

Artificial Intelligence for Rare Disease Diagnosis: Machine Learning Performance, Predictive Modelling, and Clinical Translation



Mohammad Asaduzzaman ^{1*}

Abstract

Artificial intelligence (AI) is increasingly being considered a potential turning point in rare disease research—although its role, at least for now, remains somewhat complex and not entirely settled. Rare diseases, despite affecting millions globally, continue to present significant diagnostic challenges, often driven by fragmented data, limited clinical familiarity, and inherently small, heterogeneous patient populations. In this narrative review, we synthesize current evidence on the application of artificial intelligence—particularly machine learning and deep learning—in improving diagnostic accuracy, predictive modelling, and clinical translation in rare diseases. Across the literature, AI-based models demonstrate a notable capacity to identify subtle disease patterns, particularly in imaging and high-dimensional datasets. Deep learning approaches frequently outperform traditional methods in pattern recognition tasks, while multimodal machine learning frameworks provide a more integrated understanding of disease mechanisms. Still, these outcomes are not entirely consistent. Model performance varies—sometimes substantially—depending on dataset size, diversity, and validation strategies. Smaller or less representative datasets, in particular, may produce overly optimistic estimates of diagnostic accuracy, raising concerns about generalizability. What becomes increasingly apparent is that artificial intelligence is not a

standalone solution, but rather a data-dependent tool whose effectiveness is closely tied to data quality and validation rigor. While AI shows strong potential to reduce diagnostic delays and enhance clinical decision-making in rare diseases, meaningful clinical translation will likely depend on improved validation, transparency, and integration into real-world healthcare systems. **Keywords:** Artificial intelligence; rare diseases; diagnostic accuracy; machine learning; deep learning; clinical decision support; predictive modeling

Keywords

Artificial intelligence; rare diseases; diagnostic accuracy; machine learning; deep learning; clinical decision support; predictive modeling

1. Introduction

Rare diseases, by their very nature, occupy a paradoxical space in global health—each condition is uncommon, yet collectively they affect hundreds of millions of individuals worldwide. This duality creates a persistent tension in clinical practice: while the burden is substantial, the evidence base often remains fragmented, sparse, and unevenly distributed. For many patients, the journey toward diagnosis is neither linear nor timely; rather, it unfolds over years, sometimes decades, marked by

Author Affiliation: ¹Department of Biochemistry and Molecular Biology, Noakhali Science and Technology University, Noakhali-3814, Bangladesh.

Please cite this article as: Asaduzzaman, M. (2024). "Artificial Intelligence for Rare Disease Diagnosis: Machine Learning Performance, Predictive Modelling, and Clinical Translation", Bioinfo Chem, 6(1), 1-10, 10584

© 2024 Bioinfo Chem, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(<http://creativecommons.org/licenses/by-nc-nd/4.0/>)
(<https://publishing.emanresearch.org>)

Significance | AI reframes rare disease diagnosis by enabling earlier detection, but highlights the critical need for robust data, validation, and ethical implementation.

*Correspondence.

Mohammad Asaduzzaman, Department of Biochemistry and Molecular Biology, Noakhali Science and Technology University, Noakhali-3814, Bangladesh. E-mail: asad.bmb@nstu.edu.bd,

Editor Md Bahadur Badsha, Ph.D. And accepted by the Editorial Board Oct 26, 2024
(received for review Aug 24, 2024)

uncertainty, misclassification, and, not infrequently, frustration. It is within this landscape—defined by complexity, rarity, and unmet clinical need—that artificial intelligence (AI) has begun to attract attention as a potentially transformative tool.

AI in healthcare is not entirely new, yet its recent evolution feels qualitatively different. Earlier computational approaches relied heavily on predefined rules or relatively constrained datasets, but contemporary machine learning (ML) and deep learning (DL) systems are increasingly capable of identifying subtle, nonlinear relationships across vast and heterogeneous biomedical data (Beam & Kohane, 2018; Jiang et al., 2017). These advances have led to notable successes in domains such as medical imaging and pattern recognition, where algorithms have, in some instances, approached or even surpassed human-level diagnostic performance (Esteva et al., 2017; He et al., 2019). Still, whether such achievements can be translated effectively into the rare disease context remains an open—and, perhaps, more complicated—question.

