Artificial Intelligence Review

https://doi.org/10.1007/s10462-026-11593-8

Article in Press

Empowering healthcare 5.0 with deep learning: techniques, trends, and future directions

Received: 12 September 2025

Accepted: 13 May 2026

Published online: 25 May 2026

Cite this article as: Maji P.K., Chakraborty S., Sadiq A. *et al.* Empowering healthcare 5.0 with deep learning: techniques, trends, and future directions. *Artif Intell Rev* (2026). https://doi.org/10.1007/s10462-026-11593-8

Paramita Kundu Maji, Sanjay Chakraborty, Afifa Sadiq, Saikat Basu & Krishnendu Ghosh

Empowering healthcare 5.0 with deep learning: techniques, trends, and future directions

Empowering Healthcare 5.0 with Deep Learning: Techniques, Trends, and Future Directions

Paramita Kundu Maji*[1], Sanjay Chakraborty*[2], Afifa Sadiq[3], Saikat Basu[4], and Krishnendu Ghosh[5]

[1]Pramita Kundu Maji is associated with the Department of Computer Science and Engineering, Indian Institute of Information Technology Dharwad, India, and Department of Computer Science and Engineering, Techno International New Town, Kolkata, India (Email: paramitakundu.job@gmail.com).

[2]Department of Computer and Information Science (IDA), REAL, AIICS, Linköping University, Sweden and Department of Computer Science and Engineering, Techno International New Town, Kolkata, India

[3]Afifa Sadiq is associated with the Department of Computer Science and Engineering (Data Science), Techno International New Town, Kolkata, India (Email: afifasadiq03@gmail.com).

[4]Saikat Basu is associated with the Department of Computer Science and Engineering, Maulana Abul Kalam Azad University of Technology, Kolkata, India (Email: saikatbasu@gmail.com).

[5]Krishnendu Ghosh is associated with the Department of Computer Science and Engineering, Indian Institute of Information Technology Dharwad, India (Email: krishnendu@iiitdwd.ac.in).

May 15, 2026

1

Abstract

The rapid advancement of intelligent technologies has driven a transformative shift in healthcare, giving rise to Healthcare 5.0, a new paradigm centered on patient-focused, digitally empowered medical services. This study presents a comprehensive review of cutting-edge Deep Learning (DL) techniques that are redefining Healthcare 5.0 through applications in disease prediction and early diagnosis, medical image analysis and radiology, and multimodal deep learning (MMDL). We analyze a broad spectrum of DL architectures, including Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTMs), Gated Recurrent Units (GRUs), Generative Adversarial Network (GANs), transformers, autoencoders, transfer learning, and MMDL, highlighting the unique features that make them particularly suited for healthcare tasks. Using bibliometric analysis of 1,342 Scopus-indexed publications (2018-2025) via VOSviewer, we uncover key research trends, thematic clusters, and emerging focus areas. Furthermore, we assess DL models across diverse real-world healthcare datasets and discuss critical challenges, including data privacy, model interpretability, and system integration. Following the PRISMA methodology, this review also synthesizes recent research on advanced DL architectures for disease diagnosis and Healthcare 5.0, alongside widely used medical imaging and multimodal datasets.The manuscript has been substantially strengthened by incorporating a comparative analysis, quantitative aggregation of reported results, an explicit novelty and contribution statement, and a systematic comparison of DL survey studies across Healthcare 4.0 and Healthcare 5.0, thereby providing a more critical, data driven, and clearly differentiated perspective beyond prior reviews.Finally, we address significant issues, including ethical considerations, privacy concerns, and

[1]Corresponding authors: Sanjay Chakraborty, and Paramita Kundu Maji (Email: sanjay.chakraborty@liu.se, paramitakundu.job@gmail.com)

model limitations, while outlining promising directions for future research. This work serves as a valuable resource for researchers and practitioners, offering both a broad overview and deep insights into the latest developments in DL for Healthcare 5.0 systems.

Keywords: Healthcare, Deep learning, Artificial intelligence, Medical imaging, Disease diagnosis, Multimodality.

1 Introduction

Healthcare 4.0 significantly enhances the quality, productivity, flexibility, cost-effectiveness, and dependability of healthcare services while primarily focusing on the business model [140, 87, 84, 195]. However, Healthcare 5.0 builds on the innovations of Healthcare 4.0 and focuses on the customer model, marking the next phase of the healthcare revolution. Healthcare 5.0 represents a comprehensive transformation that extends beyond technological innovation [265], shifting from disease-centered to patient-centered care [213]. Information systems have evolved from localized clinical applications to regional medical informatization [165], while management practices have advanced from generalized approaches to personalized, preventive healthcare [248, 85]. The practice of healthcare has evolved through the initial evidence-based medicine to industrialized and automated as well as digitally driven medical systems. The next level is Healthcare 5.0, which focuses on personalized care, continuous monitoring, and as well as improvement of well-being. It combines innovative technologies with big data analytics, Artificial Intelligence (AI), Machine Learning (ML), Deep Learning (DL), Internet of Things (IoT), cloud solutions and telemedicine[258] to facilitate real-time decision making and secure and patient-centered services.Healthcare 4.0 can be defined as a technology-driven paradigm that integrates IoT, big data analytics, cloud computing, and artificial intelligence to enhance automation, connectivity, and efficiency in healthcare delivery systems [87, 84, 195]. In contrast, Healthcare 5.0 extends this paradigm toward a human-centric approach, where advanced technologies such as deep learning are combined with human intelligence to enable personalized, preventive, and ethically grounded healthcare solutions [248, 85, 140, 258, 165]. The key distinction lies in the shift from system efficiency to patient-centric intelligence and collaborative decision-making. To clearly distinguish Healthcare 5.0 from its predecessor, Healthcare 4.0, a concise comparison is presented in Table 1, highlighting differences in technological focus, system design philosophy, and healthcare delivery models.

Table 1: Comparison between Healthcare 4.0 and Healthcare 5.0

Aspect	**Healthcare 4.0**	**Healthcare 5.0**
Core paradigm	Technology-driven, automation-focused systems [84, 87, 140]	Human-centric, personalized healthcare ecosystem [258, 165]
Focus of care	Efficiency, productivity, and cost reduction [195, 87]	Patient well-being, personalization, and sustainability [213, 258]
Role of AI & Deep Learning	Automation, prediction, and analytics [140, 84]	Collaborative intelligence with human-in-the-loop decision-making [213, 165]
Patient role	Passive data source and service recipient [140]	Active participant and decision-maker [258, 165]
Data utilization	Large-scale but fragmented data processing [195]	Integrated, context-aware, and personalized data usage [265, 85]
Ethical considerations	Limited emphasis on explainability and bias [87]	Strong focus on explainable, transparent, and trustworthy AI [213]

The implementation of Healthcare 5.0 faces significant challenges arising from the vast scale and intricate nature of healthcare data. Sources such as medical imaging, electronic health records (EHRs), wearable devices, and genomic data produce enormous, heterogeneous, high-dimensional, and mostly unstructured data [203, 29]. Conventional analysis approaches cannot handle such big and dense data. ML techniques are a branch of AI[86], presenting a potential solution to all these limitations, aiming to optimize processes, improve the efficiency of healthcare systems, enhance the quality of patient care, and increase the precision of physicians' work [140, 264, 114, 219]. They serve as a cornerstone in the advancement of clinical decision support systems, disease diagnosis methodologies, and personalized treatment strategies aimed at optimizing patient outcomes [75, 4, 124, 231, 51].

Regardless of their potential, traditional ML models have serious limitations when involved in contemporary healthcare settings. They usually use handcrafted characteristics, restricting their outcome, scalability, and flexibility to be applied to new data forms [167, 162]. In addition, ML models cannot process high-dimensional and unstructured data that include medical images and free-text clinical notes. Such limitations hamper their capacity to explore the potential of multimodal healthcare data to its fullest extent to pursue the vision of

Healthcare 5.0 [65]. To address these challenges, DL is replacing ML in the healthcare industry to handle bigger, more complex datasets, increase accuracy, and provide better data analysis tools [42, 212, 56, 259, 216]. In recent years, the healthcare domain has experienced a quick and striking improvement through DL technology, maintaining the hope for more effective and reasonably priced medical treatment. In low-resource countries with limited healthcare infrastructure, the shortage of trained radiologists and specialized training programs is further compounded by limited awareness among physicians regarding the benefits of diagnostic radiography and interventional radiology (IR) techniques [111, 208]. These problems help automated reporting and image-based diagnosis move forward. DL techniques present a promising approach to addressing the shortage of qualified physicians and experienced radiologists. These techniques facilitate automated report generation and assist radiologists in improving diagnostic efficiency. They are particularly effective in analyzing medical images such as chest X-rays, computed tomography (CT), and magnetic resonance imaging (MRI). Moreover, research over the past several years highlights that EHRs [266] have increasingly incorporated computer-aided detection (CADe) [193, 152] and computer vision (CV) techniques to advance disease prediction from medical images.

Nevertheless, a major concern with DL in healthcare is its tendency to operate as a *black box*. This arises from the non-linear and complex nature of its architectures. As a result, DL models often generate decisions without providing transparent explanations for their outcomes [41]. The absence of transparency gives rise to questions regarding the comprehension, visualization, dependability, impartiality, and ethical consequences of AI-powered healthcare operations [213, 102]. Explainable AI (XAI) presents a hopeful resolution to this obstacle by offering intelligible insights into the decision-making process [55, 118]. Explainable AI is focused on elucidating the underlying knowledge of DL's black-box model, which uncovers the decision-making process [12, 22, 77, 285, 274, 35].

Recent trend indicates that, MMDL, foundation model, federated learning, and XAI are developing into booming fields because of the applicability in personalized and privacy-conscious Healthcare 5.0 settings. Real-time monitoring, wearable data and cross-modal fusion applications are also becoming popular. Conversely, the trend of single-modality CNN-based image classification and older RNN-based architecture exhibit a downward trend, with newer transformer-based and hybrid models outperforming them and being more scalable. These changes are indicative of a bigger trend towards integrated, context-sensitive systems that can be used to predict illness more comprehensively [9, 149, 171].

This survey provides a comprehensive examination of the role of DL within the framework of Healthcare 5.0, structured around the following research questions.

- *RQ1:* In what ways do DL models support the vision of Healthcare 5.0, and how do their advantages compare with those of traditional ML approaches?
- *RQ2:* Which DL algorithms are most commonly applied in healthcare, and what key features make them particularly suitable for clinical applications?
- *RQ3:* What have recent studies done to use DL in disease prediction, imaging, and multimodal healthcare, and what is the role of datasets?
- *RQ4:* What are the primary challenges and limitations of DL in Healthcare 5.0?
- *RQ5:* What open research gaps persist, and which directions appear most promising for advancing efficient, trustworthy, and scalable DL models tailored to Healthcare 5.0?
- *RQ6:* What are the methods of the evaluation and validation of DL models in Healthcare 5.0, and how the related concerns like data leakage, bias, and reproducibility are addressed today?
- *RQ7:* What impact of interpretability, fairness and regulatory compliance on clinical trust and adoption of DL systems in Healthcare 5.0?

This survey aims to offer an in-depth review of the most recent DL algorithms employed in the healthcare field. This paper can be a valuable resource for researchers and professionals who wish to promptly familiarize themselves with the latest advancements in DL for healthcare applications. The main contributions of this study are given below.

- We present a comprehensive state-of-the-art survey of DL techniques, including multimodal architectures such as CNNs, RNNs, LSTMs, GRUs, GANs, diffusion models, and Transformer-based models, applied to Healthcare 5.0, with a particular focus on disease prediction, diagnostic modeling, and medical image analysis.
- In this study, through an extensive literature review, the problems of limited input medical datasets are addressed by the synthesis of generative models in modern healthcare.
- We conduct a bibliometric analysis using VOSviewer on 1,342 Scopus-indexed publications from 2018 to 2025. This analysis reveals dominant keywords, research hotspots, and thematic clusters in the field of DL applications in Healthcare 5.0. The resulting co-occurrence network visualization provides insight into the intellectual structure and evolving focus areas of this interdisciplinary domain.
- We present comparative tables of recent studies employing diverse DL techniques, along with systematic descriptions of datasets and literature reviews focused on disease prediction and early diagnosis, medical image analysis and radiology, and MMDL in healthcare.
- This study offers an in-depth analysis of fundamental challenges and research gaps in DL-based healthcare systems, highlighting a range of interesting new directions for future research in this field.
- We present novelty and contribution where the proposed review compared with recent related surveys across scope, methodology, data modalities, application focus, and identified gaps of this literature review and also comparative analysis between Healthcare 4.0 and Healthcare 5.0.
- We presents a critical analysis across the three domains of DL in Healthcare 5.0.
- We perform a systematic assessment of methodological quality, validation practices, bias risks, and explainable AI (XAI) adoption across the three domains of DL in Healthcare 5.0.
- We conduct quantitative aggregation of classification accuracies across disease categories using benchmark datasets.

For clarity, the mapping between the stated contributions and their corresponding sections and tables is provided in Table 2.

Table 2: Mapping of contributions to manuscript sections and tables

Contribution description	**Addressed in section(s)**	**Supporting tables/figures**
Survey of deep learning architectures for Healthcare 5.0	Section 3	Figure 7
Use of generative models to address limited medical datasets	Sections 4.1–4.3	Table 10, 11, Figure 8, 9, 10, 11, 12, 13, 14
Bibliometric analysis of DL-based Healthcare 5.0 research (2018–2025)	Section 2	Table 5, 6, 7, 8 Figure 2, 3, 4, 5, 6
Comparative analysis of studies across three healthcare domains	Sections 4.1–4.3	Table 10, 11, 12, 13, 14, 15, 16 Figure 8, 9, 10, 11, 12, 13, 14
Identification of challenges, research gaps, and future directions	Sections 5–6	Table 17
Comparison with existing surveys and Healthcare 4.0 vs 5.0	Section 1.1	Table 1, 3, 4
Assessment of methodological quality, validation, bias risks, and XAI adoption	Section 5.4	Table 19
Quantitative aggregation of classification performance across disease categories	Section 5.5	Table 18

The paper is organized as follows: Section 2 covers literature selection and bibliometric analysis; Section 3 reviews key DL algorithms in healthcare; Section 4 details the evaluation approach, focusing on disease prediction, medical imaging, and MMDL; Section 5 discusses challenges such as data availability, privacy, ethics, interpretability, and integration; Section 6 outlines future directions, including hybrid models, XAI, and federated learning; and Section 7 concludes the study. The conceptual framework is shown in Figure 1.

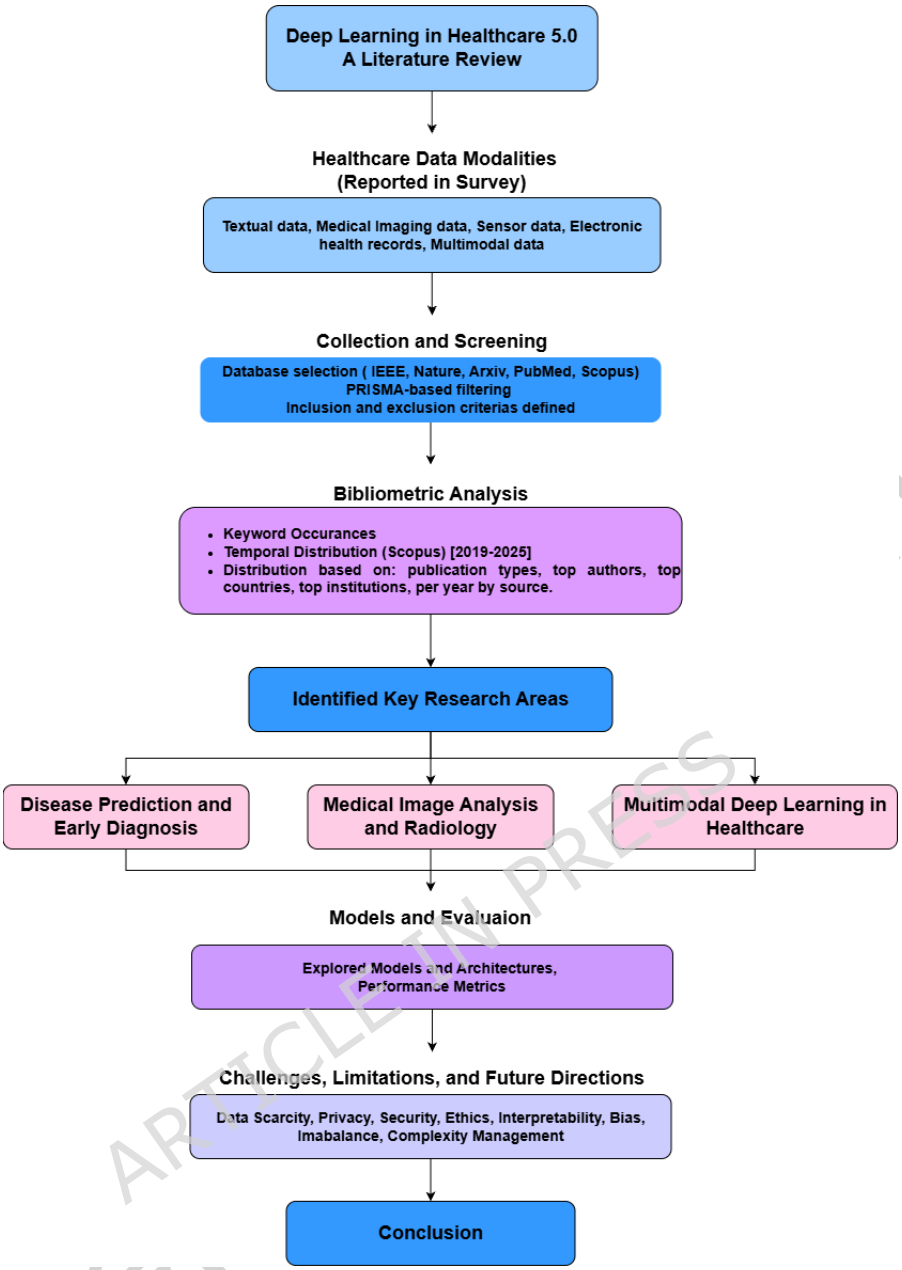


Figure 1: Conceptual framework of the study

1.1 Novelty and contribution

Table 3 systematically contrasts the proposed review with recent related surveys across scope, methodology, data modalities, application focus, and identified gaps. This table clearly highlights how the present work extends beyond existing reviews by explicitly aligning DL techniques with the Healthcare 5.0 paradigm, integrating human centric, intelligent, and sustainable healthcare perspectives. The comparison clarifies the unique contribution and novelty of the review, thereby positioning it distinctly within the current body of literature.

Table 4 presents a brief comparison between representative prior survey works and our study on deep learning-based disease diagnosis. The reviews available are primarily about Healthcare 4.0, discuss AI/DL usage extensively, and are largely qualitative with limited or partial quantitative analysis. In contrast, our research, falls under the Healthcare 5.0 paradigm, is more comprehensive and structured, by reviewing a larger number of studies, systematizes trends, challenges, and future directions, and is supported by quantitative and

Table 3: Comparison with existing reviews

Key literature work	Study type	Model types evaluated	Database used	Evaluation metrics	Pretraining analysis	Interpretability analysis	Computational complexity analysis	Comprehensiveness claim justification
[34]	Survey	Deep Neural Networks(DNNs), hybrid DNN+ML classifiers	public medical imaging datasets	Not empirical	Not covered	mentioned briefly	Qualitative remarks on Qualitative remarks on parameter counts and computational cost	The survey covers 117 studies in general in the domain of DNNs, hyperparameter optimization, and MH-based feature selection, but directly not compare it with recent similar reviews.
[74]	Systematic review	Statistical imputation, classical ML, DL, and multimodal fusion	Summarizes multiple disease areas	summarizes metrics used (e.g., imputation error, AUC)	Discusses only representation learning but not dedicated analysis	Not covered systematically	No structured complexity comparison	cover multimodal missing-data methods and applications, but no coverage in comparison to previous related surveys.
[262]	Empirical bibliometric	Bidirectional Encoder Representations from Transformers (BERT)-based text-mining pipeline	Web of Science papers (1,587) and Derwent patent records (1,314) on AI in healthcare (2018–2022)	no clinical performance metrics	first to apply Bidirectional Encoder Representations from Transformers (BERT)-based self-supervised models	No clinical interpretability analysis	qualitative cost comparison between BERT and classical text-mining methods, no detailed complexity benchmarks	Novelty by first BERT-based AI-healthcare landscape analysis and combined paper-patent view and not position against disease-specific DL reviews.
[272]	Survey	CNN, DNN, federated	Publicly available dataset	No unified empirical benchmark	Not covered	Not covered	Covered conceptually	Broad focus on edge computing, deep learning, and medical CV, but not provide sufficient comparative support of breadth.
[95]	Survey	CNN, RNN, hybrid DL, transfer learning	Multiple published works	accuracy, sensitivity, specificity, and AUC	Discusses transfer learning qualitatively	Brief discussion	No explicit analysis	Claims complete coverage of COVID-19 DL by modality and task, but does not explicitly compare this to previous coverage of COVID-DL.
[179]	Narrative review	CNNs, RNNs, transformers, Variational Autoencoders (VAEs), GANs, Graph Neural Networks (GNNs), attention and holography-based models	Publicly available datasets	Not mentioned	Not covered	Briefly covers	No detailed analysis	Claims comprehensive DL-in-medicine coverage with holography emphasis, but lacks explicit comparison to prior reviews.
Our study	Empirical comparative study	DL models (CNNs, RNNs, LSTMs, GRUs, GANs, transformers etc.)	Extensive healthcare datasets	Accuracy, positive predictive value, precision, sensitivity, specificity, AUC-ROC, recall rate, F1-score, mean, standard deviation	Comprehensive analysis	Detailed explainability analysis using multiple techniques	In-depth comparative analysis across different model architectures	First study to systematically compare mostly used DL models in healthcare across various datasets, detailed dataset descriptions, evaluation metrics, and interpretability metrics, providing a holistic and empirical assessment of strengths, limitations, and practical applicability.

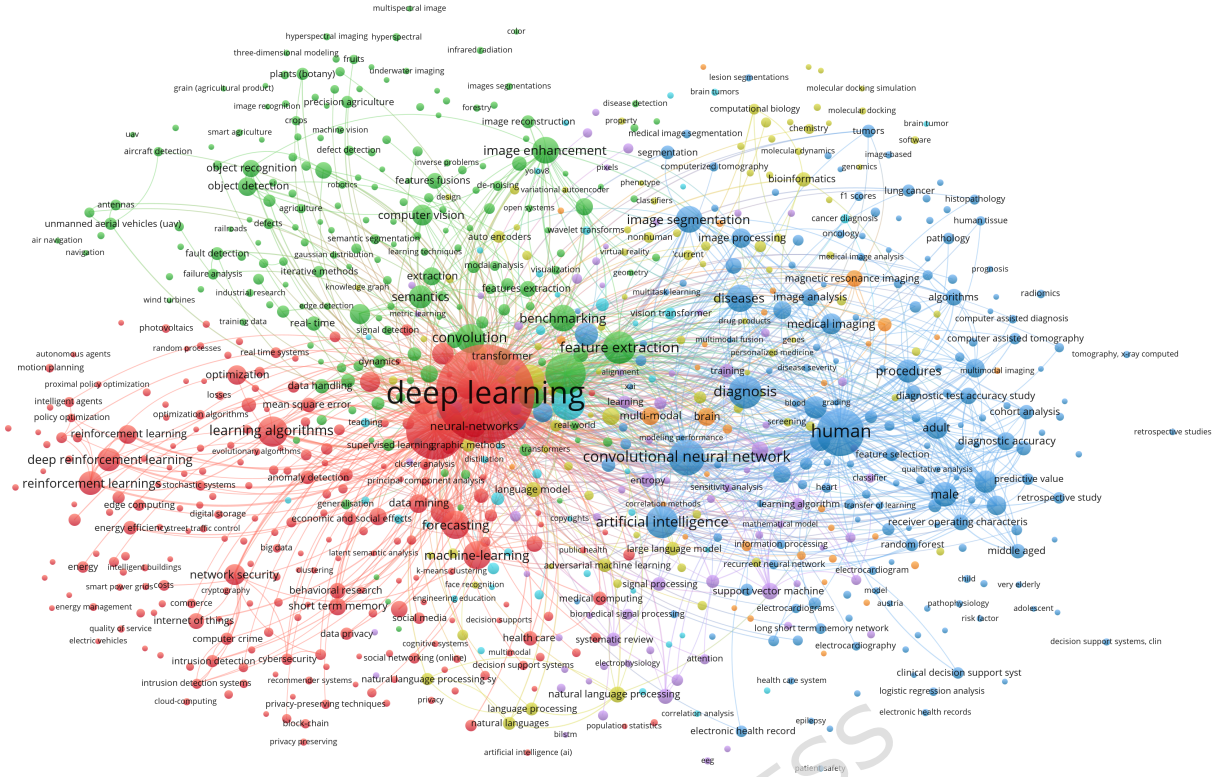


Figure 2: Network visualization of co-occurring keywords on DL techniques in Healthcare 5.0 using VOSviewer (Scopus database, year 2025-2019).

critical comparative analysis.

