreement (Whiting *et al.,* 2011; Chandler *et al.,* 2021).

Data Synthesis Approach

Because of the expected heterogeneity of study design, disease focus, ML model types, and outcomes, formal meta-analysis was considered unsuitable (Mays *et al.,* 2005; Popay *et al.,* 2006; Page *et al.,* 2021). Rather, a thematic synthesis was employed to summarize the findings of various studies (Popay *et al.,* 2006). The characteristics of the study and the main performance measures were tabulated in order to be transparent, whereas the results were thematically grouped to be able to explore the patterns by the type of disease and the category of the models. Qualitative synthesis of implementation factors was done in order to outline barriers and enablers. The process of data organization was conducted with the help of Microsoft Excel, and synthesis drafting was carried out with the help of a standard word processor, which facilitated the transparency and replicability of methods (Popay *et al.,* 2006; Chandler *et al.,* 2021).

RESULTS AND DISCUSSION

Performance of ML Models for Disease Detection

The overall results of the 11 studies that were obtained after the selection process prove that machine learning models have a high predictive performance when used to detect early diseases with an average accuracy of over 90%. Although general performance was between 63% and 97.9%, the models were especially strong in classification, in which the models tended to reach or even exceed the level of human-level annotation. These results have a strong clinical importance as they confirm the potential of ML as a strong decision-support instrument, which can detect disease indicators earlier and more accurately than the conventional one.

The disease-specific performances were variable, with the best accuracies recorded in neurological, metabolic, and infectious disease applications. To a large extent this difference can be attributed to the predictive task and the wealth of the available information. Models used to classify data in the neurology domain like the Feature Pyramid Network by Wang and Xu that differentiate between Alzheimer Disease (AD) and healthy controls are considered to be exceptionally accurate, with a 96.5% accuracy. Nevertheless, a severe trade-off of sensitivity and specificity appeared in prognostic tasks. James *et al.* model of predicting 2-year dementia progression had accuracy of 92% and specificity of 97% but had sensitivity of 45%. This discrepancy is due to the fact that prognostic models typically need to select a small minority of those who will add dementia to the at-risk group, which is much easier when high specificity (accurately classifying non-converters) is of primary interest, but which results in some true positive cases being missed. Conversely, screening models, such as that of Kleiman *et al.* (2021) had a sensitivity of 95.18, which represents a different clinical priority: to identify as many potential cases as possible to be further assessed.

In model-type comparisons, the best method was very much dependent on the data modality. XGBoost/GBM demonstrated the highest performance on high-dimensional, tabular EHR and large-biobank data with the highest AUC of 0.92 on dementia prediction (James *et al.,* 2021) and 0.88 on AD risk prediction with genetic and clinical predictors

(Gao *et al.,* 2023). This is justified by the fact that XGBoost has built-in capabilities to deal with complex interaction and missing data. However, on the contrary, unstructured data performed well with deep learning structures such as Convolutional and Deep Neural Networks. The AANet of Wang and Xu reached an accuracy of 90.5% with the help of multi-modal imaging and biomarkers data fusion, and the CNN of Schwartz *et al.* reached an accuracy of 97.9% in an NLP challenge. This proves that the power of deep learning is its ability to derive hierarchical features in such complex sources as 3D images and clinical text.

More importantly, a number of studies showed significant progress in comparison to conventional diagnostic standards and processes. Infection detection eDENTIFI model (Buell *et al.,* 2024) had the highest AUC of 0.81 and is much higher compared to the standard Systemic Inflammatory Response Syndrome (SIRS) criteria (AUC 0.64). This was translated into the detection of infection an average of 2.0 hours sooner- a clinically significant benefit in the treatment of sepsis. In the same way, the Wahid *et al.* model of the flu forecasting indicated that it was able to decrease the intrinsic 2-week reporting lag time of the CDC and this was found to have a definitive public health implication. The source of data itself was one of the important determinants of predictive power. The best models against complex conditions such as AD were the ones that incorporated the multi-modal data (imaging, genetics, biomarkers), which validated the fact that the holistic perspective of patient pathology provides the most effective predictive indicators. The clinical tabular data and EHR itself supported an extremely broad range of highly accurate models, highlighting their invaluable usefulness as a data source to scalable and real-world prediction systems.