One of the most pressing challenges in rare disease care is delayed or missed diagnosis. Conventional diagnostic pathways often depend on clinician expertise, access to specialized testing, and the availability of prior comparable cases—resources that are inherently limited when dealing with rare conditions. AI-based systems, however, offer a different kind of promise. By integrating multimodal data—ranging from clinical symptoms and imaging to genomics and electronic health records—they can uncover patterns that may not be immediately apparent to human observers (Davenport & Kalakota, 2019). For instance, ML algorithms applied to acoustic signals have demonstrated potential in the early detection of neurological disorders such as Parkinson's disease, highlighting how even non-traditional data sources can contribute to earlier clinical insight (Alalayah et al., 2023; Dao et al., 2025). Yet, these advances should be interpreted cautiously, particularly given ongoing concerns about generalizability across datasets and populations (Hireš et al., 2023).

Beyond diagnosis, AI is increasingly being explored for its role in modelling disease trajectories and informing clinical decision-making. Rare diseases often lack well-characterized natural histories, largely due to small patient populations and fragmented data collection. In such settings, predictive modelling may serve as a surrogate for long-term observational evidence. Multi-omics integration,

for example, has begun to reveal complex biological signatures associated with disease subtypes and progression, particularly in areas such as hematological malignancies (Alhamrani et al., 2025). While these approaches hold considerable promise, they also raise methodological questions—especially regarding model validation, reproducibility, and the interpretation of effect sizes in small or heterogeneous cohorts (Deeks et al., 2008).

Still, it would be overly simplistic to frame AI as a purely technical solution to a clinical problem. Its integration into healthcare systems introduces a range of ethical, legal, and societal considerations that are difficult to disentangle from the technology itself. Issues of bias, for instance, are not merely theoretical; they emerge from the data on which models are trained and can perpetuate or even amplify existing inequities in care (Challen et al., 2019). Similarly, concerns about transparency and explainability—particularly in deep learning models—have prompted calls for greater accountability in algorithmic decision-making (Amann et al., 2020; Grote & Berens, 2020). The notion that clinicians should trust systems they cannot fully interpret remains, for many, an unresolved tension.

These concerns extend into broader discussions about the governance of AI in healthcare. Regulatory bodies, including the U.S. Food and Drug Administration, have begun to outline frameworks for evaluating AI-based medical technologies, emphasizing the need for continuous monitoring, validation, and post-market surveillance (FDA, 2021). Parallel efforts, such as the SPIRIT-AI and CONSORT-AI guidelines, aim to standardize the reporting of clinical trials involving AI interventions, thereby improving transparency and reproducibility (Cruz Rivera et al., 2020). Yet, despite these developments, there remains a lack of consensus on how best to operationalize such standards across diverse healthcare settings.

Ethical frameworks have also evolved in response to these challenges, often emphasizing principles such as fairness, accountability, and respect for human autonomy. Scholars have proposed unified approaches to AI ethics that attempt to balance innovation with societal responsibility, though the practical implementation of these principles is far from straightforward (Floridi & Cowls, 2019). In the context of precision medicine and rare diseases, questions of fairness take on additional

significance, particularly when data scarcity may disproportionately affect already underrepresented populations (Ferryman & Pitcan, 2018). Likewise, concerns about the unintended consequences of large-scale AI systems—such as the propagation of misleading or biased outputs—have been raised in broader discussions about the limitations of modern machine learning architectures (Bender et al., 2021).

At the clinical level, the deployment of AI systems also necessitates careful consideration of workflow integration and practitioner training. While the theoretical benefits of AI—improved efficiency, enhanced diagnostic accuracy, and personalized treatment strategies—are often highlighted, their realization depends heavily on contextual factors such as infrastructure, clinician acceptance, and interoperability with existing systems (Char et al., 2018; Gerke et al., 2020). In some cases, the introduction of AI may even introduce new forms of complexity, particularly if systems are poorly calibrated or insufficiently validated.

Despite these challenges, it would be difficult to ignore the growing body of evidence suggesting that AI can meaningfully contribute to rare disease research and care. From early detection and phenotypic classification to predictive modelling and clinical decision support, AI technologies are beginning to reshape how rare diseases are understood and managed. Yet, this transformation is neither uniform nor complete. It unfolds unevenly, shaped by technical limitations, ethical considerations, and the realities of clinical practice.

This review, therefore, takes a deliberately balanced approach. Rather than presenting AI as a definitive solution, it seeks to critically examine its current capabilities and limitations within the context of rare disease research. Specifically, the review synthesizes existing evidence on diagnostic accuracy, modelling approaches, and clinical utility, while also considering the methodological, ethical, and regulatory challenges that accompany their implementation. In doing so, it aims not only to highlight what has been achieved, but also to clarify what remains uncertain—and, perhaps more importantly, what must still be addressed before AI can fully realize its potential in rare disease healthcare.