Table 4: Comparison of Deep Learning Survey Studies Across Healthcare 4.0 and 5.0

Key literature work	Focus	Healthcare era	Novelty/Contribution	Papers Reviewed	Quantitative Analysis
[120]	AI (ML and DL) in healthcare	Healthcare 4.0	Broad overview of AI methods, disease applications, and future challenges	200	Mostly qualitative
[174]	DL in healthcare	Healthcare 4.0	Mixed survival review with a proposed MAR-ERNN model	24	Yes
[133]	ML, DL in healthcare	Healthcare 4.0	ML/DL healthcare overview	20	Limited
Our study	DL in healthcare	Healthcare 5.0	Detailed review with categorizing trends, challenges, future directions, bibliometric analysis, critical review, comparative analysis	61	Yes

2 Literature selection and bibliometric analysis

The selection of papers for this review on DL techniques in healthcare is conducted using a systematic search strategy on Scopus to ensure the inclusion of high-quality and relevant literature. The search is performed on November, 2025, using the following query:

(TITLE-ABS-KEY (("deep learning" OR "artificial intelligence" OR "machine learning") AND ("healthcare" OR "health care" OR "medicine" OR "clinical" OR "patient monitoring" OR "clinical

decision support")) AND PUBYEAR > 2018 AND PUBYEAR < 2026 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (DOCTYPE , "cp") OR LIMIT-TO (DOCTYPE , "ar") OR LIMIT-TO (DOCTYPE , "ch") OR LIMIT-TO (DOCTYPE , "re") OR LIMIT-TO (DOCTYPE , "cr") OR LIMIT-TO (DOCTYPE , "bk") OR LIMIT-TO (DOCTYPE , "ed") OR LIMIT-TO (DOCTYPE , "sh") OR LIMIT-TO (DOCTYPE , "no") OR LIMIT-TO (DOCTYPE , "le"))

This query is designed to identify research papers specifically focused on the application of DL techniques in the context of Healthcare 5.0, targeting relevant terms in the title, abstract, and keywords. To maintain a focused yet interdisciplinary scope, publications are limited to key subject areas such as Computer Science, Engineering, Medicine, Health Professions, and Decision Sciences domains that actively contribute to the advancement of Healthcare 5.0. Moreover, standard journal articles, reviews, books, chapters, and conference papers are considered to ensure the inclusion of peer-reviewed, impactful contributions.

VOSviewer [37], a Java-based software tool widely used for constructing and visualizing bibliometric networks, is employed to perform a co-occurrence analysis of keywords extracted from 1,342 Scopus-indexed publications relevant to DL in Healthcare 5.0. Figure 2 presents the resulting keyword co-occurrence network, where each node represents a keyword appearing at least five times in the reviewed literature. The size of each node reflects its frequency, while the distance and thickness of connecting lines indicate the strength of co-occurrence relationships. This visualization provides meaningful insights into dominant research areas and the conceptual connectivity within the field. The network visualization in Figure 2, generated by VOSviewer, highlights distinct thematic clusters in deep learning research applied to healthcare. Each color represents a major domain and its related concepts.

A. Red Cluster: This prominent cluster centers on "deep learning," "electronic health record," "diagnosis," "diseases," and "machine learning," capturing the intersection of medical informatics and AI. It encompasses topics such as disease detection, diagnosis, cardiology, patient treatment, biomedical engineering, clinical research, and healthcare data analytics. This area showcases how advanced ML and neural models are used for disease prediction, EHR analysis, and personalized medicine.

B. Blue Cluster: The green nodes focus on "machine learning," "human," "prediction," "algorithm," and "article," emphasizing clinical decision-making and AI evaluation practices. It features keywords like risk assessment, support vector machine(SVM), diagnostic accuracy, epidemiology, medical imaging, and procedure optimization. TThe networhis cluster relates to research on algorithmic healthcare solutions, comparative trials, and studies involving healthcare personnel and patient outcomes.

C. Green Cluster: The green nodes focus on "computer vision," "image enhancement," "feature extraction," "object detection," and "semantics," emphasizing image analysis and visual recognition technologies. It features keywords like object recognition, benchmarking, CNN, transformer architectures, feature fusion, extraction methods, semantic segmentation, real-time systems, agriculture applications, defect detection, robotics, and precision agriculture. This cluster relates to research on advanced image processing techniques, automated visual inspection systems, and the application of DL models for feature extraction and pattern recognition across multiple domains including medical imaging and smart agriculture.

D. Yellow Cluster: This cluster focuses on "bioinformatics," "computational biology," "molecular docking," "chemistry," and related biomedical research areas. It includes keywords such as molecular dynamics, brain tumor analysis, lesion segmentation, software tools, genomics, and image-based computational methods. This area represents the intersection of AI with molecular sciences and computational biomedicine, showcasing applications in drug discovery, protein analysis, and precision medicine at the molecular level.

Overall, the diagram illustrates extensive interactions among clusters, representing integrated research efforts in medical data science, clinical AI, diagnostics, and patient-centered innovations.

Table 5 summarizes the distribution of documents used for the bibliometric analysis across different publication types. The majority of the documents are journal articles (928), followed by conference papers (276) and reviews (73). A smaller number of book chapters (22), conference reviews (29), editorials (6), and books (8) are also included. This distribution highlights the predominance of journal articles in the research landscape, reinforcing the academic rigour and depth of inquiry within the domain of Deep Learning powered Healthcare 5.0. Table 6 presents the top 10 keywords by occurrence and link strength in deep learning–based healthcare research. “Deep learning” shows the highest frequency and connectivity, confirming its central role, followed

Table 5: Statistics of 1,392 Scopus documents used for the bibliometric analysis 2019-2025

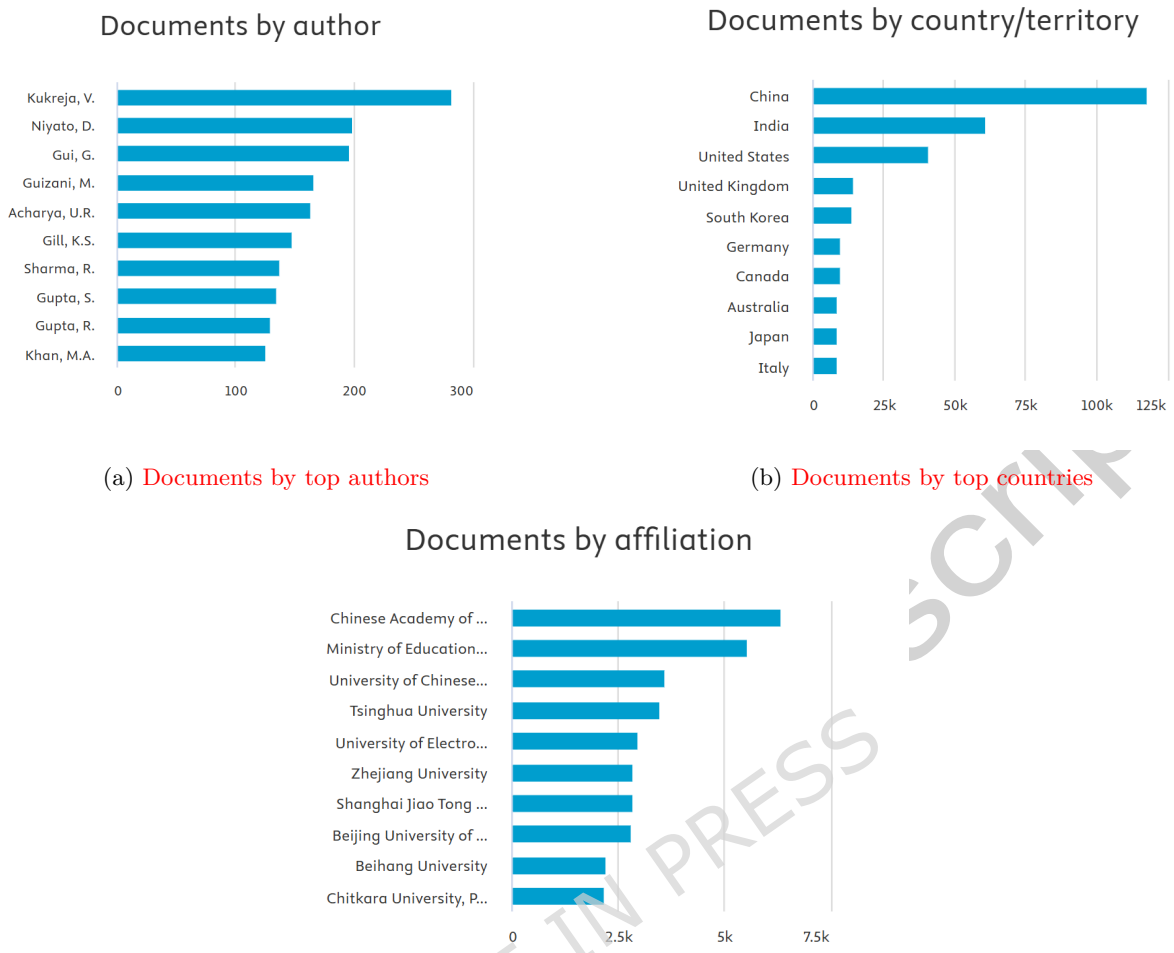
Venue	No. of Documents
Article	928
Conference	276
Review	73
Book Chapter	22
Conference Review	29
Editorial	6
Book	8

Table 6: Top 10 Keywords by Occurrence and Link Strength in Deep Learning in Healthcare

Keyword	Occurrences	Link Strength
deep learning	1299	13889
learning systems	462	5860
human	305	5314
machine learning	221	3115
convolutional neural networks	174	3052
feature extraction	145	2499
diagnosis	146	2311
classification	148	2193
artificial intelligence	140	2069
diseases	101	1747

by related AI terms such as “learning systems,” “machine learning,” and “artificial intelligence.” Keywords like “diagnosis,” “classification,” and “diseases” highlight the dominant clinical applications, while “convolutional neural networks” and “feature extraction” reflect the strong emphasis on image-driven methodologies. Figure 3 collectively summarize global research productivity across authors, countries, and institutions. The first chart, Documents by author, highlights the most prolific contributors, with Kukreja, V. leading by a significant margin, followed by Niyato, D., and Gui, G., indicating a concentration of publications among a few highly active researchers. The second chart, Documents by country/territory, shows that China dominates global research output with the highest number of documents, followed by India and the United States, reflecting the strong and rapidly expanding research ecosystems in these regions. The third chart, Documents by affiliation, further emphasizes China’s prominence, as leading institutions such as the Chinese Academy of Sciences, Ministry of Education China, and major universities like Tsinghua University and Zhejiang University produce the largest volume of documents. Together, these figures demonstrate that research productivity is heavily concentrated in a small group of top authors and is largely driven by major Chinese institutions and national research programs.

Figure 6 illustrates the systematic selection of studies using the PRISMA framework for DL techniques in Healthcare 5.0. Initially, 502 records were identified through comprehensive database searches in Scopus, with no additional records sourced from other channels. After removing duplicates, 389 unique records were retained and screened based on titles and abstracts. Studies were included if they focused on the application of DL, ML, or AI techniques in healthcare contexts (including disease prediction, medical imaging, diagnostics, and patient-centered care), were published between 2019 and 2025, and were peer-reviewed. Exclusion criteria at this stage included non-English publications, editorials without substantial research content, brief conference abstracts, and studies not directly relevant to Healthcare 5.0 or AI-enhanced medical systems. The screening process excluded 71 records that did not meet the initial inclusion criteria, leaving 288 full-text articles for detailed eligibility assessment. The full-text eligibility assessment involved evaluating these 288 articles for methodological rigor, relevance to Healthcare 5.0 paradigms, clarity of research objectives, quality of evidence, and contribution to the field. All assessments and decisions regarding inclusion/exclusion were conducted systematically by the research team. Criteria considered included research design, dataset characteristics, DL architecture employed, clinical application domain, validation methodology, and reported performance outcomes. Only studies meeting



(a) Documents by top authors

(b) Documents by top countries

(c) Documents by top institutions

Figure 3: Bibliometric analysis showing the distribution of documents across top authors, countries, and institutions from Scopus.

these rigorous standards were included in the final qualitative synthesis, ensuring a focused and high-quality set of sources for the comprehensive analysis of DL applications in Healthcare 5.0. The inclusion and exclusion criteria applied during screening and full-text assessment are summarized in Table 7. Each study was evaluated for relevance to Healthcare 5.0 and DL applications, methodological soundness, clarity of objectives, reproducibility, and quality of evidence. Only studies that satisfied these criteria were included in the final review, resulting in 288 articles for in-depth analysis.

Figure 4 shows temporal analysis of publications reveals a consistent upward trajectory in DL research applied to Healthcare 5.0 from 2019 to 2024, followed by a decline in 2025. Starting from a baseline of 22,316 documents in 2019, the field experienced steady growth, reaching 30,703 publications in 2020 (37.6% increase), 39,397 in 2021 (28.3% increase), and 49,803 in 2022 (26.4% increase). The momentum continued with 62,126 documents in 2023 (24.7% increase) and peaked at 66,319 publications in 2024 (6.7% increase), representing a nearly threefold increase from 2019 levels. However, 2025 shows a decrease to 56,795 documents (14.4% decline), which may be attributed to incomplete data collection as the year has not yet concluded at the time of this analysis. The substantial growth from 2019 to 2024 reflects the rapid advancement and increasing research interest in DL methodologies for healthcare applications, driven by improved computational resources, the availability of large-scale medical datasets, the COVID-19 pandemic's acceleration of digital health initiatives, and the demonstrated success of DL models in clinical diagnostics, medical imaging, and personalized medicine.

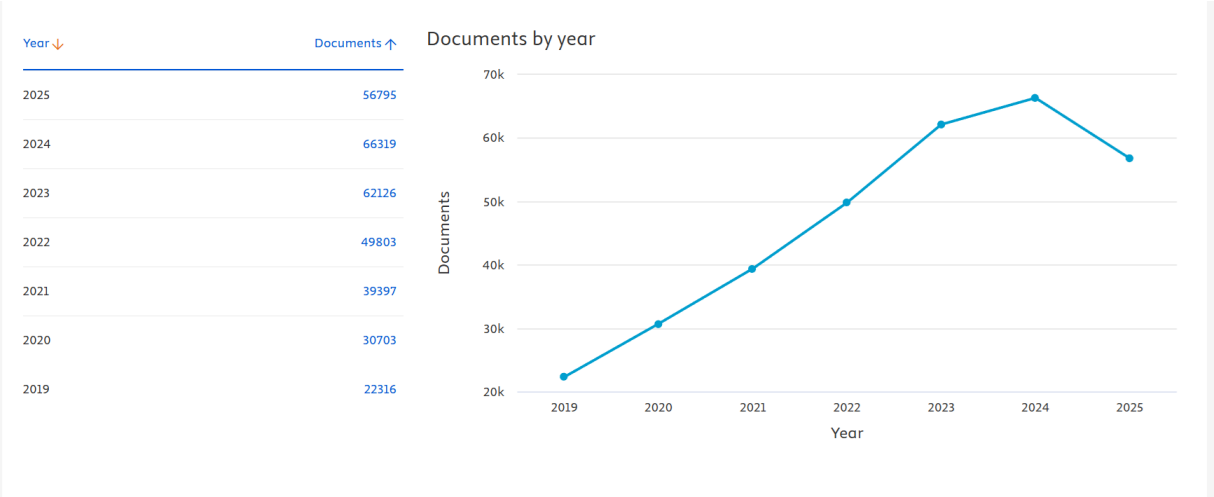


Figure 4: Temporal distribution of publications on deep learning in Healthcare 5.0 from Scopus (2019-2025)

This exponential growth pattern underscores the transformative role of DL in advancing the Healthcare 5.0 paradigm.

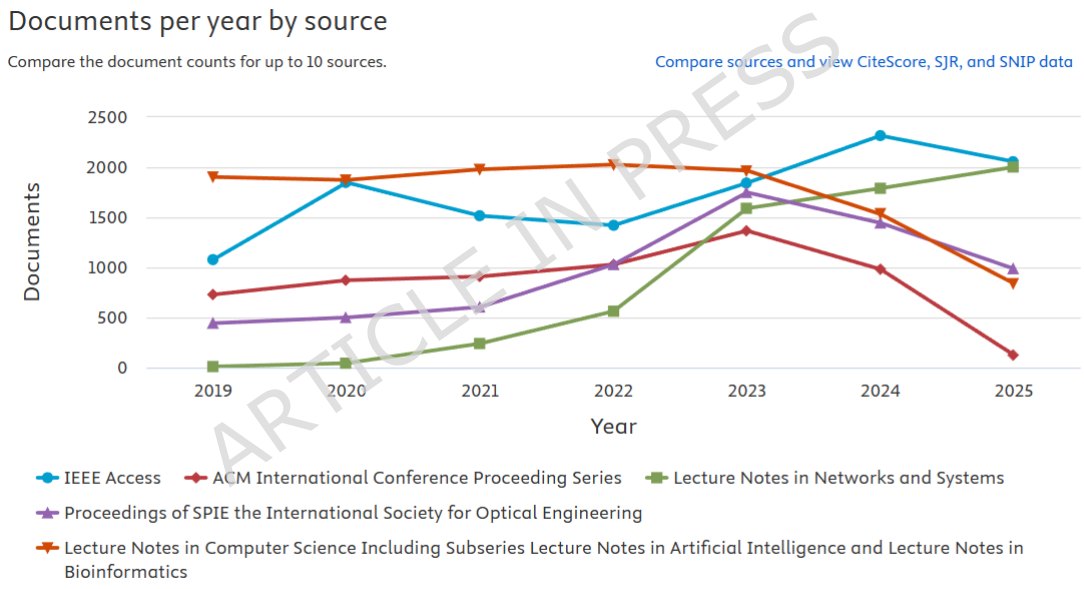


Figure 5: Publication trends by major source venues in DL for Healthcare 5.0 from Scopus (2019-2025)

Figure 5 illustrates the distribution of publications across major source venues reveals distinct patterns in research dissemination channels and evolving publication preferences within the DL healthcare domain. IEEE Access demonstrates the most dynamic growth trajectory, starting at approximately 1,100 documents in 2019, escalating to a peak of 2,300 publications in 2024, before declining to 2,000 in 2025, establishing itself as the leading venue for this interdisciplinary research. Lecture Notes in Computer Science (including subseries in Artificial Intelligence and Bioinformatics) maintained consistently high output between 1,900-2,050 documents annually from 2019 to 2024, with a notable decrease to approximately 900 in 2025, reflecting its role as a stable conference proceedings venue. Lecture Notes in Networks and Systems exhibited the most dramatic growth, emerging from near zero publications in 2019 to surpassing 2,000 documents in 2025, indicating its increasing

prominence as an outlet for systems-oriented healthcare AI research. Proceedings of SPIE (International Society for Optical Engineering) showed steady growth from 450 documents in 2019 to nearly 2,000 in 2025, highlighting the growing intersection between optical imaging technologies and DL in medical diagnostics. ACM International Conference Proceeding Series maintained moderate levels (750-1,400 documents) from 2019 to 2023, before experiencing a sharp decline to approximately 150 in 2025, suggesting a potential shift in publication venue preferences or changes in conference organization. These venue-specific trends reflect the multidisciplinary nature of DL in healthcare, spanning computer science, biomedical engineering, medical imaging, and clinical informatics communities, while also indicating evolving preferences toward open-access and rapidly published research outlets that facilitate faster knowledge dissemination in this fast-paced field.

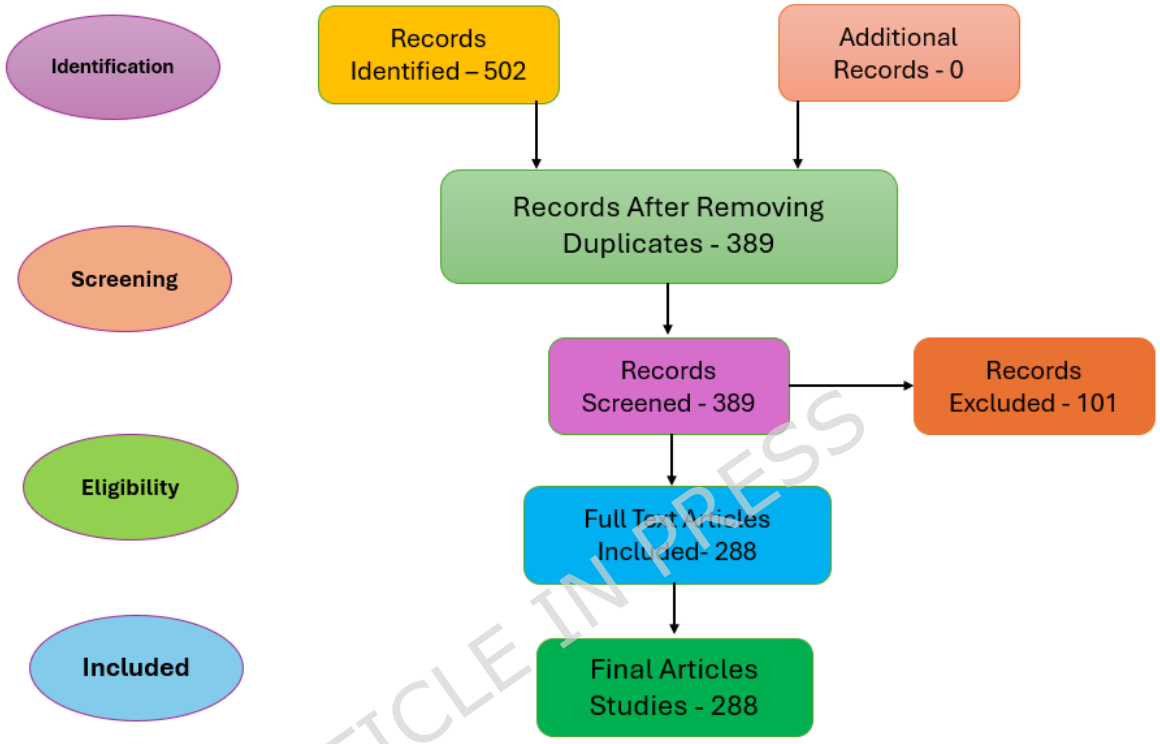


Figure 6: PRISMA flow diagram for the systematic selection of literature on Deep Learning techniques in Healthcare 5.0

Scopus is a widely used and comprehensive academic database, but its coverage carries inherent biases that can influence the results of a literature review. One of the most significant limitations is the strong dominance of English-language publications, which can lead to the underrepresentation of impactful research produced in other languages. This is particularly important for DL applications in Healthcare 5.0, as substantial work from regions such as East Asia, the Middle East, Latin America, and non-English-speaking Europe may not be fully captured. Scopus also tends to prioritize journals with higher citation indices and established publishing practices, creating a bias toward well-resourced institutions and regions with robust research infrastructures, potentially overlooking innovative healthcare AI solutions developed in resource-limited settings or published in regional medical journals. Additionally, the rapid evolution of DL techniques means that cutting-edge developments may first appear in preprints, technical reports, or specialized healthcare AI platforms that fall outside Scopus's indexing scope. Recognizing these limitations allows us to contextualize our bibliometric and PRISMA findings and highlights the need for future reviews to incorporate multilingual databases, region-specific healthcare repositories, and complementary sources such as PubMed, IEEE Xplore, and arXiv to achieve a more globally inclusive and comprehensive perspective on DL innovations in healthcare. To minimize selection bias, two reviewers independently screened titles, abstracts, and full texts using the predefined inclusion and exclusion criteria, with discrepancies resolved through discussion and consensus.