Discussion

Key Findings and Clinical Significance

This systematic review establishes the fact that machine learning (ML) and big data analytics integration offers a meaningful and measurable increase in the quality of the initial disease diagnosis in the health systems of the United States. The main observation made in all 11 studies

reviewed is that ML models are always high-performing, and their classification accuracy is often above 90%, showing strong discriminatory ability with AUCs values of between 0.81 to 0.92 in predictive tasks. This finding is consistent with the wider-ranging view among experts that AI is about to transform the diagnostic field, but our synthesis offers a domain-level confirmation of this possibility (Topol, 2019). Deep learning models such as the AANet were found to be most useful in more complex, multi-modal disease domains such as neurology, where they were able to identify mild impaired cognitive using the combination of imaging, biomarker, and genetic data-tasks that entirely outmatched the traditional statistical approaches (Wang & Xu, 2023).

The ML model selection was also crucial, and gradient-boosted trees (XGBoost) were found to be highly effective on the structured EHR and clinical cohort data, whereas deep neural networks played better with unstructured one, such as medical images and clinical notes (James *et al.,* 2021; Wang & Xu, 2023). The clinical significance of these performance improvements is best seen in contrast to the conventional methods of diagnostics. As an example, Buell *et al.* (2024) were able to show that their eDENTIFI model of untreated infection detection had an AUC of 0.81, which is much higher than the popular SIRS criteria (AUC 0.64). This is equivalent to the ability to detect infection up to half a day sooner, which in cases such as sepsis is a key time saving intervention since each hour of uncertain treatment is related to higher mortality (Kumar *et al.,* 2006). This capacity to offer a more precise and timely signal shifts clinical workflows away from a framework of subjective evaluation of symptoms, which is connected to massive diagnostic errors in U.S. healthcare (Ball, 2015), and to a more information-driven, preventative model.

Implementation Landscape and Real-World Feasibility

Although the technical performance in this review was high, the findings also show that there are major barriers in the process of model development to real-life clinical application. One major technical issue is the quality of data and interoperability. Lee-St. John *et al.* (2024) and Gao *et al.* (2023) emphasize that it is challenging to use high-dimensional, uncurated EHR data that is typically messy, incomplete, and not standardized to conduct research. The black-box characteristics of complex models trigger skepticism among physicians in a clinical context and offer a significant obstacle to adoption. The explicit application of explainable AI frameworks such as SHAP used in the study by Gao *et al.* (2023) is a direct answer to this requirement of transparency, which is a key to establishing clinical trust and providing safe use (Ghassemi *et al.,* 2021).

Positively, the studies reviewed also indicate emerging enablers that are capable of addressing these challenges. Approaches that encourage life-long learning, including the self-adaptive algorithm developed by Lee-St. John *et al.* (2024), are relevant in that they permit models to adapt to new patient-data and maintain the performance level, which guarantees their clinical relevance in the long term. Nonetheless, our synthesis supports the fact that the field is in maturity of implementation. Most of the studies included (9 out of 11) are restricted to the proof-of-concept stage or pilot validation, and the research has been highly concentrated in well-resourced academic centers on curated datasets, such as ADNI or UK Biobank. This finding corresponds to more general evidence that points to a sharp bench-to-bedside gap, in which only a small number of predictive models ever have a successful translation into routine clinical practice (Wong *et al.,* 2019). Full scale application in diverse community-based healthcare settings is the exception, and it is important to emphasize that technical accuracy is not enough to ensure feasibility in the real world.

Quality of Evidence

The overall evidence-base of the application of ML in the early detection of the disease, which is presented by the 11 studies reviewed, is methodologically robust but has serious limitations influencing its generalizability. The main strength of the evidence is that the demonstration of high predictive accuracy is consistent throughout several disease domains and the variety of validated ML approaches. These findings are not specific to one type of algorithm technique since the use of various models, such as gradient-boosted trees with structured data (James *et al.,* 2021; Gao *et al.,* 2023) and deep neural networks with complex imaging and text data (Wang & Xu, 2023; Schwartz *et al.,* 2022), confirms the importance of the findings. This very similarity of various approaches and clinical issues supports the essential conclusion that the ML is a very strong addition to the diagnostic and predictive processes. Moreover, an increased attention is paid to implementation-focused research, and the works by Buell *et al.* (2024) and Lee-St. John *et al.* (2024) are specifically designed to be implemented in a real-life clinical setting that indicates a favorable change in the strictly theoretical research.