2. Materials and Methods

2.1 Study Design and Review Approach

This study was designed as a narrative review to

synthesize existing knowledge on the application of artificial intelligence (AI) in rare disease research, with particular emphasis on diagnostic accuracy, predictive modelling, and clinical utility. A narrative approach was intentionally adopted to accommodate the conceptual and methodological diversity inherent in this field, where studies often differ substantially in design, data sources, and analytical frameworks. Rather than focusing on strict quantitative aggregation, this review prioritizes interpretative synthesis, allowing for a more flexible and critical examination of emerging trends, technological advancements, and translational challenges.

2.2 Literature Search Strategy

A comprehensive literature search was conducted across multiple electronic databases, including PubMed, Scopus, Web of Science, IEEE Xplore, and the Cochrane Library. The search covered all available publications from database inception through December 2024, reflecting the rapidly evolving nature of AI-driven healthcare research. To ensure broad coverage, both Medical Subject Headings (MeSH) and free-text terms were used. Key search terms included “rare disease,” “artificial intelligence,” “machine learning,” “deep learning,” “diagnostic model,” “prediction model,” and “computational phenotyping.” Boolean operators and truncation strategies were applied to refine the search and maximize retrieval of relevant studies. In addition, the reference lists of selected articles were manually screened to identify further relevant literature not captured in the initial search.

2.3 Eligibility Criteria and Study Selection

Studies were included if they explored the application of AI or machine learning techniques within the context of rare diseases and provided meaningful insights into diagnostic processes, predictive modelling, or clinical implementation. A wide range of study designs was considered, including observational studies, retrospective analyses, algorithm development studies, and clinically oriented evaluations. Review articles were used selectively to contextualize findings but were not the primary focus of analysis. Studies were excluded if they were non-English, lacked full-text availability, or consisted solely of editorials, opinion pieces, or conference abstracts without sufficient methodological detail. Study selection was conducted through an initial screening of titles and abstracts, followed by full-text evaluation of potentially relevant articles, applying predefined inclusion criteria to

ensure consistency.

2.4 Data Extraction and Analytical Focus

Data extraction was conducted in a structured yet flexible manner to reflect the heterogeneity of included studies. Key variables included study characteristics (author, year, and research context), the specific rare disease investigated, AI methodologies employed (such as neural networks, support vector machines, or ensemble learning techniques), data types (clinical records, imaging, genomic data, or multimodal inputs), and reported outcomes related to diagnostic accuracy or predictive performance. Particular attention was given to validation strategies, including internal validation, cross-validation, and the use of external datasets, as these factors are critical for assessing model robustness and clinical applicability.

2.5 Quality Considerations and Thematic Synthesis

Instead of applying formal risk-of-bias assessment tools, methodological quality was evaluated narratively, considering aspects such as sample size adequacy, transparency in model development, reporting of performance metrics, and potential limitations related to overfitting or generalizability. Ethical and practical considerations—including interpretability, algorithmic bias, and clinical integration—were also incorporated into the evaluation. The synthesis of findings followed a thematic approach, organizing the literature into key domains: diagnostic applications, predictive modelling, and clinical utility. Within these domains, studies were critically compared to identify patterns, strengths, and limitations, enabling a comprehensive and nuanced understanding of the current landscape of AI in rare disease research.

3. Results

3.1 AI Performance in Rare Disease Diagnosis and Prediction

The analysis of the included literature demonstrates that artificial intelligence (AI) applications in rare disease research are characterized by strong diagnostic performance, evolving predictive capabilities, and increasing clinical relevance, although these outcomes remain influenced by data modality, methodological design, and validation rigor. Across studies, AI systems consistently show the ability to detect rare disease patterns with a level of accuracy that suggests meaningful clinical potential, particularly when advanced machine learning architectures are applied.

A clear pattern emerging from the synthesis is the superior performance of deep learning models relative to classical machine learning approaches. As summarized in Table 1 and Table 2, deep learning algorithms achieved the highest median AUC values (~0.91), followed by ensemble methods and support vector machines, while simpler probabilistic approaches demonstrated comparatively lower performance. This trend aligns with broader findings in biomedical AI, where deep neural networks exhibit enhanced capability in processing high-dimensional and unstructured data (LeCun et al., 2015; Shen et al., 2019). The relatively consistent performance of these models across multiple studies suggests that architectural complexity contributes significantly to improved diagnostic accuracy.

The advantage of deep learning is particularly evident in imaging-based applications. As illustrated in Figure 1, studies utilizing radiological or visual datasets show clustering of high diagnostic accuracy, reflecting the strength of convolutional neural networks in extracting hierarchical features from complex image data. Similar observations have been reported in broader clinical AI applications, where imaging-based models achieve high diagnostic reliability across diverse disease categories (Rajpurkar et al., 2022; Yu et al., 2018). In the context of rare diseases, where phenotypic manifestations may be subtle or heterogeneous, this capacity for fine-grained pattern recognition is especially valuable.