Table 7: Inclusion and Exclusion Criteria for Study Selection in PRISMA

Criteria	**Inclusion**	**Exclusion**
Publication Year	2019–2025	Prior to 2019 or beyond 2025
Publication Type	Peer-reviewed journal articles, conference papers, reviews, and book chapters	Editorials without research content, opinion pieces, short abstracts, preprints, non-peer-reviewed materials
Focus Area	Studies focusing on DL, ML, AI applications in healthcare (disease prediction, diagnostics, medical imaging, EHR analysis)	Studies not directly related to healthcare or AI/DL applications in medical contexts
Healthcare Context	Research applicable to Healthcare 5.0, clinical decision support, patient-centered care, or medical informatics	Non-English publications, studies lacking healthcare relevance
Study Content	Studies reporting empirical results, novel architectures, frameworks, datasets, or case studies relevant to Healthcare 5.0 with clear methodologies	Studies lacking methodological detail, clear outcomes, validation, or reproducibility
Subject Areas	Computer Science, Engineering, Medicine, Health Professions, Decision Sciences	Studies from unrelated domains without healthcare applications

2.1 Summary of included studies across analytical categories

Table 8 presents the distribution of the included studies across major analytical categories, highlighting the relative emphasis of research within the surveyed literature. Among the core domains, Medical Imaging constitutes the largest proportion (27.1%, n=78), reflecting the extensive application of deep learning techniques in image-based diagnosis, including modalities such as MRI, CT, X-ray, and histopathology. This is followed by Disease Prediction (22.2%, n=64), which encompasses a wide range of clinical applications such as cardiovascular disease, diabetes, Alzheimer's disease, cancer prognosis, and IoT-enabled remote patient monitoring, indicating a strong focus on predictive analytics for early diagnosis and risk assessment. Multimodal Learning accounts for 16.3% (n=47), demonstrating a growing trend toward integrating heterogeneous data sources—such as electronic health records, medical imaging, and clinical text to improve diagnostic accuracy and decision-making. Notably, a substantial portion of the literature falls under General/Supporting studies (34.4%, n=99), which include foundational topics such as explainable AI, deep learning architectures, ethics, fairness, and federated learning, underscoring the importance of methodological and systemic advancements in healthcare AI. Overall, the table reveals a balanced yet evolving research landscape, with strong dominance in imaging and predictive modeling, alongside increasing attention to multimodal integration and supporting frameworks essential for real-world deployment.

3 Background

To provide foundational context for the subsequent discussions, this section outlines the fundamental deep learning architectures and supporting technologies relevant to Healthcare 5.0 applications. DL structures have fundamentally transformed artificial intelligence, enabling models to achieve record-breaking performance across a wide range of applications. In this section, we give an overview of the temporal analysis of the DL studies in healthcare, evolution trend with the initial development of rudimentary statistical and rules-based models, to complex, MMDL models and most important structures that have dominated contemporary AI: Generative Adversarial Networks (GANs), Transfer Learning models, transformers, federated learning, and multimodal large language models. Figure 7 shows various commonly used DL models in the healthcare domain.

Table 8: Distribution of included studies by major analytical category

Category	Scope / Representative Topics	Studies (n)	% of Total
Disease Prediction	Cardiovascular, diabetes, Alzheimer's/dementia, kidney disease, arrhythmia detection, chronic disease risk, cancer prognosis, remote patient monitoring, IoT-based prediction	64	22.2%
Medical Imaging	Computed Tomography (CT), MRI, X-ray, mammography, fundus/retinal imaging, ultrasound, histopathology; tumour/lesion detection & segmentation, computer-aided diagnosis, COVID-19 imaging	78	27.1%
Multimodal Learning	EHR + imaging fusion, text–image models, multi-source clinical data, self-supervised & transformer-based multimodal architectures, multimodal disease diagnosis & prognosis	47	16.3%
General/Supporting	*XAI methods, DL architectures, ethics, fairness, federated learning, methodology, healthcare systems background*	*99*	*34.4%*
Total		**288**	**100%**

3.1 Historical Progression and Year-by-Year Research Evolution

Prior to 2016, the dominant ML models in healthcare analytics were convincingly standard statistical models and naive ML methods, like logistic regression, decision tree, and SVM. These models worked with structured data but had difficulties with unstructured data with high dimensions such as medical images or EHRs. The major constraints were feature engineering needs and the challenge in capturing non-linear, complex relationships [215].

DL (and, in particular, CNNs in medical image and RNNs in sequential data (e.g., EHRs)) was rapidly adopted during the period between 2016 and 2018. Major papers proved that DL is better at image classification, image segmentation, and disease prediction. Indicatively, radiology and pathology have seen the use of CNNs that contributed to the breakthrough in automated detection of a disease such as cancer and diabetic retinopathy. This period was marked by the shift of the hand crafting features towards the end-to-end learning to allow more robust and scalable solutions [225].

From 2019 to 2021, research expanded beyond single-modality data to multimodal fusion, integrating imaging, genomics, and clinical notes. Attention mechanisms and transformers started gaining traction, allowing models to weigh the importance of different data sources and time points. For instance, models like Timeline introduced time decay factors and attention to capture disease progression patterns in EHRs, enabling more interpretable and temporally-aware predictions. This period also saw increased focus on interpretability, fairness, and ethical considerations, as deep learning models began to be deployed in real-world clinical settings [21, 132].

Recent years have witnessed the integration of large language models (LLMs) and foundation models, enabling zero-shot and few-shot learning for healthcare applications. There is a strong emphasis on real-time inference, federated learning for privacy-preserving analytics, and edge computing for deployment in resource-constrained environments. The focus has also shifted toward personalized medicine, with deep learning models tailored to individual patient trajectories and multimodal data streams. Additionally, longitudinal bibliometric studies highlight a surge in publications, collaborations, and interdisciplinary research, indicating a maturing ecosystem[199, 269].

3.2 Convolutional Neural Networks (CNNs)

CNNs are applied in healthcare domain due to their ability to automatically generate clinically relevant features, local spatial dependencies, large high-dimensional data, and reliable and interpretable performance across a variety of imaging modalities. For example, CNNs are extensively used in healthcare with applications in oncology [57], tumor detection [126], and disease stage classification [157], among others.

3.3 Recurrent Neural Networks (RNNs)

RNNs are especially well applied to clinical problems that deal with sequential or time- dependent or variable data like patient history, physiology, and clinical text. Their memory, dependency modeling, and future health outcome forecasting capabilities render them essential in time-sensitive medical activities such as modeling the course of diseases, real-time monitoring, and predictive analytics. For example, they are used in Alzheimer detection [180], detecting medical events [110], etc.

3.4 Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs)

LSTM and GRU are enhanced variants of RNN, which are designed to overcome the vanishing gradient problem. This makes them more suitable for capturing long term temporal dependencies in data. They find their applications in the healthcare domain, and are often used for ECG signal analysis [154], predicting cardiovascular health risk [107], patient monitoring, and prediction of the disease progression, and so on.

3.5 Generative Adversarial Network (GAN)

A GAN is a Neural Network architecture, specifically consisting of competing components: a generator (G) and a discriminator (D). The generator creates imitations, and the discriminator predicts if it is real or generated [52]. It was first introduced in the groundbreaking study "Generative Adversarial Nets" by Goodfellow et al. in the year 2014 [80]. The G and D networks play the minimax game. D tries to maximize the capability of correct prediction, 1 for real and 0 for fake, whereas G produces data to minimize D's ability to distinguish real from fake [145].

GAN finds applications in various realms of healthcare. This includes generating synthetic data as the imitated ones are very similar to the real ones, as discussed in [10]. It is commonly used for data augmentation. Yang et al. used TS-GAN (Time series GAN) to generate data for augmentation, as sensor data is not enough, and the model is able to produce data considering time and space effects [275]. Vatanparvar et al. in their study proposed a model using GAN, which hides human speech from recorded data by producing sounds similar to humans, thus preserving privacy [256]. Equation 1 represents how G tries to minimize losses and produce near results, and D tries to maximize the probability of correctly predicting real from fake.

Although the application of GANs in healthcare is very extensive in terms of data generation and augmentation, it carries with it significant risks that should not be overlooked. Mode collapse is a significant problem, in which the generator synthesizes a restricted range or repetitive patterns of the data distribution as opposed to the entire variety of the data distribution, decreasing the utility of the synthetic samples [80, 210]. The generation of hallucinated or anatomically implausible medical images is another issue, which might visually realistic yet contain errors that can mislead the further diagnostic systems [277]. These shortcomings make it evident that implementing GAN-based synthesis in healthcare settings requires strict validation, review by domain experts, and quality-control strategies.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \tag{1}$$

Where:

G = Generator network,
D = Discriminator network,
$p_{data}(x)$ = Distribution of true data
$p_z(z)$ = Prior distribution of input noise,
x = Sample of real data,
z = Random noise vector,
$D(x)$ = Discriminator probability that x is real,
$D(G(z))$ = Discriminator's probability that the generated sample is real,
$\mathbb{E}$ = Expectation.

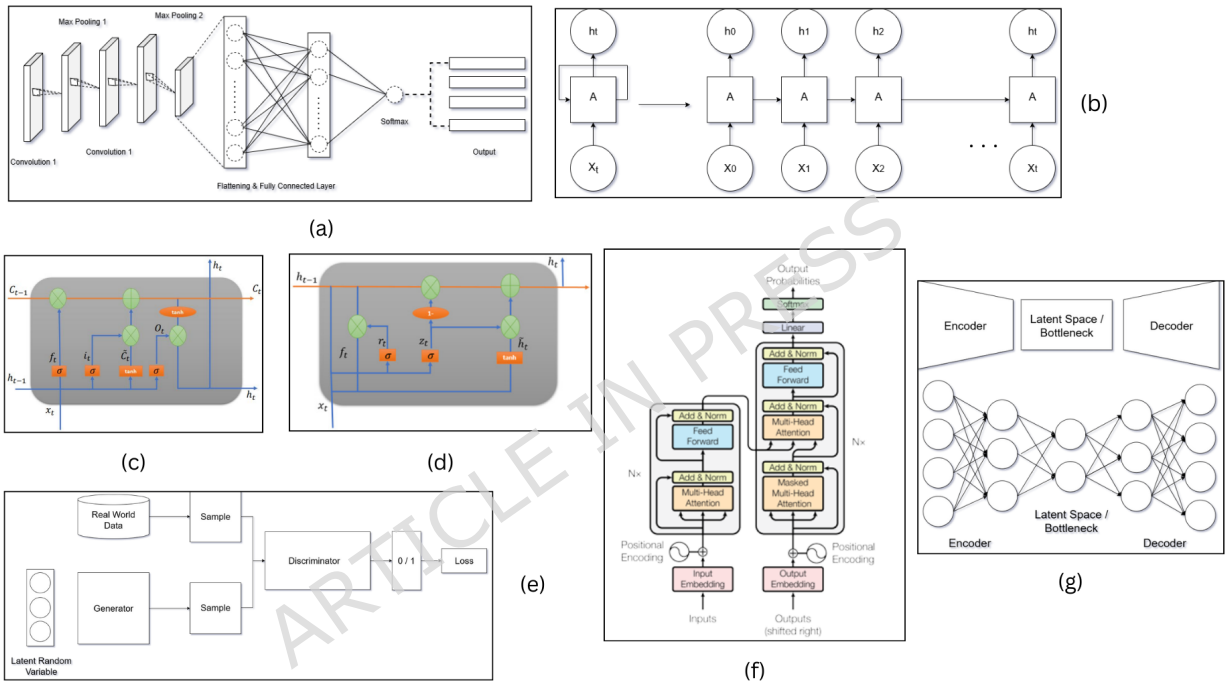


Figure 7: Commonly used DL models in the healthcare domain, (a) Diagram of a CNN, (b) Diagram of a RNN rolled out to time, (c) Diagram LSTM cell (reproduced from Mateus et al., 2021 [163] licensed under CC BY 4.0), (d) Diagram of GRU cell (reproduced from Mateus et al., 2021 [163] licensed under CC BY 4.0), (e) Diagram of GAN architecture, (f) Transformer architecture (Reproduced from Vaswani et al., 2017, [255] at Google, used under permission for scholarly activities.), (g) Diagram of Autoencoder model.

3.6 Transformer architecture

The transformer model was proposed by Vaswani et al. [255] in their seminal paper titled *Attention Is All You Need*, which brought a great shift in the way researchers think. It is an encoder-decoder model, and mainly focuses on the self-attention mechanism. Equation 2 represents the attention mechanism in transformers, compares the queries with keys to find out the most relevant information, and then takes a weighted sum of values, so the output is focused on the important parts only. It is capable of parallel processing, and multi-headed attention allows understanding different contexts of the same input. It has brought a revolution in natural language understanding and processing. The model components include encoding and embeddings of

input, which are passed to the encoder. The computer understands numbers and not words. Therefore, this step is crucial. Positional encoding allows us to generate vectors based on their position. The encoder model has a multi-head self-attention mechanism. This generates attention vectors that are fed to feed feed-forward Neural Network. Residual connections are added, and layer normalization is performed. The output and positional encoding are fed to the decoder. It has three components: masked multi-head attention, cross multi-head attention, and a position-wise feed-forward layer, two of them similar to the encoder. Masked attention masks the next words in the sequence to prevent the model from peeking, and the results depend on the outputs before the 'i'th stage. The residual connections and layer normalization steps follow. Finally linear and Softmax layers of the model. Softmax produces output in the form of probabilities for the next token.

Dosovitskiy et al., 2020, [61] introduced Vision Transformers in the landmark paper 'An image is worth 16 X 16 words' and compared its performance to CNNs. This is DL model used for computer vision tasks. The model breaks images into patches which are analogous to tokens in Natural Language Processing (NLP) tasks. These are flattened into vectors, positional encodings are added, and passed into self-attention and feed-forward blocks. Finally a class token collects information from all parts of the image, and forms a condensed representation, which is used in the final layer for prediction. This model process the entire image at once instead of scanning as done in CNN models. Its contribution healthcare includes automatically detecting diseases using X-rays images [228], detecting [188] and classifying types of cancer [8] among others.

In recent work, it has been demonstrated that transformer based architectures tend to be more effective at disease prediction than CNNs and RNNs due to their ability to treat long-range correlations, heterogeneous data, and scale to large datasets. Although CNNs are still useful in local image pattern, they can be challenging in terms of global context and intermodal association. RNNs can be applied in sequential data, but they have related problems like vanishing gradients and lack of parallelization. Transformers overcome these constraints by using self-attention, which allows enhancing feature extraction on images, text, signals, and EHR data. It renders them highly applicable to multimodal and context-aware Healthcare 5.0 applications.
Transformers are especially well-suited to clinical work since they can capture long-range dependencies through attention, combine multimodal data, transfer learners, scale effectively, and be explainable. This renders them suitable for EHR analysis, clinical text mining, multimodal fusion, drug discovery [129], and medical imaging [72] of Healthcare 5.0.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2}$$

Where:

$$\begin{aligned} Q &= \text{Query matrix} \\ K &= \text{Key matrix} \\ V &= \text{Value matrix} \\ d_k &= \text{Dimensionality of keys (needed for scaling)} \\ QK^T &= \text{Dot product of queries and keys(for measuring the similarities)} \\ \sqrt{d_k} &= \text{Scaling factor} \end{aligned}$$

3.7 Autoencoders

Autoencoders are an unsupervised Neural Network model. It is used to learn data representations efficiently, i.e., they try to represent data in as few features as possible. The architecture of an autoencoder consists of three main components: the encoder, the bottleneck (also known as the latent space), and the decoder. The task of the encoder is to compress the input data into a representation that consists only of the essential features. This representation is called the latent vector. This creates a bottleneck as the size is significantly reduced from the input to this vector. This is the smallest layer. Finally, the decoder reconstructs the input data from the latent vector representations.

3.8 Transfer Learning

The term transfer learning became popular after the survey by Pan and Yang published in 2010 [189]. It is a technique that we often use when we have available knowledge in a domain, which we use in a different but related task. This is useful as it reduces the need for having a detailed dataset for each task and also significantly saves computation time by reducing the need for training from scratch. We can either use the whole model and just change the output target, or use only part of the base model and change it based on how similar the task is compared to the base model task. The two important terminologies in Transfer Learning are: Pre-training and Fine-tuning. Pre-training is the process of training a model on a large, generic dataset. This sets the parameters in the model. Fine-tuning, on the other hand, is using these pre-trained models to provide results for a very specific task or problem. They are mainly of three types: Inductive, Transductive, and Unsupervised. Commonly used Transfer Learning models include: Residual Networks (ResNet) (image classification), MobileNet (light-weight model for mobiles) for image data, Generative Pre-trained Transformer (GPT) (text generation), BERT (text classification), BART/T5 (summarization) for NLP, etc. Transfer Learning is most suited in healthcare due to the reduced reliance on large annotated datasets, reduced cost of training, enhanced generalization, and the ability to adapt to domains, which makes it extremely useful in practical healthcare tasks such as medical imaging [112], detecting diseases [221], drug target interactions [284] and multimodal analytics.

3.9 Federated Learning

Federated learning was introduced in a paper in 2018, authored by Google researchers Hard et al. [90], who applied it to next-word generation for mobile keyboards. This is a collaborative method of training ML models, in which model goes to data, instead of collecting data in a central location. This type of learning is usually done to comply with data privacy and sharing norms of the client. Every client learns from data separately, and trains a model. The learning and updates of each model is sent to a central server, which is combined to improve the global model. This cycle is repeated multiple times. However, decentralization does not completely eliminate data security and privacy risks. Common threats to privacy include:

1. Gradient Leakage Attacks: An adversary observes and analyses the shared gradients. The attacker initializes a dummy input and iteratively updates the gradients to minimize the difference between the dummy input and to those of the original victim data. This can be curbed by gradient sparsification, which involves sending only the most important gradients from client to the server.

2. Membership Inference Attacks: The attacker tries to determine if a specific data record was used in the training set and potentially exposing the identity of the target. Differential privacy can help prevent such attacks by ensuring that no individual data point significantly affects the final model.

3. Data Reconstruction Attacks: Attackers can reconstruct data samples from shared gradients using GAN. Differntial privacy also helps mitigate such attacks.

In healthcare domain, federated learning is used to collaboratively train models from disease prediction, while keeping patient data private within the premises [108].

3.10 Multimodal transformer model

A transformer model consists of an encoder-decoder architecture. The encoder maps the input into a sequence of representations, and the decoder generates an output sequence from those representations. A multimodal transformer architecture takes inputs of multiple modalities and converts the input tokens and embedding. It then performs self and cross attention. Self-attention is the process of finding relations within the same modality, while cross-attention is finding relationships across modalities. Multimodal transformer architecture can be single-stream, multi-stream, and hybrid-stream. In the single-stream model, all modalities are concatenated in one sequence. In a multi-stream architecture, separate transformers are used per modality. A hybrid-stream architecture is like the mix of single and multi-stream architectures [270].

In healthcare, it has been used to combine multiple modalities of data and make accurate predictions even if some data is missing. Multimodal Transformers are the future of health care because they combine and harmonize different clinical data (images, text, signals, EHRs) to make more accurate, holistic, and individual decisions in Healthcare 5.0 [146].

3.11 Multimodal Large Language Models

Instead of relying on text-only inputs, multimodal LLMs can reason and process multiple data types at the same time, thus, providing more context aware responses. They combine multiple types of neural networks into one unified model, and each modality is processed by a specialized encoder. These encoders convert the inputs to numerical representations, which are brought to the shared representation space to interact with each other, and jointly interpret the data.

Healthcare and medicine are also multimodal domains. Google DeepMind and Google Research researchers proposed the Medical Pathways Language Model Multimodal (Med-PaLM M)[251], which is capable of interpreting imaging data, clinical text and genomics. Med-PaLM M also outperforms specialist model and is clinically relevant. Large Language and Vision Assistant for BioMedicine (LLaVa-Med), by Microsoft [139] is a conversational multimodal assistant. This model is cost efficient and is trained to comprehend medical image data, and answer questions about them. These models demonstrate the growing demand of multimodal LLMs, and encourages further exploration in healthcare domains.

3.11.1 Segment Anything Model

Segment Anything model has been developed by Meta. It is an groundbreaking image segmentation model AI model. It performs zero-shot generalization which means can segment objects from images and videos, which it has not seen before. It has been trained on SA-1B dataset, also introduced by Meta, specifically for this task. This is the largest dataset for image segmentation to-date, with over 1.1 billion segmentation masks and, over 11 million images. It even surpasses supervised learning technique results in many cases. It is "promptable" segmentation model, hence it allows users to generate valid masks based on prompts, and with minimal human input. It architecture includes an image encoder, a prompt decoder and a mask decoder, enabling the mask computation. Mazurowski et al. [164] tested this model against 19 medical imaging datasets and concluded that its performances was variable. On some datasets it performed exceptionally well, while on other, it was moderate or even poor. Therefore, it has to be studied well before applying it in healthcare domain. Dong et al. [60] adapted SAM for medical imaging and introduced EMedSAM, which outperformed state-of-the-art models in lungs and multiple organ segmentation.

3.12 Evaluation metrics and regulatory considerations

The metrics used to evaluate DL models in healthcare should be able to measure an aspect of predictive accuracy and clinical reliability. Such common measures are accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUROC) which is frequently used to estimate discrimination performance in clinical prediction problems. [197, 59].Standard metrics are used to test binary or multi-class predictions in disease diagnosis and early detection.

The accuracy represents the ratio of the samples that have been classified correctly but can be misleading in case of imbalanced samples [233]. Precision is a measure of consistency of the model to detect the presence of positive cases, and is defined as the fraction of times that the model is correct [196]. Recall (sensitivity) is used to demonstrate the ability of the model to represent true positive cases, which is essential when a missed cases can be risky for disease diagnosis. The F1-score is a measure that balances between recall and precision and applies to cases where the distribution of the classes is not equal or the costs of errors are uneven. The ROC curve illustrates the trade-off between sensitivity and specificity across thresholds, while AUC indicates how well a model separates classes independent of decision boundaries [66]. Lastly, The confusion matrix generates a total breakdown of prediction results and helps identify systematic model errors. It assists in identifying the weakness of the model and it assists in making specific changes. These measurements jointly provide a holistic view of model performance, and allow more objective and clinically useful analysis of DL systems in Healthcare 5.0. Loss is used as a guide to model training by quantifying prediction errors in optimization, e.g. cross-entropy or mean squared error [79].

The analysis of medical imaging is based on out-of-classification segmentation and detection measures.In radiology, tumors or organs Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) measure overlap in radiology segmentations [103]. Lesion detection in CT/MRI with sensitivity and specificity and mean Average Precision (mAP) of object detection tasks [172]. These measures point out the compromises in Healthcare 5.0, where the high-resolution 3D imaging requires the effective calculation[175].

Multimodal fusion needs to be evaluated in a composite fashion. Concordance Correlation Coefficient (CCC)

measures continuous predictions such as times of survival [225]. Calibration measures (e.g., Brier score, Expected Calibration Error) are used to make sure that the predicted probabilities are consistent with the results which is crucial when considering clinical decision-thresholds [1]. New metrics encompass fairness-conscious (demographic parity) and uncertainty quantification (predictive entropy) to successful Healthcare 5.0 implementation [64]. In addition to the performance indicators, healthcare AI should be in line with the regulatory requirements. The guidelines on Good Machine Learning Practice (GMLP) by the U.S. FDA focus on the quality, transparency, and sound evaluation of the various phases of the model, whereas the 2021 guidelines by the WHO center on safety, fairness, and responsibility of clinical AI implementation [186]. Including these considerations provides the foundation for understanding how the models are evaluated before they can proceed to the real world.

3.13 Model auditing frameworks for clinical AI

The auditing frameworks have emerged as one of the significant components of clinical AI since they assist in ensuring that the models act consistently with various groups of patients and in various clinical contexts. Model audits involve systematic checks for bias, robustness, explainability, and data governance to ensure that systems behave consistenly within patient subgroups and across different clinical conditions. There are a number of proven structures that aid in this process. The proposed regulatory framework of the FDA of AI/ML-based Software as a Medical Device includes Predetermined Change Control Plan to guarantee the continued monitoring and safety auditing after deploying the system. The regulatory guidance issued by World Health Organization (WHO) on the international level identifies such measures as independent model assessment, documentation requirements, and ongoing audit cycles as the key protections of clinical AI [186]. Tools that are research-driven also aid in this process. The Model Cards offered by Google are used to standardize its transparency reports, whereas the IBM's AI Fairness 360 toolkit analyze the bias and performance of their models across subgroups in a structured way [169, 27]. These models assist in connecting model creation to reliable clinical practices.

3.14 Explainability in Clinical Deep Learning

Explainability in clinical deep learning can be broadly categorized into different approaches based on their design principles and application contexts.

3.14.1 Categories of Explainability Approaches

Explainability assists clinicians to recognize DL models arrive at their predictions. It consists of two primary strategies: interpretable-by-design models and post-hoc strategies that study black-box systems in terms of feature attributions, saliency maps, concept scores, counterfactuals, or surrogate rules [207, 153]. Explainability aids in the auditing of models, emphasis of failure modes, and transparency in clinical contexts in Healthcare 5.0 [169, 12].