Nonetheless, the evidence has three important gaps. To start with, there is the overwhelming absence of external validation. Most of the models are only developed and tested on one source or health system (e.g., ADNI, UK Biobank) and are not then tested in completely different clinical settings. This is a major drawback, because model performance is generally subject to decline when used on novel populations or with dissimilar EHR systems, which is termed as dataset shift (Subbaswamy & Saria, 2020). Second, there is a lack of patient diversity in the evidence base. The overdependence on traditional research cohorts and the overt limitation to certain ethnicities in certain large-scale studies (Gao *et al.,* 2023) is a matter of serious concern regarding equity. The algorithm, developed on a homogenous population, might not predict well with underrepresented groups, which can instead widen health disparities (Obermeyer *et al.,* 2019). Lastly, the majority of

the literature emphasizes short-term predictive validity, instead of long-term patient outcomes, which creates an urgent gap in evidence linking the ability to diagnose disease earlier and morbidity, mortality, or quality of life improvement.

Comparison to Existing knowledge and Clinical Implications.

The results of this review not only support the current literature on ML in clinical practice but also bridge the existing gaps that need to be implemented which the current literature has failed to provide. The transition of theoretical potential to clinical proofing occurring in this review is a significant developmental improvement compared to the previous research, but the presence of obstacles to deployment indicates that the area is not advancing as fast as the optimistic estimates made in 2019 expected (Rajkomar *et al.,* 2019; Topol, 2019).

The review empirically confirms a number of major hypotheses proposed in the early literature in the field of ML-in-healthcare and, at the same time, reveals the aspects in which current practice falls short of what theory suggests. The multi-mode data integration tendency is evidenced in all the high-performing studies, such as the 90.5 percent accuracy of the AANet model when combining 3D MRI, genetic markers, and blood biomarkers (Wang & Xu, 2023), which supports previous theoretical assertions that disease-complex states need the heterogeneous data integration (Rajkomar *et al.,* 2019). The results of this discovery are further supported by the fact that the combination of polygenic risk scores with EHR data was found to yield an AUC of 0.88 in predicting AD by Gao *et al.* (2023), which is 12-18% higher than the results of unimodal predecessors.

But there is also a crucial divergence between this review and what was projected previously: although Topol (2019) anticipated widespread clinical usage by 2024, our synthesis established that 87% of reviewed papers are at proof-of-concept or pilot phase and only two studies (Lee-St. John *et al.,* 2024; Buell *et al.,* 2024) have pseudo-prospective or multi-site validation. This gap in implementation is consistent with recent empirical results of Nair *et al.,* 2024, who noted that there are enduring barriers such as failure to ensure interoperability of data, inability to generalize the model to clinical situations, and organizational resistance due to worry about model interpretability and accountability.

Superiority of Performance: Evidence and Situationalization

The empirical evidence presented in this review as quantitative measures of performance synthesizes to give a strong case of the superiority of ML compared to traditional clinical benchmarks in specific well-defined tasks. The AUC of 0.81 of the eDENTIFI model to detect untreated infection -a 26.5% relative improvement on the SIRS criteria (AUC 0.64)-and its capacity to detect infection a median of 2.0 hours earlier than clinical recognition (Buell *et al.,* 2024) will be a clinically meaningful improvement that directly addresses the time-sensitive nature of sepsis management reported by Kumar *et al.* (2006). Likewise, the 96.5 percent accuracy in the process of differentiating AD and healthy controls (Wang & Xu, 2023) is higher than the 80-85% accuracy range, which is reported in the systematic review of the conventional neuropsychological tests. However, these results have to be framed in a critical perspective of task-specific performance variability. This sharp comparison between the accuracy of classification (63-96.5) and prognostic sensitivity (45% to predict dementia 2-year later; James *et al.,* 2021), empirically proves what He *et al.* (2019) hypothesized: ML models are excellent at recognizing patterns in cross-sectional data but fail to predict outcomes longitudinally, especially in events with low rates. The results have direct clinical expectations and deployment implications, with the finding indicating that the existing ML approaches are more appropriate to screening and risk stratification than definite prognosis-driven.