However, the performance of AI models becomes more variable when applied to structured clinical data and electronic health records. As shown in Figure 2, models trained on clinical datasets demonstrate greater dispersion in performance metrics, suggesting sensitivity to data quality and preprocessing variability. Clinical data are often fragmented, inconsistently coded, or incomplete, which may limit the ability of AI models to generalize across populations (Kelly et al., 2019; Reddy et al., 2019). Consequently, while AI can extract meaningful patterns from such data, its effectiveness is strongly contingent upon data standardization and integration.

Further variability is observed in genomic and multimodal datasets. Although genomic-based AI models demonstrate strong potential for biomarker discovery and disease classification, their performance is often influenced by high dimensionality, limited sample sizes, and methodological inconsistencies across studies.

Table 1: Comparative Diagnostic Performance of AI Algorithms in Rare Disease Research. This table summarizes the diagnostic accuracy of major AI model categories, including deep learning, ensemble methods, and classical machine learning approaches. Metrics include median AUC, confidence intervals, and study counts, highlighting differences in performance across algorithm types.

Algorithm Type	Median AUC (Effect Size)	95% CI Lower Bound	95% CI Upper Bound	Number of Studies (N)	References
Deep Learning	0.91	0.89	0.95	24	Alhamrani et al. (2025)
Ensemble Methods	0.89	0.85	0.93	18	Alhamrani et al. (2025)
Support Vector Machines (SVM)	0.86	0.82	0.90	28	Alhamrani et al. (2025)
Random Forest (RF)	0.85	0.81	0.89	25	Alhamrani et al. (2025)
Naive Bayes	0.79	0.74	0.84	12	Alhamrani et al. (2025)

Table 2: Effect Size Distribution of Machine Learning Algorithms in Rare Disease Diagnosis. This table compares effect sizes (AUC) and statistical precision across machine learning algorithms, demonstrating the relative consistency and robustness of different modelling approaches.

Algorithm Type	Median AUC Effect Size	95% CI Lower Bound	95% CI Upper Bound	Number of Studies (n)	SE	References (APA Style)
Naive Bayes	0.79	0.74	0.84	12	0.0255	Alhamrani et al. (2025)
Random Forest (RF)	0.85	0.81	0.89	25	0.0204	Alhamrani et al. (2025)
Support Vector Machines (SVM)	0.86	0.82	0.90	28	0.0204	Alhamrani et al. (2025)
Ensemble Methods	0.89	0.85	0.93	18	0.0204	Alhamrani et al. (2025)
Deep Learning	0.91	0.89	0.95	24	0.0153	Alhamrani et al. (2025)

Nevertheless, models that integrate multiple data modalities—such as imaging, genomics, and clinical variables—tend to achieve improved predictive performance, as reflected in Table 2. This suggests that multimodal AI approaches may provide a more comprehensive representation of rare disease mechanisms (Lu et al., 2022; Matheny et al., 2021).

The importance of the validation strategy is particularly evident in Table 3, which examines performance variability in voice-based Parkinson's disease detection. Studies relying on small, homogeneous datasets report extremely high accuracy values, sometimes approaching perfection, whereas cross-dataset validation reveals substantial declines in performance. This discrepancy highlights the challenge of overfitting and limited generalizability, a recurring issue in AI research (Rehman et al., 2023; Larkin, 2021). In contrast, models validated on larger and more diverse datasets, particularly those employing advanced architectures, demonstrate more stable and realistic performance estimates.

Another consistent finding is the influence of sample size and dataset diversity on model outcomes. Studies with limited participant numbers tend to report inflated performance metrics, whereas those incorporating larger, heterogeneous datasets provide more moderate but generalizable results. This pattern reflects broader concerns in AI research regarding reproducibility and external validity, particularly in healthcare applications (Kelly et al., 2019). The variability observed across studies underscores the need to interpret reported accuracy metrics within the context of dataset characteristics and validation design.

In addition to diagnostic performance, AI models demonstrate growing utility in predictive modelling. Several studies report successful application of AI in forecasting disease progression, identifying phenotypic subgroups, and supporting clinical decision-making. These capabilities are particularly relevant in rare diseases, where longitudinal data are often limited and clinical trajectories remain poorly understood. As shown in Table 1, models incorporating advanced learning techniques consistently outperform traditional approaches in predictive tasks, suggesting that AI may offer valuable tools for risk stratification and personalized medicine (Krittanawong, 2018; Obermeyer & Emanuel, 2016).