3.14.2 Model-specific explainability techniques

Various model families depend on different explanation approaches. Grad-CAM, saliency, and LRP are common CNNs tools designed to demonstrate which parts of an image that impacting a diagnosis [218, 211]. Temporal attention and relevance curves are explained by sequential models like RNNs, LSTMs, and GRUs that rely on risk prediction after clarification by clinical events. The attention-map inspection and token-level attribution are used to detect influential text or image patches by transformers [45]. Autoencoders have been described using reconstruction-error maps and analysis of latent space, whereas GANs are used to inspect latent space and measure diversity to identify unrealistic generations [70, 277]. The transfer learning models have the advantage of layer-wise attribution in order to distinguish between pretrained and fine-tuned behaviors [189]. Federated learning involves site-level aggregation of explanations in order to identify heterogeneity without information sharing. Cross-modal attention and multimodal counterfactuals are used by multimodal transformers and LLMs in demonstrating the impact of each modality on a decision [250, 16]. SAM and foundation segmentation models are based on interpretable masks that can be directly checked by clinicians [128]. GNNs are described by subgraph and node-importance attribution to show the relational predictors of predictions [278].

3.14.3 General-purpose explainability methods

In all these types of models, general methods like Local Interpretable Model-Agnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP), Integrated Gradients, Testing with Concept Activation Vectors (TCAV), surrogate rules and counterfactual explanations, have remained the focus to generate legible, clinician-friendly interpretations [207, 127, 242, 260]. Table 9 summarizes key explainability techniques used in healthcare deep learning, along with their model compatibility, typical applications, and associated limitations.

Table 9: Key explainability techniques in healthcare deep learning: methods, applications, and limitations

Method	Type	Model used	Typical application	Key limitation
SHAP	Post-hoc, Model-agnostic	CNN, RNN, Transformers	Feature importance in disease prediction	High computational cost
LIME	Post-hoc, Model-agnostic	CNN, RNN, Multimodal	Local prediction explanation	Sensitive to parameter settings
Gradient-weighted Class Activation Mapping (Grad-CAM)	Post-hoc, Model-specific	CNN	Region visualization in MRI, CT, X-ray	Produces coarse heatmaps
Integrated Gradients	Post-hoc, Model-specific	CNN, RNN, Transformers	Feature attribution in structured data	Requires baseline selection
TCAV	Post-hoc, Concept-based	CNN	Concept-level clinical interpretation	Requires curated concept data
Saliency Maps	Post-hoc, Model-specific	CNN	Pixel-level importance visualization	Noisy outputs
Layer-wise Relevance Propagation (LRP)	Post-hoc, Model-specific	CNN, DNN	Layer relevance explanation	Sensitive to architecture
Deep Learning Important FeaTures (DeepLIFT)	Post-hoc, Model-specific	CNN, RNN	Neuron contribution analysis	Baseline dependency
Attention Mechanisms	Interpretable-by-design	RNN, Transformers, Multimodal	Identify important features/modalities	May not reflect true importance
Counterfactual Explanations	Post-hoc, Instance-based	CNN, RNN, Multimodal	Minimal-change decision explanation	Hard to ensure realistic outputs
Surrogate Models	Post-hoc, Model-agnostic	Any black-box model	Simplified interpretable approximation	Reduced model fidelity
GNN Explainability	Post-hoc, Model-specific	Graph Neural Networks	Node/subgraph importance analysis	High computational complexity

4 Deep Learning in Healthcare 5.0

Building upon the foundational concepts presented in the Background section, the following section reviews the major applications of deep learning in key Healthcare 5.0 domains. Deep Learning in Healthcare 5.0 represents the integration of advanced AI techniques with next-generation healthcare systems, emphasizing personalization, real-time decision-making, and human-centric care [74, 86]. Unlike previous phases, Healthcare 5.0 leverages DL models, such as CNNs, RNNs, and transformers [30, 280] for precise diagnostics, medical image analysis, predictive modeling, and patient monitoring, often powered by big data from IoT devices and EHRs[42, 176]. It promotes smart, connected, and autonomous healthcare ecosystems, enabling early disease detection, robotic-assisted surgery, and intelligent health assistants, all while ensuring patient data privacy and ethical AI deployment [198].

Table 10 presents an organized (year-wise) summary of the exhibited works for a concise overview; Tables 11 present the details of the datasets utilized. For clarity, the dataset descriptions are divided across three tables. Additionally, Table 17 summarizes the challenges and limitations reported by the authors, along with the identified advantages and disadvantages of the studies. A taxonomy of deep learning architectures commonly used in healthcare, including CNNs, RNNs, Transformers, and multimodal fusion models is represented. This taxonomy gives an overview of the structure of the reviewed model families in the subsequent subsections and is shown in Figure 8.

Explainable AI (XAI) techniques are incorporated to enhance the interpretability and transparency of DL models in Healthcare 5.0. Likely common examples of XAI tools include LIME, SHAP, Grad-CAM, and DeepLIFT , which motivate interpretability by highlighting which aspects of the features or input area of the

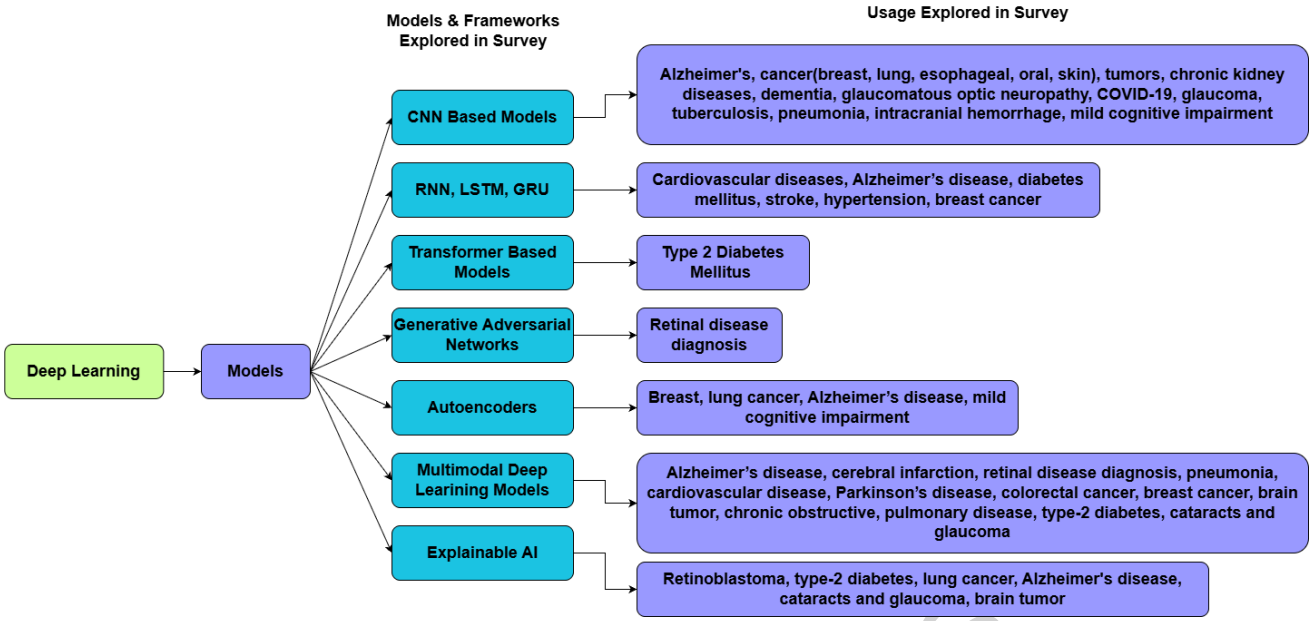


Figure 8: Diagram showing the taxonomy of deep learning architectures in healthcare 5.0, illustrating the core deep learning models and frameworks and their specific usage.

model have the most significant influence regarding predictions. For instance, Grad-CAM identifies important areas in medical images that can be used to make a diagnosis, whereas SHAP and LIME measure the importance of features in unstructured and structured data. These tools are used by clinicians to comprehend model reasoning, confirm the prediction with clinical knowledge, and gain some confidence in AI-assisted decision-making.

4.1 Disease Prediction and Early Diagnosis

DL is an emerging method that can be used in healthcare to anticipate and detect diseases before they occur and diagnose them at an early stage [19]. It is recognized for its ability to revolutionize the industry by enhancing the precision and timeliness of diagnoses, resulting in enhanced therapy and improved patient care [168, 179]. DL can improve the precise prediction of diseases in patients, leading to enhanced quality of care. Computer-aided diagnosis (CADx) is becoming increasingly vital for disease diagnosis, particularly in regions with restricted healthcare access [43, 93, 245]. Neural networks in DL evolve and enhance with new data, rendering them more effective than conventional algorithms in identifying illnesses such as cancer [234], Alzheimer's disease (AD), heart disease [109, 185, 46], brain tumors, etc. [160]. These AI-driven systems provide enhanced precision, less human error, and economical, accessible healthcare solutions [182, 236]. The continuous progress in CADx significantly influences preventive healthcare and early disease diagnosis, enhancing patient outcomes worldwide [117, 245, 5]. We have showcased recent research aimed at predicting and diagnosing various diseases via diverse deep-learning techniques.

This study presents a review of recent advancements in the field of disease prediction and early diagnosis. Acharya et al. [3] developed a deep CNN model capable of autonomously detecting and classifying electrocardiogram (ECG) signals into five distinct heartbeat categories. Jiao et al. [116] employed a deep CNN to obtain deep representations of breast mass images. Employing a 3D CNN, Backstrom et al. [18] demonstrated excellent performance in detecting AD on a large dataset. Feng et al. [67] tackled the difficulties associated with leveraging 3D-CNN and fully stacked bidirectional LSTM (FSBi-LSTM) for AD diagnosis in the context of limited imaging data. Feng et al. [69] proposed a novel brain imaging framework that integrates 3D-CNN with SVM to improve the diagnostic accuracy of AD. Cinar et al. [50] employed a CNN model to detect brain tumors from MRI scans. Nancy et al. [176] employed bidirectional Long Short-Term Memory (Bi-LSTM) models to effectively predict the risk of heart disease. Similarly, Yashudas et al. [276] utilized Bidirectional Gated Recurrent Units (BiGRU) to enable timely identification, medical intervention, and nutritional guidance for cardiovas-

cular conditions. Mandava and Vinta improved the precision of cardiovascular disease prediction through a DL-based modified DenseNet201 (MDenseNet201) model [161]. Mahmud et al. [158] proposed an interpretable AI framework for diagnosing AD by analyzing MRI data with a modified CNN. Similarly, Aldughayfiq et al. [7] enhanced retinoblastoma diagnosis using InceptionV3 with transfer learning, integrating XAI to improve model transparency. Wani et al. [263] introduced an interpretable hybrid DL approach, ConvXGB, combining CNN and XGBoost for explainable lung cancer detection. In the domain of infectious diseases, Alazab et al. [6] developed an automated CNN-based system to accurately detect COVID-19 from chest X-ray images and forecast future cases, recoveries, and deaths using historical data. Singh et al. [232] designed a DNN for early detection of chronic kidney disease (CKD), identifying key predictive features via Recursive Feature Elimination (RFE). Janghel and Rathore (2021) [113] proposed a CNN-based diagnostic model for early Alzheimer's detection using Functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET) images, applying a 3D-to-2D preprocessing conversion before feature extraction. Tanim et al. [244] employed XAI methodologies through a Deep Neural Network (DeepNetX2) to improve interpretability and accuracy in diabetes diagnosis. Ramu et al. [202] enhanced CKD detection accuracy by adopting a hybrid CNN-SVM framework designed to mitigate overfitting, computational inefficiency, and class imbalance. Shehta et al. [223] developed a classification model for blood cancer, achieving high diagnostic accuracy and supporting early intervention to improve survival outcomes. Ortiz et al. [187] advanced skin cancer diagnosis with an Ensemble DL model that combined multiple CNN architectures to boost classification accuracy on dermoscopic images. Choi et al. [49] improved early heart failure detection by leveraging temporal relationships in EHR. Furthermore, Maji et al. introduced an interactive cognitive assessment platform that integrates modified CNNs (MOD-1D-CNN for health metrics and MOD-2D-CNN for facial image analysis) with gamified evaluation strategies to support the early diagnosis of dementia. Figure 9 provides a consolidated visual summary of how different DL architectures perform across multiple disease prediction and early diagnosis tasks, addressing cross-task performance variation.

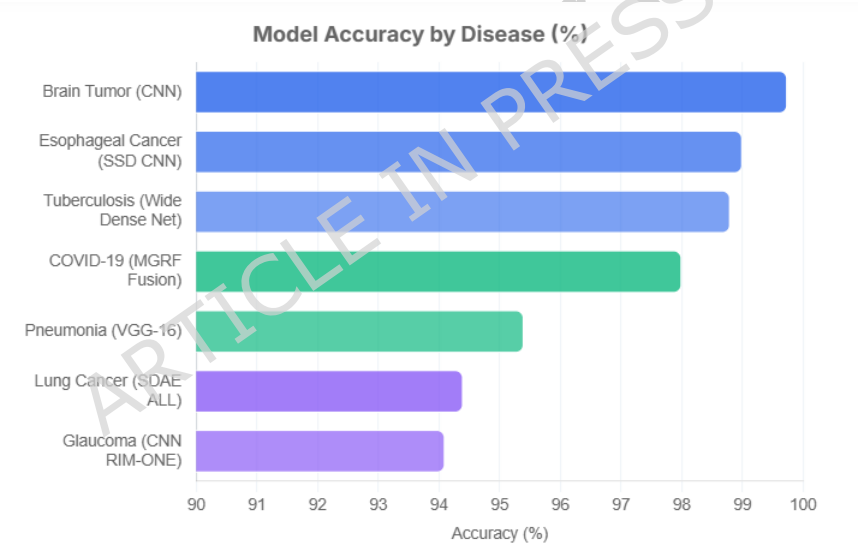


Figure 9: Comparative accuracy of DL based models for disease prediction and early diagnosis across multiple disease conditions

4.2 Medical Imaging Analysis and Radiology

The discovery of medical image analysis came from the invention of the microscope by Leeuwenhoek in the 17th century [166]. Then X-rays and CT scans set the foundation for more advanced systems. With the advancement of technology, AI was integrated into the medical image analysis process. It has become a highly studied and researched field, and large amounts of data known as big data contribute greatly to research [214]. The boom in ML and DL research during the 2010s revolutionized this field; hence, the demand for automated and accurate systems grew. These systems provided interpretations of complex images in the medical domain and were then used in diagnostics to provide better treatment and detect serious issues early on [224, 40]. Figure 10 was

adapted from the work of Anwar et al. [14], which shows the various imaging techniques used in medical imaging.

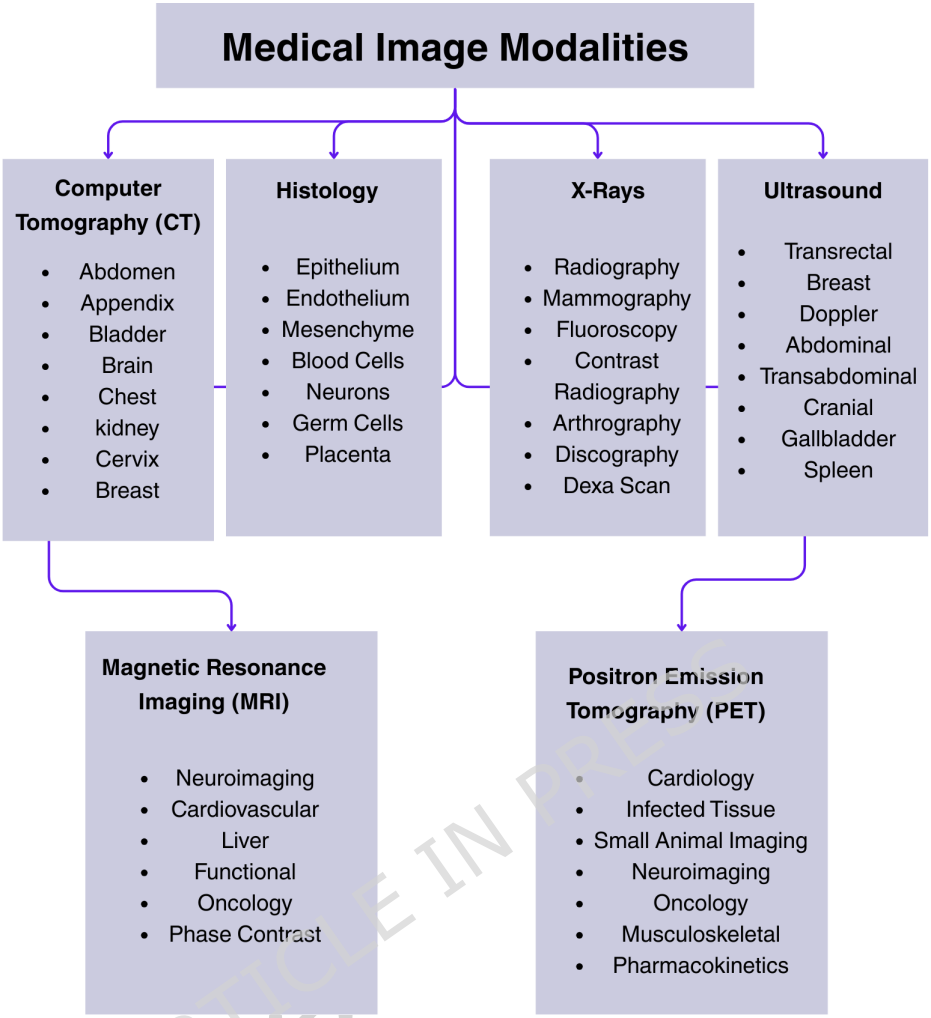


Figure 10: Common modalities used in medical image analysis

Several studies have used different methodologies in the domain of medical image analysis; for example, Hua et al. [101] used DL frameworks to eliminate explicit feature extraction in classifying pulmonary nodules in CT scans as benign or malignant. Gulshan et al. [82] developed and validated a DL algorithm for the automated detection of referable diabetic retinopathy (RDR) and diabetic macular edema (DME) using retinal fundus images. Cheng et al. [47] explored a CADx to simultaneously identify benign and malignant lesions and nodules. Zilly et al. [287] provide a computationally efficient and precise technique for the automatic segmentation of the optic disc and the optic cup of the fundus images to facilitate early detection of glaucoma. Horie et al. [97] evaluated the diagnostic performance of a CNN in detecting esophageal cancer from endoscopic images. Li et al. [147] evaluated the efficacy of a DL algorithm in identifying referable glaucomatous optic neuropathy (GON) using color fundus images. Jeyaraj and Nadar [115] created an automated, computer-assisted diagnostic method for the early identification of oral cancer. Song et al. [235] categorize thyroid nodules as benign or malignant to avoid unnecessary fine needle aspiration (FNA). Khamparia et al. [125] created a DL Internet of Health Things (IoHT)-driven system for automated skin cancer identification and classification from dermoscopic images to improve accuracy and reduce diagnostic errors. Pandit and Banday [190] created a cost-effective, automatic chest X-ray COVID-19 detection model. Kuchana et al. [131] aimed to improve COVID-19 diagnosis and tracking by DL-based lung CT scan segmentation. Elsharkawy et al. [63] developed a computer-assisted diagnostic (CAD) system to assess pulmonary function and predict mortality risk in COVID-19 patients through

chest X-ray analysis. Chattopadhyay et al. [44] created a tumor detection model for speedier diagnosis and treatment. Naz et al. [178] classify pulmonary conditions and explain classification results to help radiologists discover early disease. Huy et al. [105] developed a model to detect tuberculosis with high accuracy, specificity, and sensitivity. Sharma et al. [222] developed a system for early detection and classification of pneumonia using DL and transfer learning using chest X-ray images. Mahmud et al. [156] developed a CNN-based model for detecting brain tumors accurately in the early stages using MRI images. Odusami et al. [184] developed a model for early detection of Alzheimer's disease, combining MRI and PET images for a better analysis. Hroub et al. [99] developed an explainable DL lung disease prediction tool from chest X-rays. Kar et al. [121] employed cutting-edge DL techniques to automatically detect intracranial hemorrhage (ICH). Tawfeek et al. [246] developed a DL model to detect lung cancer.

Huy et al. [105] in their paper, proposed a CBAMWDNET hybrid DL model with two components: Convolutional Block Attention Model and Wide DenseNet(CBAMWDNet). CBAM is an attention mechanism used with CNN to improve model performance by focusing only on important parts and ignoring irrelevant details. This is made possible by channel and spatial attention in CBAM, which highlights which features are important and where in the input data the features are located, respectively. This makes the model interpretable and also reduces overfitting. WDN enables the extraction of detailed features as it widens the layers instead of deepening them.
In short, this approach combines the focusing feature of CBAM and the representation power of Wide DenseNet into a model that produces accuracy as high as 98.8% on different datasets. The architecture of the model has been shown in Figure 11.

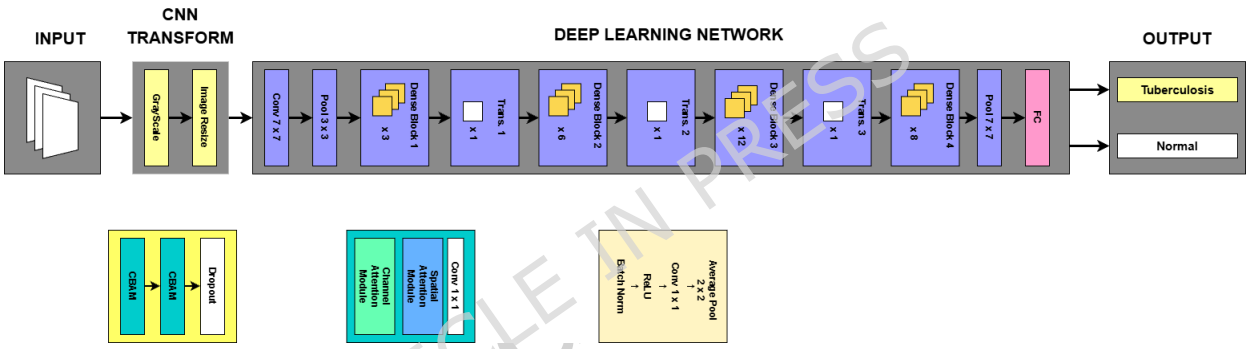


Figure 11: CBAMWDNet Architecture diagram for "An Improved Densenet Deep Neural Network Model for Tuberculosis Detection Using Chest X-Ray Images", Vo Trong Quang Huy; Chih-Min Lin[105]

Elsharkawy et al., in their paper [63] built a computer-aided diagnostic system combining the Markov-Gibbs Random Field (MGRF) model with neural networks to predict the severity of lung infection of patients due to COVID-19. The MGRF model calculates the Gibbs energy at varied distances in infected areas to detect how uneven or abnormal the tissues are. These values are then converted to graphs called the CDF—cumulative distribution functions—which are then input into the neural networks. The neural network analyzes the inputs and classifies the infection as mild or severe. The model achieved an accuracy of 98% and could classify all severe cases correctly during testing. The model's architecture has been represented in Figure 12.

4.3 Multimodal Deep Learning in Healthcare

In the era of smart healthcare, medical systems increasingly rely on diverse data modalities ranging from imaging (e.g., MRI, CT, and histopathology) to EHR, wearable sensor signals, and clinical notes[31].

There are usually three types of multimodal fusion strategies adopted by MMDL. The input stage in early fusion takes raw or low-level data across several modalities and thus, the model learns joint representations early on. Intermediate fusion takes a combination of features that are independently learned across both modalities and provides a balance between common learning and representations of the modalities. Late fusion is an algorithm that combines the end predictions or decision scores of individual models, and is applicable when

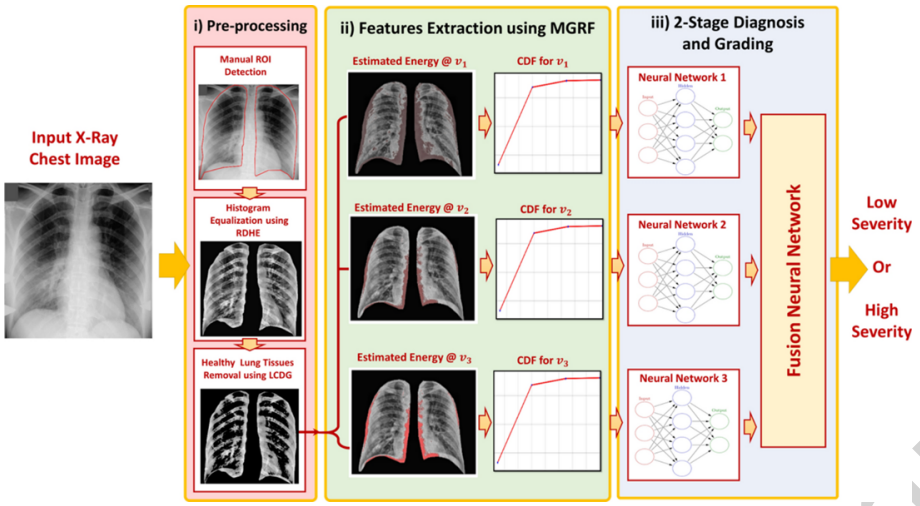


Figure 12: Pipeline of the CAD System of the MGRFNN model (adapted from Elsharkawy et al. [63] licensed under CC BY 4.0.)

modalities are strongly dissimilar in structure or quality. Such fusion approaches are useful in generating more situational and multi-faceted models that can be adapted easily to the Healthcare 5.0 setting.