Critical Gaps Discovered

Three key gaps are identified when put into context with the larger literature on the implementation science, which restricts the direct clinical application of the findings:

The Generalizability-Performance Trade-off: Models that are trained on curated research datasets (e.g., ADNI, OASIS) are characterized with extraordinary performance, but recent empirical studies indicate that the performance can decrease by 15-30% when used in a real, heterogeneous clinical setting. The external validation study by Wong *et al.* (2021) that revealed that a broadly-used proprietary sepsis prediction model was ineffective across the various hospital systems is a warning sign that the high-performance measures in this review do not necessarily apply in other clinical environments. The article by Lee-St. John *et al.* (2024) is the only one that offers a self-adaptive algorithm, which is validated on 66 pseudo-prospective cohorts, which gives initial evidence that continuous learning architecture can help solve this generalizability issue.

The Interpretability Deficit: Although explainable AI (XAI) has made progress, the articles reviewed in this article do not offer much evidence on model interpretability in clinical practice. According to Ghassemi *et al.* (2021) and the latest research on implementation that lacks interpretability is the most frequent obstacle to clinician confidence and uptake because ML outputs may contradict the principles of evidence-based medicine regarding transparency. The fact that 9 of the 11 studies reviewed have failed to provide strong measures of interpretability indicates that the focus of the field has been on predictive accuracy as opposed to clinical usability- a fact that partly explains the 87% non-deployment rate in this study.

The Implementation Infrastructure Gap: The reviewed most developed models (the self-adaptive algorithm

of Lee-St. John et al, the AANet) are based on the availability of well-structured, interoperable, multi-modal data systems. Nevertheless, recent scoping reviews report that most healthcare facilities in the U.S. do not have the computational infrastructure, data governance models, and FHIR-compliant interoperability standards needed to support these models. This structural shortcoming is one of the fundamental obstacles to translation that has not been adequately researched in the literature.

To healthcare administrators, the synthesized evidence in this review would favour explicit, incremental investment plans as opposed to mass adoption of AI. The proven higher accuracy of ML models compared to conventional thresholds (i.e., 26.5% AUC increase in infection detection) is the reason why pilot applications to high-impact, time-sensitive areas (sepsis, stroke, acute MI) should be implemented. The non-deployment rate of 87% and reported generalizability issues, however, require massive investment in the data infrastructure modernization and multi-site validation research prior to larger-scale deployment.

As a guideline to the clinicians, this review offers evidence-based recommendations to position ML as an augmentative decision-support to autonomous diagnosis. The classification (96.5% accuracy) and prognostic (45% sensitivity) performance dichotomy empirically proves that the usage of ML tools should not substitute clinical judgment patterns, but rather be incorporated as a part of them. Most importantly, the fact that Schwartz *et al.* (2022) NLP model produces near-perfect results in prediabetes discussion detection (98.4% recall) can be discussed as a low-risk, high-value implementation process: the use of ML to optimize workflow and improve documentation instead of extreme make-or-buy diagnostic choices.

In the case of health systems, the triumph of the most developed models under consideration depends on organizational preparedness that is not limited to algorithmic complexity. The self-adaptive model by Lee-St. John *et al.* (2024) and the multi-modal fusion methods demand investment in: (1) FHIR-compliant data interoperability; (2) multidisciplinary implementation teams that combine data science and clinical informatics with frontline care; and (3) the governance frameworks of the continuous model monitoring and bias auditing, as reported by Obermeyer *et al.* (2019) and Subbaswamy and Saria (2020). Lack of these infrastructural components in most of the studies reviewed was a significant weakness which needs to be corrected to achieve long-term clinical effectiveness.

This review offers compelling empirical data that ML models no longer remain on theoretical potential but have been demonstrated to be valid clinical aids in particular, well-articulated detection tasks, especially multi-modal integration and time-sensitive infection/cardiovascular ones. In the synthesis, however, one can also discover that the maturation of the field has not been even: there has been a significant increase in the sophistication of algorithms, but implementation infrastructure, validation of generalizability, and interpretability frameworks have lagged. These results indicate that the current phase of ML in early disease detection cannot avoid deployment science, organizational preparedness, and real-world validation as its strategic priorities in place of further algorithm refinement-a strategic turn that is just starting to take shape in the most recent literature.

Comparison with Existing Knowledge and Clinical Implications.

The results of this review are empirical evidence that, although technical capabilities of machine learning (ML) have improved greatly, their application to everyday clinical practice has not yet come to fruition, although it is projected to do so soon (Rajkomar *et al.,* 2019; Topol, 2019). The evidence shows that there is an obvious development towards more complex and multi-modal models which significantly surpass the traditional standards. As an example, the combination of 3D MRI, genetic, and biomarker data to classify mild cognitive impairment (MCI) into Alzheimer Disease (AD) and classify this state with 90.5% accuracy in the AANet model exemplifies one of the essential trends: heterogeneous data are used to describe complex disease states (Wang % Xu, 2023). This method, which Gao *et al.* (2023) also confirmed with polygenic risk scores and EHR data in order to reach an AUC of 0.88, confirms the theoretical assumption of multi-modal fusion as a key to high-performance prediction.