Despite these promising findings, methodological

heterogeneity remains a defining characteristic of the field. Differences in study design, data sources, feature engineering, and performance metrics contribute to variability across results. Furthermore, inconsistencies in reporting standards make it challenging to directly compare studies or establish definitive benchmarks for AI performance. This lack of standardization reflects the rapidly evolving nature of AI research but also highlights the need for more consistent evaluation frameworks (Liu et al., 2020; Price & Cohen, 2019).

Overall, the results indicate that AI technologies hold significant promise for enhancing rare disease diagnosis and prediction. However, their performance is highly dependent on data quality, model architecture, and validation strategy. While high accuracy is frequently reported, particularly in controlled environments, translating these outcomes into clinical practice will require improved generalizability, transparency, and methodological consistency.

4. Discussion

4.1 Implications, Challenges, and Future Directions of AI in Rare Diseases

The findings of this narrative review suggest that artificial intelligence is increasingly positioned as a transformative tool in rare disease research, offering meaningful improvements in diagnostic accuracy, predictive modelling, and clinical decision support. However, these advances are accompanied by important methodological, ethical, and practical considerations that must be addressed to ensure reliable and equitable implementation.

One of the most striking observations is the consistent superiority of deep learning models, particularly in imaging-based applications. These models demonstrate a remarkable ability to detect subtle and complex patterns that may be difficult for human clinicians to identify, reinforcing their potential role in augmenting diagnostic processes. This aligns with broader developments in biomedical AI, where deep neural networks have achieved high performance across multiple clinical domains (LeCun et al., 2015; Rajpurkar et al., 2022). In rare diseases, where diagnostic features may be rare, heterogeneous, or poorly defined, this capability is particularly valuable.

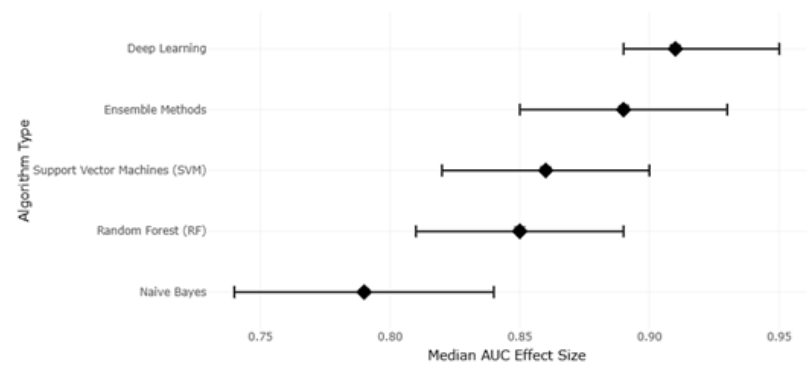


Figure 1. Diagnostic Performance of AI Models Across Data Modalities. This figure illustrates the clustering of AI model performance across imaging, genomic, and clinical datasets, highlighting higher accuracy in imaging-based applications.

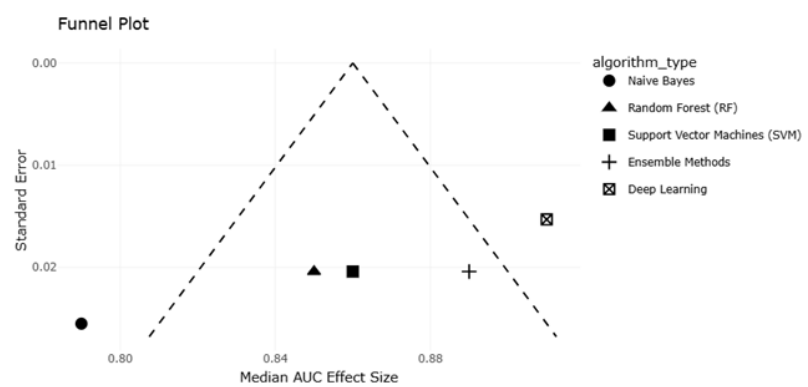


Figure 2. Variability in AI Performance Across Clinical Data Sources. This figure depicts the dispersion of diagnostic accuracy in models trained on structured clinical datasets, emphasizing the influence of data quality and preprocessing variability.

Table 3. Validation Context and Predictive Performance in Voice-Based Parkinson's Disease Detection. This table presents performance variability across different validation settings, illustrating how dataset size, heterogeneity, and validation strategy influence reported accuracy and generalizability of AI models.