The data types offer a complete picture of the patient, with imaging data recording the spatial and structural patterns [31], EHR giving the comprehensive clinical history, wearable devices providing real-time physiological measurements, and clinical notes offer contextual insights. DL models use comparatively heterogeneous sources to combine them, which can be broadly divided into feature-level and decision-level fusion. During feature-level fusion, the features extracted in separate modalities (i.e. CNN features on images and LSTM embeddings on EHR sequences, transformer embeddings from text) are inputted together into a common latent space where they are jointly learned. On the other hand, decision-level fusion takes the result of the unimodal models and combines them to obtain a final consensus prediction. The preprocessing operations (alignment, normalization and synchronization of data) are necessary to allow compatibility of modalities and to improve the overall predictive capabilities of the system.
MMDL enables the fusion of these heterogeneous sources to create richer, context-aware representations that drive improvements in diagnostics, prognostic modeling, and personalized treatment planning. For example, joint image-text models have been explored for tasks like visual question answering and automated clinical report generation, as surveyed in a scoping review of biomedical image–text integration [241]. In oncology, combining imaging modalities such as CT/MRI with structured clinical biomarkers has shown promise in improving liver cancer detection and prognosis [230]. More recently, comprehensive reviews highlight the rise of transformer-based fusion architectures (including intermediate fusion strategies), which consistently outperform unimodal baselines across tasks such as computer-assisted diagnosis and survival prediction [81], and emphasize improved AUC through multimodal integration in clinical applications [217]. These works underline how MMDL lies at the core of intelligent and adaptive healthcare systems, capable of enabling real-time decision support, automated reporting, and patient-centric care pathways.

In our study, we have reviewed some of the recent works in the MMDL domain. In healthcare and related fields, there is a growing demand for such models. Figure 13 represents different kinds of data that are frequently used in MMDL models. Hsu et al. [100] trained a shared space between image and text using unsupervised (and semi-supervised) learning, using a condition of cross-modal retrieval and representation learning in medical imaging. Hao et al. [89] performed chronic disease prediction using multimodal hospital data (Fine-grained text features and structured data). Lee et al. [137] forecasted the progression from Mild cognitive impairment to AD by integrating longitudinal and multimodal biomarkers. Li et al. [143] created a self-supervised feature learning algorithm utilizing multimodal retinal images that will be robust to defeating disease diagnosis without imposing manual labeling. Venugopalan et al. Venugopalan et al. [257] proposed a MMDL framework for precise staging of AD by integrating imaging data, genetic data (single-nucleotide polymorphisms (SNPs)), and clinical test records. Sheu et al. [226] developed a pneumonia status prediction system to assess whether

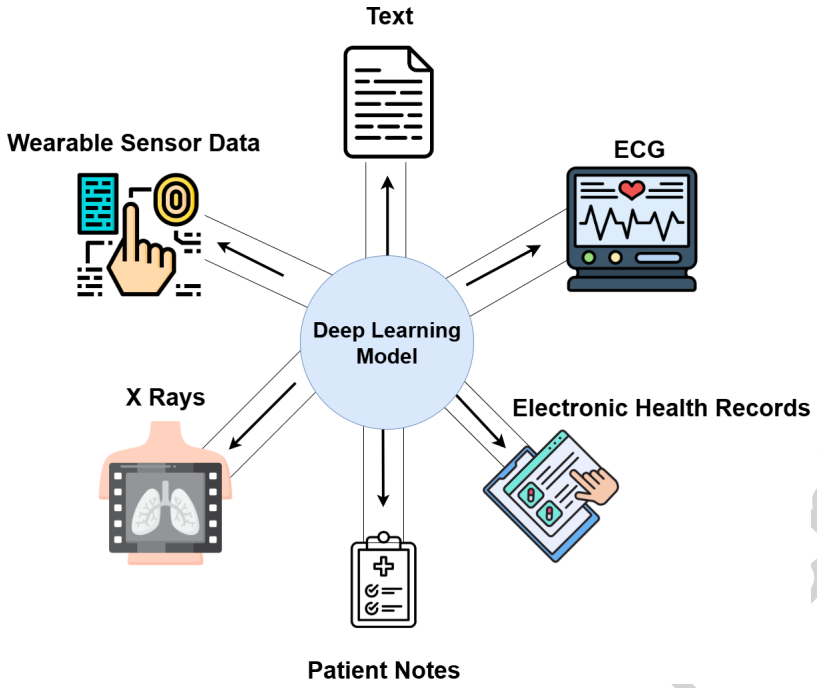


Figure 13: Commonly used modalities in MMDL

early-stage patients can be discharged within or after 7 days of hospitalization. Lee et al. estimated current Cardiovascular Disease (CVD) risk by integrating fundus images and clinical data; they evaluated model generalizability across datasets [138]. Zhang et al. [283] detected Parkinson's disease by using unsupervised, uncertainty-aware learning to extract informative representations from multivariate time series. Intriago et al. [106] proposed early identification of SCC individuals at elevated risk for developing Alzheimer's disease by multimodal MRI representation learning. Li et al. [141] enhanced disease prediction by independently optimizing feature alignment and fusion processes within an MMDL framework. Thiruvenkadam et al. [247] improved brain tumor detection using CNN on fused multi-modal MRI with XAI support. Mahalakshmi et al. [155] utilized multi-modal fusion to enhance diagnostic accuracy and facilitate personalized healthcare; breast cancer detection can be made early and precise. Krishnaveni et al. [130] achieved a more thorough classification of Chronic Obstructive Pulmonary Disease (COPD) stages through the integration of chest X-ray imaging and pulmonary function test data. Ding et al. [58] employed Large Language Multimodal Models (LLMMs) for the prediction of chronic diseases, specifically diabetes. Ranjith et al. [204] designed a multi-modal framework that facilitates accurate, interpretable, and scalable early detection of ocular diseases. Kumar et al. [134] developed a robust, scalable end-to-end system for early-stage Alzheimer's diagnosis, enabling timely intervention and improved patient outcomes. Gezimati et al. [73] established an MMDL framework for breast cancer diagnosis utilizing infrared and (theoretically) terahertz imaging. Shi et al. [227] enhanced glioma MRI segmentation via SAM fine-tuning with modality-specific fusion adapters. Tsai et al. [249] designed an effective multitask learning model for simultaneous prediction of multiple chronic diseases using multimodal patient information. Figure 14 shows how DL models deal with multiple modalities and how the data are processed, fused, and learned.

Table 10: Comparison of recent studies that use various DL techniques

Key literature	Publication year	Disease name	Risk prediction model	Accuracy (%)
1. Disease Prediction and Early Diagnosis				
[187]	2016	Alzheimer's disease	CNNs, VGG16, ResNet, and Inception	acc: 0.90 ± 0.09, sen: 0.86 ± 0.12, spec: 0.94 ± 0.10, AUC: 0.95
[49]	2016	Heart failure	RNN, GRU	AUCs ranged from 0.777 (12-month window) to 0.883 (18-month window)

Table continues on next page

Table continued from previous page

Key literature	Publication year	Disease name	Risk prediction model	Accuracy (%)
[3]	2017	Heart disease	Deep CNN	Original data: acc: 93.47%, ppv: 97.87%, sen: 96.01%, spec: 91.64% Noise free Data: acc: 94.03%, ppv: 97.86%, sen: 96.71%, spec: 91.54%
[116]	2018	Breast cancer	Deep CNN (CNNs layers, metric learning layers)	DDSM database: 97.4%, MIAS database: 96.7%
[18]	2018	AD	3D-CNN	AD vs. NC : 98.74% (detection rate: 100% and false alarm rate: 2.4%)
[67]	2019	AD	3D-CNN, FSBi-LSTM, softmax	AD vs. NC classification(%): acc: 94.82, sen: 97.7, spec:92.45, AUC: 96.76 pMCI vs. NC classification(%): acc: 86.36, sen: 83.33, spec:88.78, AUC: 91.11 MCI vs. NC classification(%): acc: 65.35, sen: 70.59, spec: 59.63, AUC: 69.17
[69]	2020	AD	3D-CNN-SVM	AD versus NC (%): acc: 99.10 ± 1.13, sen: 99.80 ± 0.37, spec: 98.40 ± 1.17 AD versus MCI (%): acc: 89.40 ± 6.90, sen: 86.70 ± 9.10, spec: 84.00 ± 4.80 MCI versus NC (%): acc: 98.90 ± 2.78, sen: 98.90 ± 3.69, spec: 98.80 ± 0.63 ternary classification: 95.74 ± 2.31%
[50]	2020	Brain tumor	CNN, Alexnet, Resnet50, InceptionV3, GoogleNet and Densenet201	acc: 97.01%, sen: 94.7%, spec: 100%, f-Measure: 96.9%
[6]	2020	COVID-19	CNN with VGG16 architecture	Australia region: prediction accuracies for confirmations, recoveries, and deaths 99.94%, 90.29%, and 94.18% Jordan region: obtained prediction accuracies 99.03%, 79.39%, and 86.82%
[176]	2022	Heart disease	Fuzzy Inference System (FIS), Bi-LSTM	acc: 98.86%, prec: 98.9%, sen: 98.8%, spec: 98.89%, f-measure: 98.86%
[232]	2022	CKD	DNN	acc:100% recall:100% prec:100% f-Measure:100%
[7]	2023	Retinoblastoma	Transfer Learning Inceptionv3 with LIME and SHAP	acc: 97%, prec: 98.8%, recall: 99.6%, f1-score: 99.2%
[276]	2024	CVD	BiGRU, IoT network	acc: 99.9%, f1-score: 99.53%
[161]	2024	CVD	MDensNet201-IDRSRNet, Relief, LASSO, IDRSNet	spec: 98.85%, sen: 98.90%, prec: 98.93%, recall: 99.01%, f1-score: 98.95%, acc: 99.12%
[158]	2024	AD	VGG16, VGG19, DenseNet169, DenseNet201, Ensemble 1 (VGG16, VGG19), Ensemble 2 (DenseNet169, DenseNet20), proposed model (EfficientNetB3, CNN), XAI method (saliency maps, Grad-CAM)	acc:96%, prec: 89%, recall: 93%, f1 score: 91%
[263]	2024	Lung cancer	CNN, XGBoost with SHAP	acc: 97.43%, sen:98.71%, and f1-score: 98.08%
[244]	2025	Type-2 diabetes	DeepNetX2	without feature selection: PIMA dataset: acc: 92.21%, prec: 94% , recall: 92% , f1-score: 92% Local Private dataset: acc: 95.74%, prec: 96% , recall: 96% , f1-score: 96% Type-2 diabetics dataset: acc: 99.50%, prec: 100% , recall: 99% , f1-score: 100% with feature selection: PIMA dataset: acc: 94.81%, precision: 95% , recall: 95% , f1-score: 95% Local Private dataset: acc: 97.87%, precision: 98% , recall: 98% , f1-score: 98% Type-2 diabetics dataset: acc: 97.50%, precision: 98% , recall: 97% , f1-score: 97%
[202]	2025	CKD	Hybrid CNN-SVM model	CKD: precision: 86%, recall: 100%, f1-score: 92% Non-CKD: precision: 100%, Recall: 75%, f1-score: 86% overall accuracy: 96%

Table continues on next page

Table continued from previous page

Key literature	Publication year	Disease name	Risk prediction model	Accuracy (%)
[223]	2025	Blood cancer	ResNetRS50, RegNetX016, AlexNet, Convnext, EfficientNet, InceptionV3, Xception, VGG19	ResNetRS50: accuracy: 97%, recall: 99%, f1 score: 98% RegNetX016: accuracy: 96.4%, recall: 99%, f1-score: 98% AlexNetacc: 87.9%, recall: 90%, f1-score: 88% ConvNext: accuracy: 94.2%, recall: 94.6%, f1-score: 95% EfficientNet: accuracy: 93.0%, recall: 88%, f1-score: 91% Inception_v3: accuracy: 92.7%, recall: 88%, f1-score: 91% Xception: accuracy: 91%, recall: 88%, f1-score: 92% VGG19: accuracy: 93.5%, recall: 93.6%, f1-score: 93.3%
[159]	2025	Dementia	MOD-1D-CNN and MOD-2D-CNN	MOD-1D-CNN: accuracy: 70%, precision: 68%, recall: 69%, f1-score: 68% MOD-2D-CNN: accuracy: 95%, precision: 93%, recall: 93%, f1-score: 93%
2. Medical Imaging Analysis and Radiology				
[101]	2015	Lung cancer	Deep Learning frameworks: DBN, CNN,	Sensitivity 1. CNN: 73.4% 2. DBN: 73.3%
[82]	2016	Diabetic retinopathy	Deep Learning algorithm	EyePACS-1: sensitivity: 97.5%, specificity: 98.1% Messidor-2: sensitivity: 96.1%, specificity: 98.5%
[47]	2016	Breast and lung cancer	Stacked Denoising Autoencoder (SDAE)	Lung CT: i. Single strategy: 87.4% ± 3.3 accuracy ii. All strategy: 94.4% ± 3.2 accuracy Breast US: 82.4% ± 4.5 accuracy
[287]	2017	Glaucoma	CNN	94.1%
[97]	2018	Esophageal cancer	Single shot multi-box detector CNN	Superficial cancer: 99% advanced cancer: 92% ESCC diagnosis: 99% EAC diagnosis: 90%
[147]	2018	Glaucomatous optic neuropathy	Deep CNN with Inception-v3 model	AUC of 0.986 sensitivity of 95.6%, specificity of 92.0%
[115]	2018	Oral cancer	Regression-based partitioned deep CNN	91.4% (small dataset), 94.5% (large dataset); sensitivity: 0.94. Specificity: 0.98
[235]	2019	Thyroid nodules (identifying benign and cancerous)	Deep Learning algorithm (DLA) using Inception-V3	Internal dataset: sensitivity: 95.2%, Negative Predictive Value (NPV): 95.5%. External dataset: Sensitivity: 94.0%, NPV: 90.3%.
[125]	2020	Skin cancer	Pretrained CNN models (VGG19, ResNet50, Inception V3, and SqueezeNet) and Transfer learning	99.60%
[190]	2020	COVID-19	VGG-16 CNN	Binary classification: Accuracy: 96%, sensitivity: 92.64%, specificity: 97.27% Multi-class classification: Accuracy: 92.53%, sensitivity: 86.7%, specificity: 95.1%.
[131]	2020	COVID-19	2D CNN with U-Net	F1-score: 97.31%
[63]	2021	COVID-19	MGRF model and a Neural Network-based fusion system for analysis	The CAD system achieved: sensitivity: 100% specificity: 97% ± 3% accuracy: 98% ± 2% DSC: 98% ± 2%
[44]	2022	Brain tumor	CNN with RMSProp optimizer and Softmax activation in final layer	99.74%
[178]	2023	Pulmonary diseases (Nodules, tuberculosis, pneumonia, COVID-19, and pulmonary edema)	ResNet50 CNN Model, LIME	COVID-CT dataset: 93% COVID Net dataset: 97%
[105]	2023	Tuberculosis	CBAMWDnet	Accuracy: 98.80% sensitivity: 94.28% specificity: 95.7% F1-score: 96.35% precision: 98.50%

Table continues on next page

Table continued from previous page

Key literature	Publication year	Disease name	Risk prediction model	Accuracy (%)
[222]	2023	Pneumonia	Neural Networks with VGG-16	1. Dataset 1: 92.15% accuracy 2. Dataset 2: 95.4% accuracy
[156]	2023	Brain tumor	CNN	Accuracy: 93.3% AUC: 98.43% Recall: 91.1% Loss: 0.25
[184]	2023	Alzheimer's disease	In-3-channel ResNet18	Accuracy: 73.90% specificity: 94.32% sensitivity: 66.74%
[99]	2024	Lung disease (COVID-19 and pneumonia)	Inception-V3, CROP, ROT, HF	Pneumonia dataset: 94.55% Covid19-&-Pneumonia dataset: 97.44%
[121]	2024	Intracranial hemorrhage (ICH)	CNN	Accuracy: 99.76% precision: 99.53% recall: 99.68%
[246]	2025	Lung cancer	CNN	Accuracy: 98.5% specificity: 88.7 % sensitivity: 89 % precision: 89.2% recall: 99.68%
3. Multimodal Deep Learning in Healthcare				
[150]	2015	AD	Stacked Autoencoder (SAE) with a softmax logistic regressor	Binary classification (MR-only): Accuracy :82.59%, sensitivity: 86.83% Multiclass classification (MR-only): Accuracy: 46.30%, specificity: 77.78% Multiclass classification (MR and PET): Accuracy: 53.79%, specificity: 86.98%
[100]	2018	Thoracic disease (Retrieval relevance according to ICD-9, this information does not apply at the training time however)	DenseNet-121:Image features, Deep Averaging Network (DAN) encoder: Textual, Embedding Alignment (EA), Adv(Adversarial alignment), EA and Adv (Semi-supervised), Principal Component Analysis (PCA): dimensionality reduction	Retrieval tasks: Text to image, image to text Measure: MRR, cosine similarity, nDCG@1/10/100 Best unsupervised bi-gram EA obtained nDCG@100
[89]	2019	Cerebral infarction	MD-RCNN (Multimodal Recurrent Convolutional Neural Network) : [Deep Belief Network (DBN), RCNN (Recurrent Convolutional Neural Network)]	Accuracy: 96%, Recall: 98.08%
[137]	2019	AD	Multimodal RNN (multi-GRU + logistic regression)	Accuracy: 81% Sensitivity: 84 % Specificity: 80 %
[143]	2020	Retinal disease diagnosis [Age-related macular degeneration (AMD), Pathologic Myopia (PM), Diabetic Retinopathy (DR)]	CycleGAN, ResNet18 Backbone	Unsupervised: 1. Ichallenge-AMD dataset: AUC: 74.58% , Accuracy: 86.58% , Precision: 83.20%, Recall : 74.58% , F1-score: 77.33% 2. Ichallenge-PM dataset: AUC: 98.55%, accuracy: 98.65% , precision: 98.60%, recall : 98.55% , F1-score: 98.57% Supervised: 1. Ichallenge-AMD dataset: AUC:77.19% 2. Ichallenge-PM dataset: AUC: 98.04
[257]	2021	AD, mild cognitive disorders (MCI), controls	3D-CNN, Auto-encoders	(MRI, EHR, and SNP)- internal cross-validation accuracy-79% and external test set accuracy-78%
[226]	2022	Pneumonia	Deep Learning Classification (MDA-PSP) system [Hybrid DNN with Batch Normalization (Dense-BN)]	Dense-BN accuracy: 0.77 (vitals), 0.92 (CXR), 0.75 (combined with weights)
[138]	2023	CVD	Multimodal Deep Neural Network (DNN + CNN: DenseNet-169)	AUROC: SMC: 0.781 (95% CI 0.766-0.798) UK Biobank: 0.872 (95% CI 0.857-0.886) for multimodal model

Table continues on next page

Table continued from previous page

Key literature	Publication year	Disease name	Risk prediction model	Accuracy (%)
[283]	2023	Parkinson's disease	TL-TAGCN: Time-contrastive learning with Topology Adaptive Graph Convolutional Network	1. PPMI data set: Accuracy: 94.6 % Precision: 0.963±0.011 Recall rate: 0.969±0.017 F1 score: 0.964±0.006 2. PS data set: Accuracy: 92.9 % Precision: 0.969±0.062 Recall rate: 0.945±0.045 F1 score: 0.955±0.027
[141]	2023	Colorectal cancer AD	DeAF: A two-stage Deep Learning framework: MA-SimSiam (unsupervised multimodal alignment) and SAF (Self-Attention Fusion) module	1. Colorectal cancer: Accuracy: 55.15 % 2. ADNI: Accuracy: 69.9 %
[106]	2023	AD	Self-Supervised Multimodal Representation Learning (SSMRL) framework: Balanced Random Forest (BRF) + CyCLIP cost function + 3D CNN (sMRI) + BiLSTM (fMRI)	AUC scores: 1.sMRI: 98.30 % ± 1.53 % (multi) 2. fMRI: 97.70 % ± 1.47 % (multi)
[247]	2024	Brain tumor	Five-layer CNN incorporating dropout and max pooling, assessed using XAI techniques	98.9% validation accuracy after 40 training epochs
[155]	2024	Breast cancer	CNN, RNN	Metrics shown as 1.0; exact values unspecified
[130]	2024	Chronic obstructive pulmonary disease (COPD)	MultimodNet: FCN (for PFTs) + DenseNet121 (for CXR)	Accuracy: Initial stage: 95.231% Progressive stage: 95.008% Complicated stage: 94.989% Critical stage: 94.256%
[58]	2024	Type 2 Diabetes Mellitus (T2DM)	LLMs (LLMs) including BiomedBERT, ClinicalBERT, and GPT-2 DNNs, and SHAP	Multimodal : (A) textual laboratory values, (B) clinical notes, and (C) laboratory values BiomedBERT-93, ClinicalBERT-90, SciFive-93, RoBERTa-92, Flan-T5-base-220M-93, Flan-T5-large-770M-92, BERT-93, GPT-2-92
[204]	2025	Cataracts and glaucoma	Hybrid MMDL framework: ResNet/EfficientNet, attention mechanisms, CNN, RCNN, GNN and explainability tools: SHAP and Grad-CAM	Accuracy: 95 % (cataract), 92 % (glaucoma) AUC-ROC: 0.97 (cataract), 0.94 (glaucoma)
[134]	2025	AD	DenseNet (primary); CNN, VGG-16, MobileNet (comparative baseline models), Adam optimizer	DenseNet accuracy: 81.5% MobileNet accuracy: 57.4% CNN accuracy: 50.8% VGG-16 accuracy: 26.7 %
[73]	2025	Breast cancer	Custom 23-layer CNN; decision-level (late) fusion architecture	Validation Accuracy:96.6% Sensitivity: 96.3% AUC: 0.99
[227]	2025	Glioma (brain tumor)	SAM [ViT-based image encoder, mask decoder, and prompt encoder]	FAHZU private dataset: Dice score: 88.81% and HD95: 9.83 BraTS2020 public dataset: Dice score 89.25% and HD95: 10.03
[249]	2025	Diabetes mellitus, heart disease, stroke, hypertension	Multimodal Multitask Learning (MTL) [Multimodal Attention Network for Dementia (MAND) with logistic regression (LR), multilayer perceptron (MLP), LSTM, and Multi-Head Self-Attention (MHSA), CTR models: Factorization Machines (FM) and Deep Convolutional Network (DCN)]	Best AUC: 0.9346 (MAND-LSTM for stroke)

acc = accuracy, ppv = positive predictive value, prec = precision, sen = sensitivity, spec = specificity, AD = Alzheimer's disease NC = Normal control, pMCI = progressive state, MCI = Mild cognitive impairment.

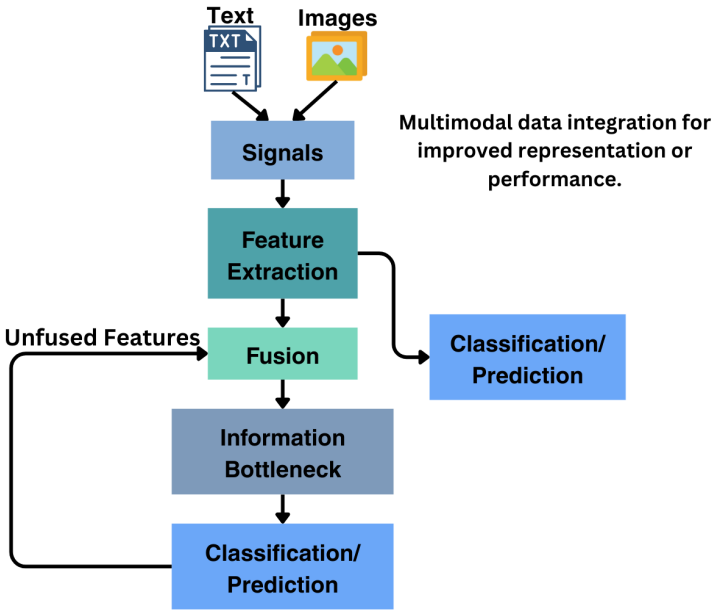


Figure 14: Deep learning architecture for multimodal fusion.