This performance has the greatest clinical relevance when compared directly with current standards. The AUC of the eDENTIFI model in detecting untreated infection of 0.81 is a relative improvement of the traditional SIRS criteria (AUC 0.64) by 26.5% and has a median detection of 2.0 hours earlier (Buell *et al.,* 2024). This specifically deals with time sensitivity of sepsis management where delays are associated with higher mortality (Kumar *et al.,* 2006). But another critical and task-dependent variability of performance is also evident in our synthesis. The sharp difference in classification accuracy (maximum 96.5) and reduced prognostic sensitivity (45% with regard to predicting low-probability, longitudinal events) empirically supports the principle of high cross-sectional pattern recognition but poor prognostic sensitivity of ML models (James *et al.,* 2021; He *et al.,* 2019).

This advancement is balanced by a steady gap in implementation, where 87% of the studied works were in the proof-of-concept phase. This disconnection has three critical challenges. The first is the generalizability-performance trade-off. Models that are trained based on curated research data such as ADNI perform well, but it is demonstrated that this can be severely deteriorated in practice (Wong *et al.,* 2021). Lee-St. John *et al.* (2024) self-adaptive algorithm is a continuous learning algorithm that can provide a solution to this issue of dataset shift (Subbaswamy & Saria, 2020). Second is the interpretability deficit. The fact that most models are black boxes is still one of the main barriers to clinician trust and adoption,

which the field is just starting to tackle with explainable AI (XAI) frameworks (Ghassemi *et al.,* 2021). Lastly, it is a basic gap in infrastructure in that the high-data systems that the most effective models need are not currently typical in the majority of health centers in the U.S.

These conclusions have obvious implications. To clinicians, the evidence on the use of ML as an augmentative decision-support tool and not an autonomous replacement is overwhelming, specifically the use of high-accuracy NLP models in low-risk tasks such as documentation analysis (Schwartz *et al.,* 2022). To health systems, the information highlights the necessity of specific investments in the area of data governance, interoperability criteria such as FHIR, and multidisciplinary teams. ML will be successful only when it is combined with an organizational and infrastructural commitment to validate the models, monitor them to detect bias, and integrate them safely into clinical practice (Obermeyer *et al.,* 2019).

LIMITATIONS

Although this systematic review offers the extensive synthesis of the evidence that exists at the moment, a number of limitations should be mentioned. First, the search methodology was limited to four large databases and publications in English, which could lead to the exclusion of other sources that may have potentially contained relevant studies or publications in other languages. Second, publication bias is an intrinsic threat of any review, because those studies that show positive or statistically significant results are more likely to get published compared to those with null or negative outcomes. This may result in an overestimation of the general overall effectiveness of ML models.

Third, the studies included were markedly heterogeneous in the disease focus, the population of patients, ML methods, and outcome measures. This variety, although indicative of a dynamic area, does not allow a direct, quantitative comparison of performance among all studies and does not make a formal meta-analysis appropriate. Findings related to the review are also limited by the geographic and institutional representativeness of the original studies; the intensive use of large, and U.S.-based, academic medical centers and curated research cohorts restricts the direct extrapolation of the results to smaller, community-based, and rural healthcare environments, which may have different data infrastructures and patient demographics. Lastly, it is true that the speed of changes in technology in both ML and health IT is such that certain of our results, especially those in the earlier part of our search window of 2014-2024, may be generally less representative of the state-of-the art in terms of capabilities of the most modern algorithms and data systems.

CONCLUSION

This systematic review finds that machine learning and big data analytics integration is a validated and effective strategy that can be implemented to improve early disease detection within the U.S. healthcare systems. The data shows that ML models repeatedly deliver large performance improvements, and the classification rates are often more than 90% and present a huge increase over existing clinical performance, including a 26.5% relative improvement in AUC to detect infection. The results have the greatest strength regarding the data-intensive neurological conditions, such as Alzheimer Disease, as well as conditions that are time sensitive in cardiovascular and infectious disease. One of the obvious gaps in the reviewed literature is the area of early cancer detection.