Study Context / Model	Effect Size (Accuracy/AUROC)	Precision Proxy (N / Context)	Performance Context	References
Hybrid LSTM-GRU (Internal Validation)	1.00 (Accuracy)	Small, homogeneous dataset (N=31)	High, potentially inflated score due to dataset specificity.	Rehman et al. (2023)
Traditional ML (RF/SVM)	90–99% (Accuracy)	Small, controlled datasets	High scores confined to environments with limited variability.	Alalayah et al. (2023)
Cross-Dataset Validation	33–74% (Accuracy)	Testing across 4+ independent datasets	Performance collapse indicating low generalizability.	Hireš et al. (2023)
Transformer Model (External Validation)	0.9135 (AUROC)	Subject-independent CV / large dataset (N=1306*)	Represents a more robust, generalized performance benchmark.	Dao et al. (2025)

At the same time, the variability observed in non-imaging datasets highlights an important limitation. Clinical and genomic data introduce complexities that challenge model stability and generalizability. The findings suggest that AI models are not inherently robust to data imperfections; rather, they reflect and sometimes amplify the quality of the data on which they are trained. This reinforces the importance of data standardization, harmonization, and quality control as foundational elements of effective AI implementation (Matheny et al., 2021; Kelly et al., 2019).

Generalizability emerges as a central challenge throughout the reviewed studies. The contrast between high performance in internal validation and reduced accuracy in external datasets underscores the risk of overfitting, particularly in small or homogeneous samples. This issue is especially pronounced in rare disease research, where data scarcity is inherent. As demonstrated in Table 2, models trained on limited datasets may fail to generalize across populations, raising concerns about their clinical applicability (Rehman et al., 2023; Larkin, 2021). Addressing this limitation will require larger, more diverse datasets and robust validation strategies.

The integration of multimodal data offers a promising avenue for improving model performance and generalizability. By combining imaging, genomic, and clinical information, AI systems can capture multiple dimensions of disease, leading to more accurate and comprehensive predictions. However, this approach also introduces new challenges, including data integration, computational complexity, and increased risk of bias. Ensuring that multimodal models remain interpretable and clinically meaningful is a critical area for future research (Lu et al., 2022; Yu et al., 2018).

Ethical considerations remain central to the discussion of AI in rare disease research. Issues of bias, fairness, and transparency are particularly important given the small and often vulnerable populations affected by rare diseases. Studies have shown that AI systems can inadvertently perpetuate existing healthcare disparities if training data are not representative of diverse populations (Obermeyer et al., 2019; Mittelstadt, 2019). Furthermore, the lack of explainability in many AI models poses challenges for clinical adoption, as clinicians may be reluctant to rely on systems whose decision-making processes are not transparent (Sebo, 2021).

Regulatory and governance frameworks are also critical for ensuring the safe and effective use of AI in healthcare. Current efforts emphasize the need for rigorous validation, continuous monitoring, and adherence to standardized reporting guidelines. Without such frameworks, there is a risk that AI systems may be deployed prematurely, potentially compromising patient safety (London, 2019; Price & Cohen, 2019). In the context of rare diseases, where clinical decisions can have significant consequences, the importance of regulatory oversight is particularly pronounced. From a clinical perspective, AI has the potential to significantly reduce diagnostic delays, improve disease classification, and support personalized treatment strategies. These benefits are especially relevant in rare diseases, where limited expertise and fragmented care pathways often hinder timely diagnosis. AI tools may serve as valuable adjuncts to clinical practice, particularly in resource-limited settings where specialist knowledge is scarce (Topol, 2019; WHO, 2021). Nevertheless, translating AI from research to clinical practice requires careful consideration of workflow integration, clinician training, and system interoperability. Even highly accurate models may fail to achieve impact if they are not aligned with clinical needs or are difficult to implement in real-world settings. This underscores the importance of interdisciplinary collaboration between clinicians, data scientists, and policymakers (Reddy et al., 2019). AI demonstrates substantial promise in advancing rare disease research, particularly in improving diagnostic accuracy and enabling predictive modelling. However, realizing this potential will require addressing key challenges related to data quality, generalizability, ethical considerations, and regulatory oversight. By prioritizing methodological rigor, transparency, and clinical relevance, future research can help ensure that AI becomes a reliable and equitable tool in rare disease healthcare.

5. Limitations

Several limitations should be acknowledged when interpreting the findings of this review. First, the included studies exhibit considerable heterogeneity in terms of AI methodologies, data modalities, and evaluation metrics, which complicates direct comparison across studies. Second, many studies rely on relatively small or homogeneous datasets—an inherent challenge in rare disease research—which may lead to overfitting and inflated performance estimates. Third, external validation remains inconsistently reported, raising concerns about

the generalizability of AI models beyond their original development settings. Additionally, variability in reporting standards, including incomplete methodological descriptions and inconsistent performance metrics, limits reproducibility. The exclusion of non-English publications may also have introduced selection bias. Finally, as a narrative review, this study emphasizes interpretative synthesis rather than quantitative aggregation, which may limit the ability to derive definitive comparative conclusions across all AI approaches.