Table 11: Dataset description

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
1. Disease Prediction and Early Diagnosis						
[187]	ISIC archive dataset (International Skin Imaging Collaboration)	Contains dermoscopic images of skin lesions	900 dermoscopic images	Images are labeled as benign or malignant skin lesions. The dataset is specifically curated for skin cancer classification.	https://challenge.isic-archive.com/landing/2016/41/?utm_source=chatgpt.com	Python, TensorFlow, Keras, OpenCV, and NumPy
[49]	Sutter Palo Alto Medical Foundation (Sutter-PAMF) primary care patients	EHRs	32,787 patient records (3,884 HF cases and 28,903 controls)	Time-stamped clinical data such as: Diagnoses, medication orders, procedure orders	-	Theano, Skip-gram
[3]	PhysioBank MIT-BIH arrhythmia database	Imbalanced ECG data	104.3 MB	-	https://physionet.org/content/mitdb/1.0.0/	MATLAB (Matlab, Natick, MA, USA) software
[116]	DDSM and MIAS database	Mammography (X-ray imaging)	SM database- 600 images MIAS database- 1.51 GB	-	DDSM database- http://www.eng.usf.edu/cvprg/mammography/database.html MIAS database https://www.repository.cam.ac.uk/handle/1810/250394	Deep Learning toolbox, MatConvNet and CUDA
[18]	ADNI	Brain MRI scans	-	subjects- 340 MRI brain scans-1198 (AD- 600, NC-598)	https://adni.loni.usc.edu/data-samples/adni-data/	Python, Keras Deep Learning library built on TensorFlow backend
[67]	ADNI	MRI, PET	-	AD-93, pMCI-76, sMCI-128, NC-100	https://adni.loni.usc.edu/	Keras library, TensorFlow as backend

Table continues on next page

Table continued from previous page

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
[69]	ADNI	Brain MRI images	-	ubjects- 469 MRI samples-3127	https://adni.loni.usc.edu/	Python version 2.7.12, Keras , TensorFlow
[50]	Kaggle Repository	Brain MRI images	-	without tumor-98 with tumor-155	https://www.kaggle.com/datasets	Matlab
[6]	Kaggle	Chest X-ray images	Original dataset: 97 MB 128 images Augmented dataset: 1000 images	Two types of data: A healthy dataset (chest X-ray images of healthy persons) and A COVID-19 dataset (chest X-ray images of COVID-19 patients)	https://www.kaggle.com/datasets/nabeelsajid917/covid-19-x-ray-10000-images	Python, ARIMA, Fbprophet, ImageDataGenerator, Keras, LSTM, Matplotlib, NumPy, Pandas, Scikit, SciPy, TensorFlow
[176]	Cleveland and Hungarian dataset UCI machine learning repository	Physiological data, electronic clinical data (ECD)	658 Kb	-	https://archive.ics.uci.edu/dataset/45/heart+disease	Wireless body sensor networks, TensorFlow ML package, Apache Spark, and Cassandra
[232]	UCI Machine Learning repository	Clinical data (patient health records)	400 patients	24 clinical features. 250 instances of CKD and 150 instances of non-CKD	https://archive.ics.uci.edu/	Python Scikit-learn
[7]	Retinoblastoma fundus images-MathWorks Retinoblastoma dataset, Google images non-retinoblastoma fundus images-Messidor dataset	Fundus images	-	800 fundus images (400 retinoblastoma and 400 normal images)	-	Python with Keras, Lime, and SHAP libraries for model training and explainable AI methods, and Google Colab with a Tesla T4 GPU for computation
[276]	1. Framingham's heart dataset: US National Institutes of Health's (NIH) Biospecimen Information Coordination Center and Data Warehouse 2. Statlog heart dataset: Kaggle	Epidemiological data, physiological data	1. Framingham's heart dataset: 2. Statlog heart dataset: 141 Kb	4240 records, 15 attributes	1. Statlog heart dataset: https://archive.ics.uci.edu/dataset/98/statlog+project	biosensors, Cloud Simulator (CloudSim)
[161]	Hungarian, Cleveland, Long Beach, VA, Switzerland	Statlog dataset: UCI	1. Hungarian, Cleveland, Long Beach, VA, Switzerland dataset:125.9 KB 2.Statlog dataset: 4.3 KB	303 records, 14 attributes	Hungarian, Cleveland, Long Beach, VA, Switzerland dataset: http://www.eng.usf.edu/cvprg/mammography/database.html Statlog dataset: https://www.repository.cam.ac.uk/handle/1810/250394	Python
[158]	Kaggle Oasis	MRI scans	36 MB	Mild dementia-896 images moderate dementia-64 images Non-dementia- 3200 images Very mild dementia-2240 images	-	Python, TensorFlow, Keras, OpenCV, Scikit-learn, Google Colab

Table continues on next page

Table continued from previous page

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
[263]	Survey Lung Cancer (SLC)	Numerical data and categorical data	-	309 samples with 16 features	-	TensorFlow, Keras
[244]	1. Type-2 diabetic dataset (IEEEDataPort) from Frankfurt hospital, Germany 2. Local private dataset from the Pabna diabetes hospital 3. PIMA Indian diabetes dataset	Numerical data and categorical data	1.Type-2 Diabetes dataset: 60.60 KB 2. Local private dataset 3. PIMA Indian diabetes dataset: 45.8 KB	1. Type-2 diabetes dataset: 2000 individuals (684 diabetic, 1316 non-diabetic) 2. Local private dataset: 465 participants (373 diabetic, 92 non-diabetic) 3. PIMA Indian diabetes dataset: 768 records (268 diabetic, 500 non-diabetic)	1. Type-2 diabetes dataset: https://iee-dataport.org/documents/type-2-diabetes-dataset 2. Local private dataset: - 3. PIMA Indian diabetes dataset: https://data.mendeley.com/datasets/7zcc8v6hvp/1	Python, TensorFlow, Keras, and Scikit-learn, SHAP, LIME for XAI techniques, assembled on Kaggle's free tier
[202]	Clinical database	Numerical data and categorical data	-	8 essential clinical features including age, blood pressure, specific gravity, albumin, sugar levels, red and white blood cell counts, hemoglobin, blood urea, serum creatinine, and other relevant factors	-	-
[223]	Chinese National Medical centre (C-NMC) dataset	Medical images of blood cancer (leukemia, lymphoma, and multiple myeloma)	15,135 images total, 10.44 GB	Open-access dataset of blood cancer images, Labeled images showing specific forms of blood cancer	https://www.cancerimagingarchive.net/collection/c-nmc-2019/	Google Colab, Python, TensorFlow
[159]	1. Health metric data -Apollo diagnostic center and hospitals. 2. Facial image dataset collected by the research team	-	Health metric data: 1000 samples. Facial image dataset: 1800 images; 900 demented, 900 non-demented	Quantitative health metrics (diabetic status, age, blood oxygen level, heart rate, body temperature, weight) and Facial images	-	Python programming within Jupyter notebook environment, Javascript, HTML, and CSS
2. Medical Imaging Analysis and Radiology						
[101]	Lung Image Database Consortium (LIDC) dataset	CT images	Data of 1010 patients	Nodules with size larger than 3mm selected. Total 2545 selected	1. https://pmc.ncbi.nlm.nih.gov/articles/PMC3041807/ 2. https://pubmed.ncbi.nlm.nih.gov/15333795/	-
[82]	EyePACS-1, Messidor-2 datasets	Retinal fundus images	1. EyePACS-1:9963 images from 4997 patients 2. Messidor-2: 1748 images from 847 patients	EyePACS and 3 hospitals in India for development and Messidor-2 for validation	Messidor: https://www.ias-iss.org/ojs/IAS/article/view/1155	StatsModels, SciPy, Python
[47]	Breast US from Taipei Veterans general Hospital, LIDC: Lung image database consortium	Breast ultrasound and CT scans	1. 520 breast US images from 520 patients 2. 1400 nodules from LIDC	1. Breast US dataset: 520 images (245 malignant, 275 benign) 2. Lung CT: 1400 (700 benign, 700 malignant)	1. https://pmc.ncbi.nlm.nih.gov/articles/PMC3041807/ 2. https://pubmed.ncbi.nlm.nih.gov/15333795/	
[287]	DRISHTI-GS dataset and RIM-ONE v3 dataset	Retinal fundus images	DRISTI GS- 1,016 MB	DRISHTI-GS dataset includes 50 images, RIM-ONE v3 has 159 stereo fundus images.	DRISTI-GS : https://cvit.iiit.ac.in/projects/mip/drishti-gs/mip-dataset2/Home.php RIM-ONE v3: http://medimrg.webs.ull.es/research/retinal-imaging/rim-one/	MATLAB

Table continues on next page

Table continued from previous page

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
[97]	Cancer Institute Hospital, Japan	Endoscopic images with white-light imaging (WLI) and narrow-band imaging (NBI) at high resolution	Training: 8428 images; testing: 735 images,	Training: 8428 images of esophageal cancer lesions (397 ESCC and 32 EAC) Test: 1118 images (162 cancerous and 376 non-cancerous)	-	Framework: Caffe Deep Learning framework algorithm: Single shot multiBox detector (SSD).
[147]	LabelMe (Healgoo Ltd. LabelMe dataset),	Retinal fundus photographs		48 116 fundus photographs	-	Stata software version 14.0 (for statistical analyses)
[115]	BioGPS data portal: UCI Machine Learning repository, TCIA archive: Standard GDC data set: Standard	Hyperspectral imaging data encompassing malignant, benign, and normal tissue areas	1300 image patches	BioGPS: 100 image patches TCIA: 500 image patches GDC: 700 image patches	TCIA Archive: https://www.cancerimagingarchive.net/ GDC data set: https://portal.gdc.cancer.gov/	GoogLeNet Inception V3 architecture, executed on hardware with an NVIDIA GeForce GPU
[235]	Internal dataset: Seoul Metropolitan Government Seoul National University (SMG-SNU) Boramae Medical Center, Korea External dataset: Kuma Hospital, Japan	Ultrasonographic (US) thyroid nodule images in TIFF or DICOM format	1513 images	Training data: 1358 US images (670 benign, 688 malignant) Internal test set: 55 US images (34 benign, 21 malignant) External Test Set: 100 US images (50 benign, 50 malignant)	https://cdn-links.lww.com/permalink/md/c/md_2019_04_02_chai_md-d-18-06935_sdc1.pdf	Python, NumPy
[125]	International Skin Imaging Collaboration (ISIC) image archive	Skin lesion images	Augmented dataset: 5,000 images (3,800 training, 1,200 testing)	Benign: 1,900 (training), 600 (testing) Malignant: 1,900 (training), 600 (testing)	https://www.isic-archive.com/	Python, MongoDB, TensorFlow, .NET framework (C#)
[190]	1. Cohen's COVID-19 image data collection, 2. Kaggle competition data set of chest x-rays	Chest X-ray images		A total of 1,428 X-ray images:224 positive for COVID-19, 700 instances of bacterial pneumonia, 504 healthy cases	1. https://github.com/ieee8023/covid-chestxray-dataset 2. https://www.kaggle.com/datasets/andrewmvd/convid19-x-rays	
[131]	1. Kaggle dataset 2. Github dataset	CT scan images	1. Kaggle dataset: 1 GB 2. Github dataset: 309.3 MB	1. Kaggle dataset: 20 complete chest CT scans (301 images each) 2. Github dataset: 929 CT slices from over 50 patients	1. https://www.kaggle.com/datasets/andrewmvd/covid19-ct-scans 2. http://medicalsegmentation.com/covid19/	MedSeg annotation tool, python, TensorFlow
[63]	1. Publicly available archives of COVID-19 positive cases 2. COVID-19 Open Research Dataset Challenge (CORD-19) 3. Data from the University of Louisville and Mansoura University	Chest X-ray images	1. Publicly available archives of COVID-19 positive cases: 525 MB 2. COVID-19 Open Research Dataset Challenge (CORD-19): 20 GB	X-rays of 200 COVID-19-positive individuals, 100 of whom died and 100 recovered	1. https://github.com/ieee8023/covid-chestxray-dataset	
[44]	2020 BraTS dataset	2D MRI images	7 GB	2892 images (2473 training and 273 testing, 9:1)	https://www.med.upenn.edu/cbica/brats2020/data.html	Python, TensorFlow, Keras

Table continues on next page

Table continued from previous page

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
[178]	COVID-CT, COVID-NET	Computed Tomography (CT) scans and chest radiographs (X-rays)		COVID-CT dataset: 800 images, 349 CT scans with COVID-positive results (216 persons), and 397 CT scans with COVID-negative (397 persons). COVID-NET Dataset: 19,000 chest radiographs (CXRs)	COVID-CT: https://arxiv.org/abs/2003.13865	Python, Google Colaboratory (COLAB), TensorFlow, LIME
[105]	1. Montgomery dataset 2. Shenzhen dataset 3. Tuberculosis (TB) chest X-ray dataset	Chest X-Ray images	5000 CXR images	3906 in normal category, 1094 in TB category from 3 datasets	Datasets 1 and 2. https://pmc.ncbi.nlm.nih.gov/articles/PMC4256233/ Dataset 3. https://arxiv.org/abs/2007.14895	
[222]	1. Dataset 1: Kaggle 2. Dataset 2: Kaggle and GitHub	CXR radiographs	1. 5856 images from dataset 1 2. 6436 images from dataset 2	1. Pediatric CXR, having normal vs. pneumonia classes 2. Multi-class images having normal, pneumonia and COVID-19 images	1. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia	Orange 3.31.1 simulator used.
[156]	Brain tumor dataset: Kaggle	MRI images	3264 images	The dataset contains images in 4 classes: 937 images of Meningioma, 500 of no tumor, 900 of pituitary tumor, and 926 of glioma tumor	https://www.kaggle.com/code/sayedgomaa/brain-tumor/notebook	Google Colab Pro+
[184]	ADNI	MRI, PET images	824 images (412 MRI, 412 PET)	The dataset contains spatially normalized MRI and PET images, which consist of early and late MCI (Mild Cognitive Impairment) groups	https://adni.loni.usc.edu/	Pytorch, Nvidia GTX1660
[99]	Pneumonia dataset: Kaggle Covid19-&-Pneumonia dataset Kaggle	Chest X-Ray image data	1. Pneumonia dataset: 2.31 GB 2. Covid19-&-Pneumonia dataset: 2.14 GB	Pneumonia dataset: 5856 images, a training set (5232 images) and a testing set (624 images). Covid-19and pneumonia dataset: 6432 images in a training set (5144 images) and a testing set (1288 images)	1. https://www.kaggle.com/datasets/paultimothymooney/chest-xray-pneumonia?resource=download	Python, PyTorch, vit_pytorch API, pytorch-grad-cam
[121]	Kaggle	-	-	82 patients, around 30 image segments each	-	Python, TensorFlow, Keras, OpenCV, NumPy
[246]	Kaggle	Chest CT images	613 images	It has normal and affected groups of images. The affected group has squamous, large cell, and adenocarcinoma categories	https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images	Python, TensorFlow, ResNet-50, Modified AlexNet, VGGNet, Keras, Pandas, Matplotlib, NumPy, Seaborn, Sklearn
3. Multimodal Deep Learning in Healthcare						
[150]	Alzheimer's Disease Neuroimaging Initiative (ADNI)	T1-weighted MRI, FDG-PET scans	758 MR-only patients (180 AD, 204 NC, 214 ncMC, 160 cMCI) multimodal patients (MRI and PET): 331	Total brain ROIs utilized 83, PET aligns with MRIs	https://adni.loni.usc.edu/	MATLAB 2013a, Image Registration Toolkit (IRTK)
[100]	Medical Information Mart for Intensive Care (MIMIC) dataset	X-ray images, Radiology reports (Text: Findings and impressions)	473,057 chest X-rays images and 206,563 reports of 63,478 patients	473,057 chest X-rays images and 206,563 reports of 63,478 patients	-	-

Table continues on next page

Table continued from previous page

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
[89]	Grade-A, second-class hospital in Wuhan, China	-	20,320,848 data from 31,919 patients (2013–2015)	multimodal clinical data from 706 patients, 606 for training and 100 for testing	-	Python framework
[137]	ADNI	Neuroimaging data, cerebrospinal fluid biomarkers, cognitive performance metrics, and participant demographic characteristics	1,618 ADNI participants: 307 MCI-converter, 558 MCI non-converter, 415 cognitively normal older adult controls (CN), 338 AD	Longitudinal and cross-sectional data	http://adni.loni.usc.edu	TensorFlow, PyTorch
[143]	1. Ichallenge-AMD dataset 2. Ichallenge-PM dataset 3. EyePACS dataset: Kaggle's Diabetic Retinopathy Detection Challenge (EyePACS) dataset 4. Fundus-FFA	Color fundus images and Fundus Fluorescein Angiography (FFA) images	1. Ichallenge-AMD dataset: 1200 images 2. Ichallenge-PM dataset: 1200 images 3. EyePACS dataset: 88.29 GB (88,702 images 4. Fundus-FFA dataset: 59 subjects (30 healthy, 29 DR patients)	1. Ichallenge-AMD dataset: AMD patients (23%) and non-AMD patients (77%) 2. Ichallenge-PM dataset: PM cases (50%), Non-PM (50%), and PM cases (50%) 3. EyePACS dataset: DR grading ranges from 0–4	1. EyePACS dataset: https://www.kaggle.com/c/diabetic-retinopathy-detection/data	PyTorch
[257]	ADNI (ADNI1, ADNI2, ADNI GO, and ADNI3	Clinical data, imaging (MRI, PET), Genetic	Genetic data-3 million	SNP-808 patient, MRI imaging-503 patients, clinical and neurological test data -2004 patients	https://adni.loni.usc.edu/	-
[226]	Taichung Veterans General Hospital (TCVGH), Taiwan	Vital signs, Chest X-ray images	3972 patient records for pneumonia	Adults (> 18); days 1 to 3; CXR and vitals (2014 to 2018) ICU, short stays, deaths excluded	-	Python (Keras, OpenCV, Pandas, Matplotlib, Seaborn, NumPy)
[138]	1. Samsung Medical Center (SMC) dataset, Korea 2. UK Biobank dataset	-	SMC: 517 patients (2954 images) for internal validation; 2026 patients (3518 images) for development UK Biobank: 11,091 patients (11,298 images) for external validation	Retrospective EMR and fundus data (2010–2016); CVD defined by ICD-10; externally validated with UK Biobank	UK Biobank: https://www.ukbiobank.ac.uk/use-our-data/apply-for-access/	TensorFlow (version 2.9.0), R (version 3.6.3)
[283]	The Parkinson's Progression Markers Initiative (PPMI) dataset Parkinson Speech dataset (PS): Department of Neurology in Cerrahpasa Faculty of Medicine, Istanbul University	Multivariate time series (from sensor data, voice recordings, clinical attributes)	1. PPMI dataset: 683 subjects, 15,798 records, 212 features 2. PS dataset: 68 subjects, 1208 records, 26 features	1. PPMI dataset: Clinical assessments 2. PS dataset: Acoustic voice features (e.g., jitter, shimmer, pitch)	PPMI dataset: http://www.ppmi-info.org	TAGCN architecture using a PyTorch-based framework
[141]	1. Colorectal cancer dataset (Sun Yat-sen University, Fudan University, Cancer Hospital, Zhujiang Hospital of Southern Medical University, Fujian Cancer Hospital, and Fujian Medical University Cancer) ADNI dataset	Multimodal (medical images and clinical data)	1. Colorectal cancer dataset: 185 patients 2. ADNI dataset: 1675 samples	1. Colorectal cancer dataset: Includes CT images and clinical data 2. ADNI dataset: MRI images, volume of anatomical brain structures (138-dimension), phenotypic information (2-dimension), and genetic data	ADNI dataset: http://adni.loni.usc.edu	PyTorch

Table continues on next page

Table continued from previous page

Key literature work	Database used	Data types	Database size	Details of the dataset	Link	Software/ tools used
[106]	Dataset from different pathologies	Brain imaging data (labeled data and unlabeled data)	1. Labeled dataset: 487 participants (31 AD, 456 NC) 2. Unlabeled dataset: 468 participants (177 SCC, others with psychiatric conditions)	Multimodal MRI (Structural MRI (sMRI), Functional MRI (fMRI)) data from individuals aged 18–98, encompassing SCC, AD, and psychiatric groups	-	FastSurfer, fMRIprep
[247]	BraTS (Brain Tumor Segmentation Challenge dataset)	Multimodal MRI brain images (T1c, T2, FLAIR)	3300 images	Labeled MRI slices fused at the pixel level and categorized into tumor and non-tumor classes. 10 volumes of each MRI multimodality image	-	Google Colaboratory
[155]	-	Mammograms (image), prescription data, blood reports (textual/sequential)	-	Contains benign/malignant data; augmented mammograms; sequential clinical records	-	TensorFlow, Keras
[130]	1. Chest X-ray14 dataset 2. COPDGene dataset	-	Chest X-ray14 dataset: Over 100,000 images	CXR resized to 224×224; PFTs normalized; data aligned and balanced using augmentation and SMOTE	1. Chest X-ray14 dataset: `https://nihcc.app.box.com/v/ChestXray-NIHCC`	TensorFlow, Keras

4.4 Comparative analysis of DL models in healthcare tasks

Over the past decade, the sphere of deep learning (DL) models has achieved profound improvements in the healthcare sector with regard to predicting diseases, especially as a part of Healthcare 5.0. Comparisons of various diseases, including AD, kidney disease, lung disease, heart disease and breast cancer, indicate that DL models like CNN, LSTM, hybrid and ensemble models are always more effective than the traditional machine learning models in medical imaging and clinical data. A brief comparison of model based on the prominent studies is provided in the Tables 12, 13, 14, 15, 16 below and the main distinctions in terms of Model/Method, accuracy, disease domain are pointed out.

5 Challenges and Limitations

Despite the remarkable progress achieved in deep learning–based healthcare systems, several methodological, ethical, and deployment-related challenges remain that must be addressed for reliable real-world implementation. DL methods have changed the scenario of many healthcare sectors, dealing with high percentages of improvement in medical image analysis, disease prediction, and remote patient monitoring. However, use of such techniques and implementation in the clinical setting has several obstacles and hindrances [220, 13, 122]. These problems include issues relating to the quality of data and the capacity to interpret results, computationally intensive demands, and the necessity of large amounts of data [181]. It is important to recognize these problems for effective use of DL in healthcare environments.

The literature review presented in Table 17 covers all three areas: disease prediction and early diagnosis, medical imaging and radiology, and MMDL, summarizing the advantages, disadvantages, and research gaps of existing Deep Learning approaches. This section outlines the key challenges and unresolved issues identified across the reviewed studies.

5.1 Challenges and Limitations for Disease Prediction and Early Diagnosis

Early detection of illnesses and anticipating illnesses are critical elements of health care, emphasizing the improvement of patient outcomes by specifying the illnesses during the first stages [74]. However, several challenges and limitations hinder the efficacy of the predictive models.