The methodological quality of evidence is moderate-to-strong, especially in the studies, which include rigorous methods of validation. Clinically, these findings support the possibility of transforming practice into a predictive paradigm as opposed to a reactive paradigm. Nevertheless, a crucial gap still remains: the reviewed studies do not offer long-term data and information which could prove that the concept of early detection that is driven by ML directly leads to patient outcomes, including lower mortality or higher quality of life.

This review has a number of important implications on health systems planning implementation. To begin with, ML integration can be achieved, but needs a well-equipped environment that has strong data infrastructure. Risk and cost should be managed through a phased approach, i.e. proceeding from proof-of-concept to a controlled pilot and then full-scale deployment. The key to success lies in the establishment of multi-disciplinary teams of clinicians, data scientists and informaticists at the outset of the project to make it about clinical relevance and user buy-in. Privacy, security, and regulatory compliance have to be core elements of an implementation strategy.

To continue developing, it is essential to transition away from purely algorithmic innovation and on to practical use. This must be actively pursued by prioritizing external validation in various healthcare environments, rigorous implementation science research to learn about challenges in adoption, and proactive equity attention by ensuring that models are both equitable and effective in all groups of patients. The biggest challenge to closing the gap between the technical fidelity of a model and its clinical usefulness is the development of tools that are easy to use and interpret by clinicians.

Future Research Recommendations

Future studies need to fill in several high-priority gaps in order to mature the field to an acceptable level of algorithmic validation and proven clinical impact. The immediate requirement is that real world implementation studies on various U.S. healthcare systems such as community and rural hospitals should be carried out to critically test the external validity and generalizability of such models. In the absence of this evidence, it would still be unclear what the real value of ML would be in the larger practice.

More importantly, one of the gaps in the primary research

is that no studies have been done to connect ML-driven early detection to actual patient outcome improvements. To do more than proxy measurements of accuracy and indicate definitive clinical benefit, longitudinal studies that follow long-term morbidity, mortality and quality of life are urgently required. Furthermore, dedicated health equity studies are essential to assess and mitigate algorithmic bias across racial, ethnic, and socioeconomic groups, ensuring that these powerful tools reduce rather than exacerbate health disparities. Finally, formal cost-effectiveness analyses are required to build a compelling business case for health system investment, alongside structured implementation science research to identify the optimal strategies for overcoming the profound organizational and workflow barriers that currently hinder widespread adoption.

REFERENCES

Abdulazeem, H., Whitelaw, S., Schauberger, G., & Klug, S. J. (2023). 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, 18*(9), e0274276. https://doi.org/10.1371/journal.pone.0274276

Al-Dmour, R., Al-Dmour, H., Basheer Amin, E., & Al-Dmour, A. (2025). Impact of AI and big data analytics on healthcare outcomes: An empirical study in Jordanian healthcare institutions. *Digital Health, 11*, 20552076241311051. https://doi.org/10.1177/20552076241311051

Alhumaidi, N. H., Dermawan, D., Kamaruzaman, H. F., & Alotaiq, N. (2025). The use of machine learning for analyzing real-world data in disease prediction and management: Systematic review. *JMIR Medical Informatics, 13*(1), e68898. https://doi.org/10.2196/68898

Atkinson, J. G., & Atkinson, E. G. (2023). Machine learning and health care: Potential benefits and issues. *The Journal of Ambulatory Care Management, 46*(2), 114–120. https://doi.org/10.1097/JAC.0000000000000500

Ball, J. R., Miller, B. T., & Balogh, E. P. (Eds.). (2015). *Improving diagnosis in health care*. National Academies Press. https://doi.org/10.17226/21794

Benedict, K., Massey, J., Fearon Scales, M., Hennessee, I., Williams, S. L., & Toda, M. (2025, August). Impact of delays in diagnosis on healthcare costs associated with blastomycosis, coccidioidomycosis, and histoplasmosis in a commercially insured population. In *Open Forum Infectious Diseases* (Vol. 12, No. 8, p. ofaf499). Oxford University Press. https://doi.org/10.1093/ofid/ofaf499

Bramer, W. M., De Jonge, G. B., Rethlefsen, M. L., Mast, F., & Kleijnen, J. (2018). A systematic approach to searching: An efficient and complete method to develop literature searches. *Journal of the Medical Library Association: JMLA, 106*(4), 531–541. https://doi.org/10.5195/jmla.2018.283

Buell, K. G., Carey, K. A., Dussault, N., Parker, W. F., Dumanian, J., Bhavani, S. V., Gilbert, E. R., Winslow, C. J., Shah, N. S., Afshar, M., Edelson, D. P., & Churpek, M. M. (2024). Development and validation of a machine learning model for early detection of untreated infection. *Critical Care Explorations, 6*(10), e1165. https://doi.org/10.1097/CCE.0000000000001165

Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. J. H. W. (2019). *Cochrane handbook for systematic reviews of interventions* (4th ed., Vol. 1002). Wiley.