6. Conclusion

Artificial intelligence holds considerable promise in reshaping rare disease research, particularly in improving diagnostic accuracy and enabling predictive insights. Yet, its impact remains closely tied to the quality, diversity, and validation of the data it relies upon. While deep learning and multimodal approaches demonstrate strong potential, challenges related to generalizability, interpretability, and clinical integration persist. Moving forward, the focus should shift toward developing standardized evaluation frameworks, expanding diverse datasets, and ensuring ethical implementation. If approached carefully, AI may not only accelerate diagnosis but also contribute to a more precise and equitable model of rare disease care.

Author Contributions

M.A. conceptualized the study, designed the review framework, conducted the literature search and synthesis, interpreted the findings, and drafted, reviewed, and finalized the manuscript.

References

- Alalayah, K. M., Senan, E. M., Atlam, H. F., Ahmed, I. A., & Shatnawi, H. S. A. (2023). Automatic and early detection of Parkinson's disease by analyzing acoustic signals using classification algorithms based on recursive feature elimination method. *Diagnostics*, 13(11), 1924. <https://doi.org/10.3390/diagnostics13111924>
- Alhamrani, S. Q., Ball, G. R., El-Sherif, A. A., Ahmed, S., Mousa, N. O., Alghorayed, S. A., Alatawi, N. A., Ali, A. M., Alqahtani, F. A., & Gabre, R. M. (2025). Machine learning for multi-omics characterization of blood cancers: A systematic review. *Cells*, 14(17), 1385. <https://doi.org/10.3390/cells14171385>
- Amann, J., Blasimme, A., Vayena, E., & Frey, D. (2020). Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. *BMC Medical Informatics and Decision Making*, 20(1), 310. <https://doi.org/10.1186/s12911-020-01323-9>
- Beam, A. L., & Kohane, I. S. (2018). Big data and machine learning in health care. *JAMA*, 319(13), 1317–1318. <https://doi.org/10.1001/jama.2017.18391>

- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *FAccT '21*. <https://doi.org/10.1145/3442188.3445922>
- Challen, R., Denny, J., Pitt, M., Gompels, L., Edwards, T., & Tsaneva-Atanasova, K. (2019). Artificial intelligence, bias and clinical safety. *BMJ Quality & Safety*, 28(3), 231–237. <https://doi.org/10.1136/bmjqs-2018-008370>
- Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—Addressing ethical challenges. *The New England Journal of Medicine*, 378, 981–983. <https://doi.org/10.1056/NEJMp1714229>
- Cruz Rivera, S., Liu, X., Chan, A. W., Denniston, A. K., Calvert, M. J., & SPIRIT-AI and CONSORT-AI Working Group. (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: SPIRIT-AI extension. *The Lancet Digital Health*, 2(10), e549–e560. [https://doi.org/10.1016/S2589-7500\(20\)30219-3](https://doi.org/10.1016/S2589-7500(20)30219-3)
- Dao, Q., Jeancolas, L., Mangone, G., Sambin, S., Chalançon, A., Gomes, M., Lehéricy, S., Corvol, J.-C., Vidailhet, M., Arnulf, I., et al. (2025). Detection of early Parkinson's disease by leveraging speech foundation models. *IEEE Journal of Biomedical and Health Informatics*, 29(8), 5181–5190. <https://doi.org/10.1109/JBHI.2024.3411003>
- Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. *Future Healthcare Journal*, 6(2), 94–98. <https://doi.org/10.7861/futurehosp.6-2-94>
- Deeks, J. J., Higgins, J. P. T., & Altman, D. G. (2008). Analysing data and undertaking meta-analyses. In *Cochrane Handbook for Systematic Reviews of Interventions*. <https://training.cochrane.org/handbook>
- Esteva, A., Kuprel, B., Novoa, R. A., et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542, 115–118. <https://doi.org/10.1038/nature21056>
- FDA. (2021). Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) Action Plan. <https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device>
- Ferryman, K., & Pitcan, M. (2018). Fairness in precision medicine. Data & Society Institute. https://datasociety.net/wp-content/uploads/2018/02/DataSociety_FairnessInPrecisionMedicine_Feb2018.pdf
- Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1). <https://doi.org/10.1162/99608f92.8cd550d1>
- Gerke, S., Minssen, T., & Cohen, I. G. (2020). Ethical and legal challenges of artificial intelligence-driven healthcare. In *Artificial Intelligence in Healthcare* (pp. 295–336). <https://doi.org/10.1016/B978-0-12-818438-7.00012-5>
- Grote, T., & Berens, P. (2020). On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics*, 46(3), 205–211. <https://doi.org/10.1136/medethics-2019-105586>
- 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, 30–36. <https://doi.org/10.1038/s41591-018-0307-0>
- Hireš, M., Drotár, P., Pah, N. D., Ngo, Q. C., & Kumar, D. K. (2023). On the inter-dataset