Table 12: Summary of comparative table for neurodegenerative disease diagnosis (alzheimer's disease, dementia, and parkinson disease)

Key literature work	Model	Accuracy (%)
[187]	CNNs, VGG16, ResNet, and Inception	90%
[18]	3D-CNN	98.74%
[67]	3D-CNN, FSBi-LSTM, softmax	94.82%
[69]	3D-CNN-SVM	99.10%
[158]	VGG16, VGG19, DenseNet169, DenseNet201, Ensemble 1 (VGG16, VGG19), Ensemble 2 (DenseNet169, DenseNet20), proposed model (Efficient-NetB3, CNN), XAI method (saliency maps, Grad-CAM)	96%
[184]	In-3-channel ResNet18	Accuracy: 73.90%, specificity: 94.32%, sensitivity: 66.74%
[150]	SAE with a soft- max logistic regressor	Binary classification (MR-only): Accuracy :82.59%, sensitivity:86.83% Multiclass classification (MR-only): Accuracy: 46.30%, specificity: 77.78% Multiclass classification (MR and PET): Accuracy: 53.79%, specificity: 86.98%
[184]	Multimodal RNN (multi-GRU + logistic regression)	Accuracy: 81% Sensitivity: 84 % Specificity: 80 %
[106]	Self-Supervised Multimodal Representation Learning (SSMRL) framework: Balanced Random Forest (BRF) + CyCLIP cost function + 3D CNN (sMRI) + BiLSTM (fMRI)	AUC scores: 1.sMRI: 98.30 % ± 1.53 % (multi) 2. fMRI: 97.70 % ± 1.47 % (multi)
[134]	DenseNet (primary); CNN, VGG-16, MobileNet (comparative baseline models), Adam optimizer	DenseNet accuracy: 81.5% MobileNet accuracy: 57.4% CNN accuracy: 50.8% VGG-16 accuracy: 26.7 %
[113]	VGG-16 CNN + SVM/KNN/Tree	99.9%(fMRI)
[159]	MOD-CNN [MOD-1D-CNN and MOD-2D-CNN]	MOD-1D-CNN : 70.50%, MOD-2D-CNN : 95.72%
[257]	3D-CNN, Auto-encoders	(MRI, EHR, and SNP)- internal cross-validation accuracy-79% and external test set accuracy-78%
[283]	TL-TAGCN	1. PPMI data set: Accuracy: 94.6 % Precision: 0.963±0.011 Recall rate: 0.969±0.017 F1 score: 0.964±0.006 2. PS data set: Accuracy: 92.9 % Precision: 0.969±0.062 Recall rate: 0.945±0.045 F1 score: 0.955±0.027

Table 13: Summary of comparative table for heart disease diagnosis

Key literature work	Model	Accuracy (%)
[49]	RNN, GRU	UCs ranged from 0.777 (12- month window) to 0.883 (18-month window)
[3]	Deep CNN	Noise free Data: 94.03%, Origi- nal data: 93.47%
[176]	Fuzzy Inference System (FIS), Bi-LSTM	98.86%
[276]	Bi-GRU, IoT network	99.90%
[161]	MDensNet201-IDRSRNet, Relief, LASSO, IDRSNet	99.12%
[138]	DNN + CNN: DenseNet-169	AUROC: SMC: 0.781 (95% CI 0.766-0.798) UK Biobank: 0.872 (95% CI 0.857-0.886) for multimodal model
[205]	AttGRU-Hybrid Deep Learning	95.42%
[33]	Hybrid ETCXGB (CNN-XGBoost)	99.26%
[17]	Pure DL (MLP, CNN, RNN)	84.3% (AUC)

5.1.1 Challenges in Data Handling

- *Preprocessing of data:* Real-world healthcare data often contains some inconsistency, incompleteness, and noise, which make the DLA very challenging to implement. Therefore, feature extraction needs additional processing that can result in enhanced diagnosis skills. This indicates that there has been an inadequate and ineffective way of learning the model efficacy in detecting the disease at its early stages [49, 116, 176].

Table 14: Summary of comparative table for lung disease (pneumonia, COVID-19, cancer) diagnosis

Key literature work	Model	Accuracy (%)
[263]	CNN, XGBoost with SHAP	97.43%
[101]	DBN, CNN	Sensitivity 1. CNN: 73.4% 2. DBN: 73.3%
[6]	CNN with VGG16 architecture	Australia region: 94.80% Jordan region: 88.43%
[47]	Stacked Denoising Autoencoder (SDAE)	Lung CT: i. Single strategy: 87.4% ± 3.3 accuracy ii. All strategy: 94.4% ± 3.2 accuracy Breast US: 82.4% ± 4.5 accuracy
[190]	VGG-16 CNN	Binary classification: Accuracy: 96%, sensitivity: 92.64%, specificity: 97.27% Multi-class classification: Accuracy: 92.53%, sensitivity: 86.7%, specificity: 95.1%.
[131]	2D CNN with U-Net	F1-score: 97.31%
[63]	MGRF model and a Neural Network-based fusion system	sensitivity: 100% specificity: 97% ± 3% accuracy: 98% ± 2% Dice Similarity Coefficient (DSC): 98% ± 2%
[178]	ResNet50 CNN Model, LIME	COVID-CT dataset: 93% COVID Net dataset: 97%
[222]	Neural Networks with VGG-16	1. Dataset 1: 92.15% accuracy 2. Dataset 2: 95.4% accuracy
[99]	Inception-V3, CROP, ROT, HF	Pneumonia dataset: 94.55% Covid19-&-Pneumonia dataset: 97.44%
[246]	CNN	Accuracy: 98.5% specificity: 88.7 % sensitivity: 89 % precision: 89.2% recall: 99.68%
[226]	Dense-BN	Dense-BN accuracy: 77%(vitals), 92%(CXR), 75% (combined with weights)
[130]	MultimodNet: FCN + DenseNet121	Accuracy: Initial stage: 95.231% Progressive stage: 95.008% Complicated stage: 94.989% Critical stage: 94.256%

Table 15: Summary of comparative table for kidney disease diagnosis

Key literature work	Model	Accuracy (%)
[232]	DNN, SVM, KNN, Logistic Regression, Random Forest, Naive Bayes	100%
[202]	Hybrid CNN-SVM model	96.8%

Table 16: Summary of comparative table for breast cancer diagnosis

Key literature work	Model	Accuracy (%)
[116]	Deep CNN (CNNs layers, metric learning layers)	DDSM database: 97.4%, MIAS database: 96.7%
[155]	CNN, RNN	Metrics shown as 1.0; exact values unspecified
[73]	CNN,decision-level (late) fusion architecture	Validation Accuracy:96.6% Sensitivity: 96.3% AUC: 0.99
[5]	YOLO detector with CNN, ResNet-50	DDSM dataset: YOLO 99.17% INbreast dataset: 97.27%

- *Data quality and data complexity:* Various studies point to the problems of data quality and complexity, especially when it comes to EHRs. The types of data covered include but are not limited to numerical, categorical, and textual information; they often have missing values and noise, making it challenging to extract the data and impeding the training of DL models [187, 3, 49, 161, 244]. In most instances, the reliance on health variables entered by the user can result in the inaccuracy of the data, which harms the level of diagnostic accuracy [159].
- *Data availability:* Availability of large, high-quality datasets is a big challenge since DL models need vast

amounts of data to learn correctly. In the case of healthcare, the condition is highly problematic, as the considerations of data privacy limit access to the full dataset [3, 116, 69, 50, 6, 276, 159].

- *Data diversity and data size:* The data diversity and data size are also vital aspects in the realization of the high performance of the model. The shortage of a range of datasets that possess a wide variety of demographics and disease progression levels is a factor limiting the accuracy of the predictive models [187, 49, 116, 18, 232, 7, 161, 158, 244, 223, 202]. Indicatively, the health metic and facial dataset has low representation of different ethnicities, age groups and socio-economic backgrounds thus it limits the capability of the model to generalize within large populations [159].

- *Data imbalance:* Imbalanced datasets, where individual classes are drastically underrepresented as compared to others, pose one of the most serious issues when it comes to the development of models that represent an effective generalization across different patient groups. This has been the case especially when the datasets are used in predicting chronic illnesses and other health aspects that are associated with them [3, 263, 244]. As an illustration, the Deep Learning approach that has been suggested to predict chronic kidney disease was tested on fairly small data volumes, hence limiting its generalizability [232] or the limited access to labeled health data hampers the ability of the MOD-1D-CNN model to succeed in being generalized across various clinical settings [159].

- *Reproducibility crisis:* Reproducibility is an increasing concern in clinical AI because many studies are not publicly available with code, model checkpoints, and clinical datasets to reproduce their findings. Experiments with AI are fragile to implementation and training parameters and by empirical audits, the literature results are usually difficult to reproduce without complete provenance [104, 83]. Cross-site evaluation, in particular within the healthcare context, has shown significant amounts of generalization errors as models that are trained using data of one hospital are tested to the data of another hospital, highlighting how the heterogeneity of the data set and missing data can conceal fragile performance [281]. The field is shifting to more rigorous reporting and release practices (e.g., CONSORT-AI, TRIPOD+AI), standard outputs, and routine sharing of code, seeds, and environment descriptions to allow results to be separately validated and compared [151].

5.1.2 Methodological Constraints

- *Model interpretability and complexity:* Although Deep Learning models have high predictive power, they are often criticized because they lack interpretability and can be described as black boxes. Such a lack of transparency is an obstacle to acceptance in general clinical practice because confidence in the model predictions is fundamental in creating trust between the professionals in the medical field [116, 263, 158, 244]. Techniques like LIME and SHAP are used to provide more interpretable predictions for individual predictions. However, there are still major issues in achieving transparent and comprehensive interpretability of complex models [116, 18, 7]. Moreover, the high computational demands of such models may limit their applicability in resource-constrained settings [3, 161, 244].

- *Feature extraction and selection:* Traditional Machine Learning models mainly rely on manual feature extraction, which is resource-intensive, subjective, and error-prone [3, 161].
  In many cases, Deep Learning systems do not produce the optimal results. For example, the CNN-LSTM architecture can never accurately localize the lesions of the brain [67]. Whereas dimensionality reduction option in feature selection algorithms (e.g., Recursive Feature Elimination(RFE)) may lead to improved model performance, it does not always succeed in identifying the most influential features in disease predictions and as such, could lower accuracy and stability of derived models [232].

- *Combining multiple data modalities:* The integration of data from different modalities—such as medical imaging, EHR, and genomic data, clinical notes—has the potential to enhance the accuracy of predictive models. Nonetheless, there is also a lot of complexity in this integration to the design and interpretability of such models. [18].
  Implementing multimodal data integration requires the use of advanced and complex model designs, the development and optimization of which can be complicated in the circumstance of insufficient access to the dataset, in particular [69, 158].

5.1.3 Computational and Technological Limitations

- *Computational demands:* DL models need substantial computational resources, which can present a considerable barrier in resource-limited settings [49, 3, 116]. This challenge is particularly dominant in the requirement for high-performance hardware and specialized software to manage and process huge volumes of data effectively [276, 161, 263, 159, 244, 202].
  The complex architectures like DeepNetX2 or three-dimensional convolutional network (3D ConvNet) have extensive computation requirements that might not always be feasible and sustainable, particularly with limited technology infrastructure in a health facility [244, 18].

- *Real-time functionality and scalability concerns:* A significant percentage of models cannot be extended or scaled and would not easily apply in real-time, therefore, decreasing their applicability in dynamic and time-sensitive cases, such as in clinical practice. This limitation is especially profound in models that require a lot of extensive computational resources, making them less feasible for routine clinical practice [49, 276, 263].

- *Generalizability across different demographics:* Models trained on small-scoped, defined datasets often exhibit poor performance when deployed in diverse clinical settings, mainly due to data quality disparities, data collection procedures, and patient demographics. [69, 158, 263, 244, 202]. This difficulty is particularly evident in studies focused on retinoblastoma diagnosis, where the practical validity of AI models remains questionable [7].
  Moreover, the widespread adoption of AI technologies in healthcare is hindered by the absence of affordable and convenient methods to access high-quality clinical data, such as retinal imaging, which is necessary to achieve reliable and accurate performance of models [7].

- *Energy efficiency and carbon footprint of DL training:* Large-scale AI systems have been questioned in their impact on energy consumption and the environment because of the computational requirements of training and deploying DL models. The resource intensive nature of modern architectures, particularly transformers, multimodal models and large foundation models necessitate the heavy use of both GPUs and TPUs, which in turn leads to high electricity usage and lots of carbon emission [238, 192]. This is a major problem in the healthcare environment due to the possibility of resource-intensive training pipelines, which can be unfeasible in hospitals with a small computational infrastructure. Recent research points to ways of reducing energy and carbon expenses, including model pruning, quantization, knowledge distillation, energy-conscious neural architecture search (NAS) and the progression of lightweight clinical models optimized for edge deployment [243]. The federated and distributed learning can also decrease the necessity of centralized high-powered training because they allow updating locally at a low cost [142]. The use of energy metrics in model evaluation, including FLOPs, inference latency, and training-energy reporting, can be useful to promote sustainable, scalable, and environmentally friendly development of clinical AI.

5.1.4 Clinical validation and trustworthiness

Clinical validation of Deep Learning models is hindered by the complexity and heterogeneity of EHR data, demographic and setting-related variability, and the challenge of modeling temporal dynamics [49]. Overfitting and limited generalizability due to constrained datasets further impact performance [7, 244]. Clinical validation is imperative to ascertain that predictive models demonstrate consistent performance and reliability within real-world healthcare settings. Meaningful collaboration with healthcare professionals is essential to support the design and execution of clinical studies that assess the safety, efficacy, and practical utility of these models in routine medical practice [187, 3, 18, 69, 50, 176, 7, 161, 158, 223].

5.1.5 Ethical considerations

Ethical issues, including the imperative to ensure that models do not produce discriminatory outcomes, are of critical importance in the development of healthcare AI systems. The involvement of ethics experts throughout the development and evaluation process can aid in identifying and mitigating potential biases, thereby promoting fairness and equitable treatment across diverse patient populations[223].

5.1.6 Fairness, equity, and bias

The problem of equity issues in clinical AI arise because DL models usually reproduce demographic and socioeconomic imbalances present in real-world healthcare data [183, 201]. The underrepresentation of the minority will result in the lack of the model generalizability and the unbalanced diagnostic performance within the race, gender, and socioeconomic levels. It is also demonstrated that models can have implicitly encoded demographic characteristics that can amplify risk of biased predictions and unequal clinical outcome [76]. In addition, differences in quality of data, sensor stability and digital access may also further skew model behavior and increase already existing health inequities [38]. To overcome these challenges, it is necessary to have more representative data, subgroup-consciousness in evaluation, and algorithmic design oriented towards fairness, to achieve equitable deployment.

5.1.7 Fairness metrics

- *Demographic parity:* Demographic Parity (DP) is a measure to determine whether a model gives positive results equally on different demographic groups irrespective of the ground-truth labels. In clinical prediction problems, the value can be used to determine whether an algorithm disproportionately flags some groups as either high-risk or low-risk [91, 240, 68].
- *Equalized odds:* Equalized Odds (EO) demands that both the true positive rate (TPR) and false positive rate (FPR) of a model are equal between groups. This is especially relevant in disease prediction systems due to unequal error rates, which may result in either overdiagnosis or underdiagnosis of particular groups of patients [91, 201, 240, 148].
- *Equal opportunity:* A relaxation of equalized odds which only demands similar true positive rates across groups (e.g., sensitivity is similar across all), and is concerned with missed detections in any group. This is important in healthcare where all the patients with a condition stand equal opportunities of being rightly diagnosed and reducing disparities in access to timely treatment [91, 170, 24].
- *Calibration within groups:* Calibration across groups examines whether predicted risk scores correspond to the actual outcome probabilities are similar across various demographics. Even with high accuracy, poor calibration leads to making harmful clinical decision [170, 252].

5.1.8 Real-time constraints

Healthcare 5.0 applications which include continuous monitoring and emergency response and wearable-assisted diagnostics require real-time performance [135, 92]. Deep learning models require high availability, and can run at resource-limited IoT/edge devices and should provide low-latency predictions with clinical reliability, but most of the state-of-the-art architectures are computationally costly and cannot be implemented on-device or at the bedside. According to previous surveys on IoMT, H-IoT, and smart healthcare, the real-time problems of latency, bandwidth, and computational capabilities and power consumption are still significant unresolved problems, particularly with edge-based and safety-critical deployments [209, 135].

5.2 Challenges and Limitations in Medical Imaging Analysis and Radiology

Medical imaging analysis and radiology encounter a range of challenges and limitations that can impede accurate diagnosis and optimal treatment outcomes. These issues arise from technological constraints and human-related factors, collectively affecting the effectiveness and reliability of medical imaging within clinical practice. This section presents a comprehensive overview of the common challenges and limitations encountered in medical imaging analysis and radiology.

5.2.1 Data Constraints

- *Limited availability of training data:* DL models, particularly CNNs, often require large and diverse datasets to perform effectively. Limited data, such as a few images or the absence of rare or abnormal

cases, can restrict the model's ability to learn comprehensive features and reduce its generalizability to real-world applications [101, 47, 287, 115, 235, 246]. Like the small sample size may hinder the CNN's generalization and diagnostic accuracy for esophageal cancer [97].

- *Heterogeneity and quality of datasets:* The effectiveness of a model is influenced not only by the volume of data but also by its quality. Poor data quality can compromise model performance, while insufficiently diverse datasets may hinder the model's ability to generalize across varying populations or imaging environments [101, 82, 147, 97, 235, 115, 125, 44, 105, 246].
- *Artifacts in medical imaging:* Medical images often contain noise that impacts diagnostic accuracy. While the SDAE can address this, the persistent noisy nature of medical data remains a challenge [47].

5.2.2 Model complexity and algorithmic challenges

- *Probabilistic behavior of algorithms:* Certain algorithms demonstrate stochastic behavior as a result of point sampling within images, which can cause variability in performance across different executions [287].
- *Complexity in feature extraction:* Traditional methods require complex feature extraction and selection, which can be time-consuming and may not offer significant advantages over simpler morphological features used in clinical practice [47]. Sometimes, deep networks implicitly learn features that may be unknown or overlooked by humans, limiting the clarity and interpretability of model decisions [101, 82, 287, 125].
- *Algorithmic inefficiencies:* The performance of image processing methods like segmentation and classification is often limited by current algorithms, which struggle to manage the complexity and variability of medical images. For example, automated glioma segmentation is hindered by the complexity of MRI data, driven by variability in tumor size, shape, and location, making effective algorithm development difficult [101, 287, 97, 125, 44].
- *Integration of diverse imaging techniques:* Analyzing images from diverse modalities poses challenges due to variations in image characteristics and the necessity for modality-specific algorithms [47, 235, 115].
- *Restricted generalization capability:* DL models trained on specific datasets may lack generalizability, potentially reducing performance across diverse populations and settings, highlighting the need for further research in broader demographic contexts [147, 235].

5.2.3 Explainability and interpretability challenges

- *Explainability and accountability:* Similar interpretability concerns arise in medical imaging models, particularly when complex feature hierarchies are involved[147, 235, 246].
- *Erroneous classification outcomes:* Automated systems can yield false positives, prompting unwarranted interventions, or false negatives, leading to missed diagnoses [147]; in many cases, these errors are primarily linked to overlapping or coexisting eye conditions [97].

5.2.4 Ethical and Compliance Considerations

The application of AI in medical imaging presents regulatory and ethical challenges, such as ensuring transparency and mitigating algorithmic bias, which may impede its integration into clinical practice [47, 147, 235, 115].

5.2.5 Legal accountability and responsibility

One of the essential dilemmas in implementing the DL-based systems to Healthcare 5.0 is the establishment of clear responsibility and liability frameworks in the case of the AI wrong diagnosis of patients. Contrary to the conventional methods of clinical decision-making, AI-based diagnosis contains multifaceted interactions between software developers, medical professionals, and organizations, and it is difficult to determine

accountability, which is legally ambiguous. The current medical malpractice legislation is poorly suited to deal with the errors caused by algorithmic bias, software bugs, and model drift. The new regulatory environment of the European Union, with new AI Act and amended Product Liability Directive, makes the providers of AI systems very demanding in terms of safety, transparency, and risk management, but there still are challenges of implementation and enforcement. Moreover, the medical workers will still be responsible in the circumstances when they do not correctly incorporate AI outputs into clinical decision-making or overruling inappropriate recommendations.These unresolved legal issues create barriers to adoption, restrict innovation, and obscure pathways for error reporting and patient compensation [254, 253].

5.2.6 Real-world and Clinical Integration

- *Overburdened radiologists:* Radiologists handle large workloads, leading to fatigue and the risk of overlooking critical findings. While automated systems can ease this burden, their integration presents significant challenges [82, 246].
- *Difficulties in manual interpretation:* Current diagnostic processes rely on manual image interpretation, which is time-consuming and prone to human error [115]. This method also depends on skilled radiologists, who may be scarce, particularly in underserved regions [82, 147, 97, 125, 246].

5.2.7 Contextual Limitations

- *Computer-aided diagnostic systems:* Deep Learning shows potential in CADx but remains less explored. Conventional CADx systems' reliance on explicit feature design presents a challenge that Deep Learning seeks to overcome [101, 47].
- *Necessity for advanced methods:* There is a critical need for precise automated disease detection, as traditional methods are often inadequate. Advanced approaches, particularly those employing CNNs, are essential to enhance the accuracy and efficiency of tumor identification [101].
- *Limited healthcare access:* In numerous regions, particularly in underserved or rural areas, access to specialized healthcare services is limited, impeding early disease detection and intervention [246].

5.3 Challenges and Limitations of Multimodal Deep Learning in Healthcare

There are several limitations and challenges related to healthcare MMDL and their application that occur during data collection and pre-processing, as well as model deployment and understanding. Important challenges are data imbalance in the classes, insufficient amounts of data, prevalence of pre-processing, interpretability of models, and the challenges associated with choosing the best fusion methods. It is essential to overcome these current challenges in the deployment of advanced models in clinical practice [206, 26, 54, 273, 288].

5.3.1 Data-related Challenges

- *Data source complexity:* A healthcare system has diverse entrants of data, some of which may be an image (X-rays, CT scan), clinical measurements (PFTs), or EHR, and even at a more fundamental level, genetic information. It is a great challenge to integrate these divergent types of data that are of different formats, scales, and internal structures [257, 283, 58, 130].
- *Lack of adequate integration strategies:* Numerous current multimodal methods suffer due to inferior integration strategies, making the methods have a low diagnostic capability. Concatenation of features may fail to find the complex relationship among different modalities [137, 143, 226, 138, 247, 130, 204].
- *Use of simulated patient data:* Another major limitation of the past research with linguistic models is that they have used simulated patient data, and they lack authenticity in a large-scale clinical environment. This leaves doubts about the autonomous applications and practical deployment [89, 155].
- *Contextual information deficiency:* Most of the available schemes fail to consider the context of a piece of information when dealing with text data of medical texts or some missing values. It has the potential to result in less detailed feature extraction, thereby overlooking potentially important diagnostic

or prognostic information [89]. In particular, the status of gene mutation may be characterized by a high rate of missingness (e.g., 10.56% in total, 7.41% in gene data, and 3.15% in the other clinical data), making model training and reliability complicated [141].

- *Data quality and volume dependency:* The quality and quantity of the multimodal data are very important factors in the accuracy and effectiveness of the multimodal models. Moreover, a deep integration process itself is so crucial that all the shortcomings in the quality of data or a lack of its quantity may influence the work of the model to a great extent [89, 143, 257, 283, 141, 58, 204].
- *Single disease oriented:* Most past studies on AI imaging and multimodal imaging are single-disease oriented and do not easily integrate with others that are single-disease oriented. This discontinuity among disciplines reduces the general functionality and poses a threat of the generalizability of the models because of spectrum bias.
- *Small public dataset and scarcity of multimodal dataset:* In particular, in healthcare, the lack of a huge sample size in the datasets represents a major bottleneck for the data integration of DL-based models [143, 141, 283, 58, 204]. Indicatively, the ADNI dataset contains only a few thousand samples, and even fewer have the three modalities (imaging, EHR, and genetic data) [257, 134], and publicly shared small THz imaging datasets hinder the development of robust AI-enabled THz applications [73]. This shortage may cause suboptimal performance because of insufficient training data for the dense networked DL.

  This field is also lacking in multimodal datasets that would offer one-to-one mappings between other representations (e.g., text and images, or images that belong to different sources). This compels existing research to often use labels from image datasets as text input, which in turn results in inferior model performance [257].
- *Dependence on sensitive medical records:* Multimodal models tend to rely heavily on the information provided in the medical records of the patients, which include highly confidential medical information. Such dependency is not very practical when implementing in real-life operations because of privacy and data governance regulations. Any need to access and use such data demands strict ethical clearances and compliance with the data protection legislation, restricting its usage to the general population and wide-scale research cooperation [226].

5.3.2 Interpretability Challenges

These interpretability challenges are further intensified in multimodal architectures due to the complexity of integrating heterogeneous data sources[141, 283, 106, 247, 130, 58, 73, 204].

5.3.3 Generalization and Validation Challenges

A model trained with a particular dataset or on a set of population may fail to apply to other previously unseen data in other hospitals, areas, or even demographics. Clinical deployment Real-world deployment of the model depends crucially on ensuring its robustness and its generalizability [257, 138, 283, 130] Although there are good internal results of cross-validation, certain multimodal combinations, especially those with extremely weak overlapping data and small, perform poorly in the external validation sets. This demonstrates that combining or clustering on several datasets is a challenge in making sure that these models work well on the unseen set of data [257, 138, 155, 130].

5.3.4 Clinical Adoption and Workflow Integration

The current research suggests that XAI does not intrinsically increase the performance of clinicians and can often lead to automation bias or over-reliance when clinicians follow inaccurate AI results because it is backed by a plausible-looking explanation (e.g., a heatmap emphasizing a relevant organ when the diagnosis is incorrect). Contrarily, complicated or unclear explanations may reduce trust and raise cognitive load, and interfering optimal clinical practices instead of simplify them. It needs a more human-centered evaluation framework since technical metrics (such as SHAP fidelity) are unsuitable to be correlated with the real-world requirements of bedside decision making [123, 20, 71, 32].