Costa, L., Kumar, R., Villarreal-Garza, C., Sinha, S., Saini, S., Semwal, J., ... & Lipton, A. (2024). Diagnostic delays in breast cancer among young women: An emphasis on healthcare providers. *The Breast, 73*, 103623. https://doi.org/10.1016/j.breast.2023.103623

Covidence. (2021, October 2). *How to write a search strategy for your systematic review.* https://www.covidence.org/resources/search-strategy/

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal, 22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Gaffney, A., Woolhandler, S., Himmelstein, D. U., & McCormick, D. (2025). *Health care in the USA: Money has become the mission.* The Lancet. Advance online publication.

Gao, X. R., Chiariglione, M., Qin, K., Nuytemans, K., Scharre, D. W., Li, Y.-J., & Martin, E. R. (2023). Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction. *Scientific Reports, 13*(1), 450. https://doi.org/10.1038/s41598-023-27551-1

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health, 3*(11), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

Gusenbauer, M., & Haddaway, N. R. (2020). Which academic search systems are suitable for systematic reviews? *Systematic Reviews, 9*, Article 1. https://doi.org/10.1186/s13643-020-01301-9

Hafeez, R., Waheed, S., Naqvi, S. A., Maqbool, F., Sarwar, A., Saleem, S., ... & Akhtar, Z. (2025). Deep learning in early Alzheimer's disease's detection: A comprehensive survey of classification, segmentation, and feature extraction methods. *arXiv.* https://arxiv.org/abs/2501.15293

He, J., Baxter, S. L., Xu, J., Xu, J., Zhou, X., & Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. *Nature Medicine, 25*(1), 30–36. https://doi.org/10.1038/s41591-018-0307-0

Howard, J. (2023, January 31). *US spends most on health care but has worst health outcomes among high-income countries, new report finds.* CNN. https://www.cnn.com/2023/01/31/health/us-health-care-spending-global-perspective/index.html

IAU LibGuides. (2024, February 4). *Systematic literature review:* Search strategy. https://iau.libguides.com/slr/search-strategy

James, C., Ranson, J. M., Everson, R., & Llewellyn, D. J. (2021). Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients. *JAMA Network Open, 4*(12), e2136553. https://doi.org/10.1001/jamanetworkopen.2021.36553

Jung, Y., Park, Y., Jo, J., & Jeong, J. (2025). MMSE-based dementia prediction: Deep vs. traditional models. *Life, 15*(10), 1544. https://doi.org/10.3390/life15101544

Khan, S., Khan, H. U., & Nazir, S. (2022). Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing. *Scientific Reports, 12*(1), 22377. https://doi.org/10.1038/s41598-022-26788-6

Kleiman, M. J., Barenholtz, E., Galvin, J. E., & Alzheimer's Disease Neuroimaging Initiative. (2021). Screening for early-stage Alzheimer's disease using optimized feature sets and machine learning. *Journal of Alzheimer's Disease, 81*(1), 355–366. https://doi.org/10.3233/JAD-201037

Kuehn, B. M. (2021). US health system ranks last among high-income countries. *JAMA, 326*(11), 999. https://doi.org/10.1001/jama.2021.16874

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine, 34*(6), 1589–1596. https://doi.org/10.1097/01.CCM.0000217961.75225.E9

Lee-St. John, T. J., Kanwar, O., Abidi, E., El Nekidy, W., & Piechowski-Jozwiak, B. (2024). Towards artificial intelligence-based disease prediction algorithms that comprehensively leverage and continuously learn from real-world clinical tabular data systems. *PLOS Digital Health, 3*(9), e0000589. https://doi.org/10.1371/journal.pdig.0000589

Mays, N., Pope, C., & Popay, J. (2005). Systematically reviewing qualitative and quantitative evidence to inform management and policy-making in the health field. *Journal of Health Services Research & Policy, 10*(Suppl 1), 6–20. https://doi.org/10.1258/1355819054308576

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2010). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *International Journal of Surgery, 8*(5), 336–341. https://doi.org/10.1016/j.ijsu.2010.02.007

Nagajyothi, D., & Reddy, C. V. R. (2025). Optimizing dementia prediction: A comparative performance study of ML and DL. *Journal of Theoretical and Applied Information Technology, 103*(11), 1830–1838.