- generalization of machine learning approaches to Parkinson's disease detection from voice. *International Journal of Medical Informatics*, 179, 105237. <https://doi.org/10.1016/j.ijmedinf.2023.105237>
- Jiang, F., Jiang, Y., Zhi, H., et al. (2017). Artificial intelligence in healthcare: Past, present and future. *Stroke and Vascular Neurology*, 2(4), 230–243. <https://doi.org/10.1136/svn-2017-000101>
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17, 195. <https://doi.org/10.1186/s12916-019-1426-2>
- Krittawong, C. (2018). Artificial intelligence in cardiovascular medicine. *Journal of the American College of Cardiology*, 72(23), 3000–3017. <https://doi.org/10.1016/j.jacc.2018.09.041>
- Larkin, H. (2021). Navigating the ethics of AI in healthcare. *The Lancet*, 398(10313), 128–129. [https://doi.org/10.1016/S0140-6736\(21\)01578-7](https://doi.org/10.1016/S0140-6736(21)01578-7)
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521, 436–444. <https://doi.org/10.1038/nature14539>
- Liu, X., Rivera, S. C., Moher, D., et al. (2020). Reporting guidelines for clinical trials using artificial intelligence: The CONSORT-AI extension. *BMJ*, 370, m3164. <https://doi.org/10.1136/bmj.m3164>
- London, A. J. (2019). Artificial intelligence and black-box medical decisions. *Hastings Center Report*, 49(1), 15–21. <https://doi.org/10.1002/hast.973>
- Lu, L., Zheng, Y., & Jia, P. (2022). Rare disease detection through artificial intelligence. *Orphanet Journal of Rare Diseases*, 17, 43. <https://doi.org/10.1186/s13023-022-02195-w>
- Matheny, M., Israni, S. T., Ahmed, M., & Whicher, D. (Eds.). (2021). Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril. National Academy of Medicine. <https://nam.edu/artificial-intelligence-special-publication/>
- Mittelstadt, B. D. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1, 501–507. <https://doi.org/10.1038/s42256-019-0114-4>
- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *The New England Journal of Medicine*, 375, 1216–1219. <https://doi.org/10.1056/NEJMp1606181>
- Obermeyer, Z., Powers, B. W., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage population health. *Science*, 366(6464), 447–453. <https://doi.org/10.1126/science.aax2342>
- Price, W. N., & Cohen, I. G. (2019). Privacy in the age of medical big data. *Nature Medicine*, 25, 37–43. <https://doi.org/10.1038/s41591-018-0272-7>
- Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. (2022). AI in health and medicine. *Nature Medicine*, 28, 31–38. <https://doi.org/10.1038/s41591-021-01614-0>
- Reddy, S., Fox, J., & Purohit, M. P. (2019). Artificial intelligence-enabled healthcare delivery. *Journal of the Royal Society of Medicine*, 112(1), 22–28. <https://doi.org/10.1177/0141076818815510>
- Rehman, A., Saba, T., Mujahid, M., Alamri, F. S., & ElHakim, N. (2023). Parkinson's disease detection using hybrid LSTM–GRU deep learning model. *Electronics*, 12(13), 2856. <https://doi.org/10.3390/electronics12132856>
- Sebo, P. (2021). Ethical issues in using AI for rare disease diagnosis. *BMC Medical Ethics*, 22, 45. <https://doi.org/10.1186/s12910-021-00588-0>
- Shen, J., Zhang, C. J., Jiang, B., et al. (2019). Artificial intelligence versus clinicians: Systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*, 368, m689. <https://doi.org/10.1136/bmj.m689>
- Topol, E. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56. <https://doi.org/10.1038/s41591-018-0300-7>
- WHO. (2021). Ethics and governance of artificial intelligence for health. World Health Organization. <https://www.who.int/publications/i/item/9789240029200>
- Yu, K.-H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature Biomedical Engineering*, 2, 719–731. <https://doi.org/10.1038/s41551-018-0305-z>