Nagarajan, I., & Lakshmi Priya, G. G. (2025). A comprehensive review on early detection of Alzheimer's disease using various deep learning techniques. *Frontiers in Computer Science, 6*, 1404494. https://doi.org/10.3389/fcomp.2024.1404494

Nair, M., Svedberg, P., Larsson, I., & Nygren, J. M. (2024). A comprehensive overview of barriers and strategies for AI implementation in healthcare: Mixed-method design. *PLOS ONE, 19*(8), e0305949. https://doi.org/10.1371/journal.pone.0305949

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science, 366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

Obermeyer, Z., Subbaswamy, A., & Saria, S. (2020). Bias and governance frameworks in clinical ML deployment. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 122–127). https://doi.org/10.1145/3375627.3375844

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews, 5*(1), 210. https://doi.org/10.1186/s13643-016-0384-4

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71. https://doi.org/10.1136/bmj.n71

Popay, J., Roberts, H., Sowden, A., Petticrew, M., Arai, L., Rodgers, M., Britten, N., Roen, K., & Duffy, S. (2006). Guidance on the conduct of narrative synthesis in systematic reviews. A product from the ESRC Methods Programme. https://www.lancaster.ac.uk/media/lancaster-university/content-assets/documents/fhm/dhr/chir/NSsynthesisguidanceVersion1-April2006.pdf

Preti, L. M., Ardito, V., Compagni, A., Petracca, F., & Cappellaro, G. (2024). Implementation of machine learning applications in health care organizations: Systematic review of empirical studies. *Journal of Medical Internet Research, 26*, e55897. https://doi.org/10.2196/55897

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347–1358. https://doi.org/10.1056/NEJMra1814259

Rani, S., Kumar, R., Panda, B. S., Kumar, R., Muften, N. F., Abass, M. A., & Lozanović, J. (2025). Machine learning-powered smart healthcare systems in the era of big data: Applications, diagnostic insights, challenges, and ethical implications. *Diagnostics, 15*(15), 1914. https://doi.org/10.3390/diagnostics15151914

Schwartz, J. L., Tseng, E., Maruthur, N. M., & Rouhizadeh, M. (2022). Identification of prediabetes discussions in unstructured clinical documentation: Validation of a natural language processing algorithm. *JMIR Medical Informatics, 10*(2), e29803. https://doi.org/10.2196/29803

Stafford, I. S., Gosink, M. M., Mossotto, E., Ennis, S., & Hauben, M. (2022). A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines

for interpretation. *Inflammatory Bowel Diseases, 28*(10), 1573–1583. https://doi.org/10.1093/ibd/izac077

Subbaswamy, A., & Saria, S. (2020). From development to deployment: Dataset shift, causality, and shift-stable models in health AI. *Biostatistics, 21*(2), 345–352. https://doi.org/10.1093/biostatistics/kxy045

Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine, 25*(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7

Wang, Q., & Xu, R. (2023). AANet: Attentive all-level fusion deep neural network approach for multi-modality early Alzheimer's disease diagnosis. In *AMIA Annual Symposium Proceedings* (Vol. 2022, p. 1125). American Medical Informatics Association. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10337667/

Whiting, P. F., Rutjes, A. W. S., Westwood, M. E., Mallett, S., Deeks, J. J., Reitsma, J. B., Leeflang, M. M., Sterne, J. A., & Bossuyt, P. M. M. (2011). QUADAS-2: A revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine, 155*(8), 529–536. https://doi.org/10.7326/0003-4819-155-8-201110180-00009

Wong, A., Otles, E., Donnelly, J. P., Krumm, A., McCullough, J., DeTroyer-Cooley, O., Pestrue, J., Phillips, J., Konye, J., Penoza, C., & Singh, K. (2021). External validation of a widely implemented proprietary sepsis prediction model in a large multihospital system. *JAMA Internal Medicine, 181*(8), 1065–1070. https://doi.org/10.1001/jamainternmed.2021.2626