5.3.5 Deployment and Practical Problems

- *Computational resource:* The time and costs of training and deploying MMDL models can be prohibitively expensive, as they often demand high-powered computing resources (e.g., access to a high-performance Graphics Processing Unit (GPU)), which may make it infeasible to deploy MMDL models at most research institutions or by healthcare providers [137, 106, 155, 204, 134].
- *Real-world deployment needs:* Problems that could arise with real-world deployment, including how to fit these systems into a real-world clinical workflow and how to support scale and robustness in a variety of clinical environments, have yet to be fully addressed [106, 155, 204].
- *Clinician feedback integration:* Although the improvements in areas of feedback loops occur to capture clinical realities, ongoing adjustments and confirmation through clinician input are required in the implementation of any real-life situation incorporation [226, 106, 204].
- *Training optimization conflicts:* End-to-end multimodal models may encounter conflicts during backpropagation, as the model finds it challenging to concurrently train both the sub-model (e.g., for medical images) and the fusion module. This competing learning pressure can cause the model to under-train the sub-module, limiting its ability to capture critical features [141]
- *Loss of information:* In numerous multimodal fusion methodologies, medical images characteristics must be transformed into one-dimensional (1D) tensors to align with the format of clinical data. This procedure may lead to a substantial loss of spatial information, essential for precise diagnosis [141].

5.3.6 Cross-Cutting Ethical, Regulatory, and Fairness Challenges

The deployment of deep learning models in Healthcare 5.0 requires careful consideration of cross-cutting ethical, regulatory, and fairness-related challenges. Regulatory approval and compliance demand transparent model behavior, robust validation, and clear accountability to ensure patient safety and system reliability. Ensuring fairness across diverse patient populations and minimizing bias are essential to prevent unequal healthcare outcomes. Addressing these interconnected issues is critical for achieving trustworthy and responsible real-world implementation of AI-driven healthcare systems.

Table 17: Literature review

Key literature work	Pros	Cons	Research Gap
1. Disease Prediction and Early Diagnosis			
[187]	1. High accuracy 2. Automated diagnosis 3.Uses pre-trained models for feature extraction	1. Requires significant computational power 2. Limited dataset size, only 900 images 3. Interpretability of models remains a challenge	1. Limited dataset size 2. Need insights into feature importance or decision-making of the model 3. Not validated in real clinical settings or diverse demographic groups
[49]	1. Improved predictive performance 2. Temporal modeling 3. Scalability	1. Require significant computational resources and expertise 2. Data limitations that may limit generalizability 3. Interpretability	1. Dataset limitation 2. Further exploration into more feature representations might enhance performance 3. Do not address how predictions could be integrated into decision-making processes 4. Lacks clinical implications
[3]	1. Fully automatic 2. Unaffected by the ECG signal quality 3. Incorporates a ten-fold cross-validation technique	1. Extensive hours of training, sophisticated technology such as GPUs, computationally demanding	1. Class imbalance in datasets 2. Lack of comprehensive validation in clinical settings 3. Not a fully automated system; additional feature extraction or selection required 4. More extensive datasets are needed
[116]	1. Improved performance 2. Effective deep feature extraction capabilities 3. A reliable baseline for breast mass categorization	1. Mammography breast masses differ in appearance 2. Challenging to catalog in the original data space 3. Time-consuming	1. Restricted dataset size 2. Dependence on pretraining 3. Lack of cross-dataset validation 4. High Computational Requirements 5. Omission of Interpretability Analysis 6. Limited Integration of Clinical Parameters

Table continues on next page

Table continued from previous page

Key literature work	Pros	Cons	Research Gap
[18]	1. Efficient and simple 2. High performance 3. Automatic feature extraction 4. Large dataset	1. Performance drop on large datasets 2. Preprocessing dependent 3. Higher computation cost 4. Absence of multi-modality 5. Restricted Generalization	1. Integration of Multiple Modalities 2. Enhanced diversity in datasets 3. Enhanced architectural efficiency 4. Interpretability of predictions 5. Validation in clinical practice
[67]	1. No prior knowledge required for manual feature extraction 2. Eliminate subjectivity 3. Faster convergence 4. High accuracy	1. Limited diagnostic performance for sMCI 2. Lack of direct identification of cerebral lesion structures 3. Underutilization of longitudinal MRI data	1. Lack of use of longitudinal data 2. Direct localization of brain lesions is not feasible 3. Lacks the exploration of shared feature representations across different imaging modalities
[69]	1. Higher performance 2.Efficiency 3. Non-invasive and non-irradiating 4. Automated classification 5. 3D Feature extraction	1. Insufficient data, 2. Inadequate comparison with radiologist-based diagnoses, 3. Inability to handle unusual presentations 4. Uncertain long-term effectiveness	1. Insufficient data 2. Requires validation with follow-up imaging and external datasets 3. Limited generalizability without cross-platform evaluation 4. Performance in early or atypical AD remains unclear
[50]	1. Elevated precision 2. Improved feature extraction 3. Robust validation metrics	1. Computationally demanding 2. Limited dataset 3. Risk of overfitting	1. Necessity for an expanded dataset 2. Comparative analysis with alternative hybrid models 3. Empirical clinical validation
[6]	1. High detection accuracy 2. Increased generalization and resilience 3. Quick, economical X-ray diagnosis	1. Limited CT scan exploration 2. Small dataset size 3. Although infected, some patients test negative	1. Limited diagnostic methods 2. Need for rapid detection 3. Insufficient data availability
[176]	1. Precision and efficiency 2. Real-time monitoring 3. Remote patient surveillance 4. Effectual data handling	1. Security and privacy of data; 2. Reliance on IoT devices 3. Flexibility and cooperation 4. Latency and bandwidth usage 5. Price and availability	1. Lack of clinical validation 2. Evaluation based on synthetically augmented datasets
[232]	1. Achieved high accuracy 2. Uses RFE for feature selection, improving efficiency 3. Outperforms traditional ML classifiers in accuracy	1. Limited dataset 2. Limited feature used 3. Overfitting risk 4. Lack of real-world Testing	1. Need larger datasets 2. Need for advanced features
[7]	1. Good categorization accuracy 2. Improved readability using LIME and SHAP	1. Bias in dataset composition 2. Insufficient clinical evaluation 3. Restricted availability of imaging modalities	1. Restricted dataset heterogeneity 2. Clinical validation 3. Relevance to additional ocular disorders 4. Enhancement of explainability 5. Accessibility and economic viability
[276]	1. Remarkable precision 2. Utilization of IoT network 3. Implementation of the Bi-GRU Attention Model 4. Recommendation framework 5. Validation of cloud simulation	1. data integrity 2. Adaptive medical conditions 3. Requirement for additional testing	1. Limited integration of recommendation systems 2. Higher computational overhead 3. Lack of adaptability to dynamic real-time patient data
[161]	1. High accuracy 2. Efficient feature selection 3.Advanced pre-processing techniques 4. Combines multiple methods 5.Comprehensive testing 6.Practical application 7.Handles large datasets	1. High computational resources required 2. Potential over-fitting 3. Complexity 4. Clinical validation needed 5. Real-time application challenges 6. Dependence on quality of data 7. Limited generalizability	1. Insufficient clinical validation 2. Larger dataset needed 3. Management of missing data 4. Techniques for feature selection 5. Flexibility of interpretation 6. Level of resource consumption
[158]	1. Exceptional precision 2. Elucidation 3. Resilient Ensemble models 4. Thorough assessment	1. Complication 2. Generalizability 3. Incorporation into clinical workflows	1. Need for generalization 2. Advancement of XAI methods 3. Assessment of varied datasets 4. Integration of multimodal data 5. Lack of real-world clinical validation 6. Performance enhancement through optimization
[263]	1. Elevated predictive precision 2.Interpretability 3. Effective computation 4. Hybrid model architecture 5. Augmented trust	1. Hardware-intensive 2. Dataset imbalance 3. Accessibility challenges	1. Explainability in AI 2. Generalizability issue 3. Hardware dependence 5. Model scalability

Table continues on next page

Table continued from previous page

Key literature work	Pros	Cons	Research Gap
[244]	1. Higher accuracy 2. Enhanced transparency using Explainable AI 3. Improved interpretability and efficiency due to effective feature selection 4. Reduced computational complexity 5. Faster inference time	1. High computational complexity. 2. Extensive preprocessing is required 3. Dependent on higher-quality data 4. Difficult to generalize due to data set restriction	1. Limited dataset 2. Need for optimization to minimize model complexity 3. Limited model transparency
[202]	1. improved feature extraction 2. Better accuracy 3. Manages imbalanced data through SMOTE 4. Minimized overfitting	1. High computational complexity 2. Dataset size affects SVM's performance 3. Real-time optimization is needed	1. More optimization is required to lower computational overhead 2. Different classifiers could increase productivity 3. Need varied datasets 4. Improved generalization required
[223]	1. High accuracy 2. Comparative analysis 3. Comprehensive metrics 4. Clinical relevance 5. Automation potential	1. Ethical considerations 2. Data limitations 3. Limited clinical validation 4. Need for more explainable AI approaches 5. Need more hyperparameter optimization	1. More thorough ethical consideration is needed 2. Need larger and more diverse datasets covering all demographics 3. Clinical validation is necessary
[159]	1. Innovative design 2. User-friendly installation 3. High accuracy 4. Portability 5. Reliability and security 6. Rapid responsiveness 7. User-friendliness 8. Cost-effectiveness	1. Requires enhancements in the 2D gaming application's user interface layout. 2. Necessitates improvements in robustness and scalability of health metrics and facial imaging.	1. Restricted demographic representation 2. Inconsistencies due to variable lighting conditions 3. Insufficient availability of health-related data 4. Variability in input accuracy based on user interaction 5. Inconsistent behavioral patterns during gameplay 6. Limitations imposed by technological infrastructure
2. Medical Imaging Analysis and Radiology			
[101]	1. Handcrafted features not required. 2. Good results in classification without using texture or morphology features	1. 3D features not incorporated 2. Resizing input images leads to loss of information related to size, which is an important indicator.	1. Incorporating diverse datasets 2. Utilizing advanced Deep Learning models 3. Difficulty in feature extraction
[82]	1. High performance accuracy 2. Consistent results 3. Scalable 4. Evaluation based on Specificity and sensitivity-2 operating points	1. Limited scope as only suitable for RDR and DME 2. Algorithm may miss minute features as ophthalmologists missed them 3. High dependency on training data 4. Cannot replace traditional systems	1. Testing algorithm on more datasets 2. Integrator variability 3. Assessing feasibility in real-world situations
[47]	1. Handcrafted features not required 2. Works well with varied modalities 3. Noise-resistant	1. Limited breast US data size 2. Data quality dependency 3. Not validated across modalities	1. Requirement of a larger dataset for reliable results 2. Validation across modalities required
[287]	1. Elevated segmentation precision 2. efficient, robust, and scalable 3. Novelty	1. Dataset dependency 2. Stochastic element 3. Limited training 4. Computational limits	1. Depends on varied datasets for validation 2. Imbalanced sampling impacts 3. Entropy sampling needs improvement 4. Underperforms deep CNNs on huge datasets
[97]	1. High sensitivity (98%) for detecting esophageal cancer 2. Identification of Small Lesions 3. Rapid analysis time (27 seconds for 1118 images) 4. Capability to distinguish between superficial and advanced cancers	1. Positive predictive value relatively low (40%) 2. Misdiagnosis due to shadows and normal structures resulted in a negative predictive value of 95% 3. Dependent on image quality	1. Need for a diverse dataset 2. Real-time validation needed 3. Poor-quality image handling 4. Need for further training
[147]	1. Attained exceptional performance 2. Potentially advantageous for large-scale, cost-efficient glaucoma screening 3. Effective model construction utilizing an extensive, annotated dataset	1. Absence of interpretability 2. Constrained to photos from a certain place (China) 3. Elevated misclassification rates for instances with coexisting ocular problems (e.g., high myopia)	1. Generalizability is needed 2. Interpretability issue 3. Need for clinical validation 4. Managing confounding variables 5. Limitation in data collection and population diversity 6. Constraints related to image quality and evaluation accuracy
[115]	1. Elevated precision, sensitivity, and specificity 2. Efficient cancer segmentation via hyperspectral imaging 3. Decreases the need for specialized oncologists	1. Small data limits generalization 2. Computationally intensive 3. Not clinical workflow-integrated	1. Needs more diversified data 2. Requires real-time clinical implementation 3. Only limited comparison to advanced Deep Learning algorithms

Table continues on next page

Table continued from previous page

Key literature work	Pros	Cons	Research Gap
[235]	1. Elevated sensitivity 2. Real-time image predictions (0.07 seconds) 3. A promising FNA-reduction tool 4. Training photos need little preprocessing	1. Reduced specificity 2. Comparatively limited training dataset 3. Poor augmentation or segmentation may reduce model robustness	1. Relatively small training datasets 2. Insufficient data augmentation to simulate variability 3. Clipped nodules may overlook anatomical characteristics 4. Validation on larger, more diverse datasets is needed
[125]	1. High categorization accuracy 2. Integration with IoHT allows remote diagnosis 3. Automated detection cuts errors 4. Effective data augmentation boosts model performance	1. Intensive computer instruction 2. Lack of dataset diversity hinders generalization 3. Highly dependent on IoHT infrastructure and resources	1. Dependence on the quality of input data 2. Validation needed on varied and larger datasets 3. Consider sophisticated IoHT integration for scaling 4. Human Error in diagnosis 5. Optimize computation
[190]	1. High classification accuracy 2. Quick, affordable diagnosis 3. Non-contact diagnosis 4. Depends less on RT-PCR	1. Small dataset 2. Possible over-fitting despite data augmentation 3. Image quality and variety affect model performance	1. Limited training data limits generalization 2. More diverse, larger datasets improve robustness 3. Clinical validation needed
[131]	1. High performance 2. Potential integration into healthcare systems 3. Real-time monitoring and recovery tracking	1. Limited by the quality and variety of the training datasets 2. Need for significant computing resources 3. Different patient demographics and conditions may affect the model's efficacy 4. Human assistance may be needed for preprocessing	1. Need for a larger dataset 2. Requirement for enhanced 3D-models 3. Integration of multimodal data 4. Real-world evaluation
[63]	1. High accuracy and sensitivity in predicting mortality 2. Provides objective metrics for assessing lung infection severity	1. Limited by the variability of X-ray quality across different machines 2. Relies on a relatively small dataset for training and validation	1. Variability in the quality of chest X-ray images 2. Integration with other modalities
[44]	1. Achieved high accuracy 2. Tumor detection is automated, saving time 3. Well-manages tumor size, form, and intensity	1. Needs huge, diversified datasets to train 2. Computationally costly training	1. Limited clinical and dataset generalization 2. Improve intensity biases and MRI fluctuations
[178]	1. Very accurate pulmonary disease categorization 2. LIME makes clinical decisions more explainable 3. Effective Transfer Learning small datasets	1. Small datasets hinder generalization 2. Reliance on ResNet50-trained models 3. Explainability evaluation needs expert acceptance	1. Insufficiently diversified, huge datasets for generalization 2. Limited demographic and clinical tests 3. Additional features needed to strengthen the model
[105]	1. High accuracy without the use of pre-trained processes 2. Model generalizes well 3. Good results in few (only 50) epochs	1. Computationally intensive due to a large number of parameters 2. Limited labelled data 3. High resource demanding during training	1. Using CBAMWDNet for other diseases 2. Incremental learning and continuous training with new data 3. Optimizing performance with ensemble models. 4. Improving weights with multi-objective optimization
[222]	1. Good performance on both binary and multi-class datasets 2. High accuracy 3. Robust feature extraction using VGG-16	1. Computationally expensive, the model has a high number of parameters 2. Validation on real-world scenarios is required	1. Validation on a larger dataset is required 2. Optimize architecture 3. Use of other architecture
[156]	1. High accuracy and AUC score 2. Better performance than InceptionV3, ResNet-50, VGG16	1. Many CNN layers 2. Requires a good GPU 3. Post hoc visualization lacking on tumor areas	1. Validation with diverse datasets is required 2. Do tumor localization and post hoc visualization
[184]	1. Handles heterogeneous MRI and PET data effectively 2. XAI provides interpretability 3. Better classification performance due to the fusion of features	1. Low sensitivity 2. Requires a high GPU	1. Validation with diverse datasets required 2. Optimization required
[99]	1. High accuracy through data augmentation 2. Class activation mapping (CAM)-based explaining predictions 3. Easy, affordable implementation	1. Computer-intensive models 2. Optimized ViT (vision transformers) training requires larger datasets	1. Performance of ViTs is limited by dataset size 2. More robust augmentations and lightweight models are possible

Table continues on next page

Table continued from previous page

Key literature work	Pros	Cons	Research Gap
[121]	1. High accuracy 2. Efficiency 3. Process huge datasets fast 4. Improved interpretable mechanism 5. Robustness across classifications 6. Handling large data sets accurately	1. Limited data size 2. Highly resource-intensive	1. Limited availability of external dataset 2. Use of diverse clinical scenarios 3. Investigating methods to enhance the interpretability of the model 4. Advanced interface required
[246]	1. High accuracy 2. Better segmentation accuracy 3. End-to-end pipeline	1. Limited data size 2.Highly resource-intensive 3. Low transparency	1. Need for interpretability 2. Diverse dataset usage requirement
3. Multimodal Deep Learning in Healthcare			
[150]	1. Improved accuracy 2. Effective multimodal fusion 3. Well performance on low label data 4. High-level feature learning 5. Unsupervised capability	1. Dataset limitation 2. Computational-demanding training needed 3. Limited performances in small classes	1. Validation on a more diverse dataset is needed 2. The unlabeled data performance is not widely studied 3. Test needed for real-time applicability and scalability
[100]	1. Achieve closer to supervised performance with very few labels 2. The mixture of annotated and unannotated data is even more advantageous 3. Best performance when applying the report sections of the clinic description	1. Text embeddings trained without domain awareness are poor 2. Pre-trained Models offer limited generalizability 3. Unsupervised rocrustes refinement minimally effective 4. Low Medical Record Review (MRR) demonstrates report similarity	1. Domain-specific text encoders needed 2. Generation tasks exploration 3. Broader modality fusion inability 4. Sparse evaluation of downstream clinical activities
[89]	1. Integration of multimodal data 2. Detailed feature extraction 3. Efficient feature integration 4. Exceptional precision 5. Outstanding efficacy	1. Cost of computational 2. Dependency on data	1. Ineffective fusion strategy for multimodal information
[137]	1. Incorporates heterogeneous data modalities 2. Effectively processes irregular temporal sequences 3. Utilizes both overlapping and non-overlapping datasets 4. Strong predictive accuracy 5. Improved sensitivity and balanced accuracy	1. GRU parameters remain static during the final integration phase 2. Initial learning occurs independently across modalities 3. Risk of missing cross-modal feature interactions 4. Predictive accuracy declines over extended time horizons	1. Lacks unified optimization across data modalities 2. Cross-modal feature integration remains suboptimal 3. Future direction includes incorporating genomic and imaging data 4. Intends to design a feedback-enabled, unified GRU framework
[143]	1. Efficacy of multimodal data 2. Advantages compared to other self-supervised methods 3. Resistance to computer-generated images 4. Generalizable and transferable features	1. Limited FFA fundus images 2. Initially unsupervised, but final classification relies on labeled data	1.Fundus-FFA images are limited 2. Not fully unsupervised 3. Lacks multi-modal mutual information modeling 4. Needs to extend to the other imaging modalities 5. Synthesis required to generate another modality via adversarial learning
[257]	1. Better performance than shallow learning models 2. DL-assisted multimodality data analysis performs better than single-modality DL models 3. Can extract high-performing features.	1. Limited sample size of the ADNI dataset 2. Combined SNP and imaging modality exhibited reduced performance 3. Model is not fully trained for all modalities	1. Limited dataset 2. No evaluation on early/atypical AD 3. Limited generalizability 4. No validation on external datasets
[226]	1. Enhances prediction accuracy 2. Supports clinical decisions 3. Enables early discharge 4. Integrates multi-modal data with interpretability	1. Requires physician training 2. Limited model flexibility 3. Opaque deep layers 4. Potential data privacy issues	1. Incomplete hospital-standard analysis 2. Limited multi-modal studies 3. Absence of post-treatment recovery prediction 4. Lack of clinical-CXR integration for decision simulation
[138]	1. Enhanced accuracy 2. Non-invasive risk prediction 3. Future event association 4. Model interpretability	1. Retrospective design 2. Restricted SMC data access 3. Geographic bias 4. Need for prospective validation	1. No prospective validation 2. Geographic limitation 3. Absent real-time use 4. Need for broader multimodal integration
[283]	1. Unsupervised representation learning 2. Models complex temporal relationships 3. Incorporates uncertainty estimation 4. Learns discriminative features 5. Achieves high accuracy 6. Demonstrates broad applicability	1. Prior methods emphasize linear correlations 2. Assumes fully supervised settings 3. Ignores intra-variable temporal patterns 4. Susceptible to overconfidence and low reliability 5. Simplistic graph construction 6. Risk of representation over-smoothing	1. Fails to capture complex temporal correlations 2. Limited use of unlabeled clinical data 3. Lacks uncertainty quantification 4. Ignores temporal intra-correlations 5. Overconfident models lack reliable uncertainty scoring

Table continues on next page

Table continued from previous page

Key literature work	Pros	Cons	Research Gap
[141]	1. Enhanced clinical-image feature synergy 2. Spatial information retention 3. Superior to unimodal and prior fusion models 4. Robust unsupervised feature learning	1. Limited sample size 2. Incomplete clinical data 3. Persistently low mean accuracy	1. Small dataset impacts model stability 2. Improved clinical interpretability required 3. Future work to include larger, more diverse datasets
[106]	1. Multimodal approach improves AUC by 2.45 percent 2. Enables early risk identification in SCC cases 3. Leverages unlabeled data effectively	1. Dataset access is restricted 2. Relies solely on MRI data 3. SCC outcome prediction remains uncertain	1. Cross-site validity untested 2. Lacks PET/clinical data 3. No real-time use shown
[247]	1. Achieves high accuracy 2. Improved tumor visibility via fusion 3. Enhanced interpretability with XAI	1. Restricted to pixel-level fusion 2. Lacks segmentation methods 3. Uses limited architectural diversity	1. Current models provide limited interpretability 2. Advanced fusion techniques are required 3. Limited transparency in CNN architectures
[155]	1. Improved diagnostic accuracy 2. Integrative methodology 3. Efficient processing 4. Early-stage detection 5. Resilient architecture 6. Adaptable design	1. Limited data reliability 2. Reduced model transparency 3. High computational demand	1. Unclear dataset scope 2. Lack of clinical validation 3. Absence of real-time testing 4. Incomplete multimodal data integration
[130]	1. High diagnostic accuracy 2. Holistic analysis 3. Handling Imbalanced Data efficiently 4. Better outcomes 5. Strong data fusion	1. Limited generalization 2. Dataset dependence 3. Untested real-time use	1. No external validation 2. Lacks demographic data 3. Needs multimodal model improvement
[58]	1. Enhanced precision 2. Efficient utilization of electronic health records (EHRs) 3. Integration of multimodal data 4. comprehensibility.	1. Computational intensity 2. Restricted generalizability 3. Reliance on superior EHR quality	1. Requirement for expanded, varied datasets 2. Enhanced fusion methodologies 3. Capabilities for real-time prediction
[204]	1. Enhanced predictive accuracy 2. improved model interpretability 3. Effective integration of multimodal data 4. Strong generalizability across diverse populations 5. Increased robustness 6. High AUC-ROC performance metrics 7. Positive clinician feedback	1. Dataset is restricted 2. Clinical validation incomplete 3. computationally intensive 4. Lacks evaluation in varied clinical settings	1. Deficiency in Deep Learning advancements 2. Inadequate multi-modal data fusion 3. Restricted model interpretability 4. Challenges in early-stage disease recognition 5. Obsolete data fusion techniques 6. Limited use of explainable AI methods
[134]	1. Superior DenseNet performance 2. Enhanced feature propagation and reduced gradient issues 3. Strong model robustness 4. Multimodal framework potential 5. Effective early-stage detection	1. Limited accuracy of MobileNet 2. Suboptimal performance of VGG-16 3. Moderate results from CNN 4. High computational resource demands 5. Restricted dataset scale	1. Constraints of the current diagnostic approach 2. Limited sensitivity to early-stage detection 3. Insufficient generalizability across diverse populations
[73]	1. Safe, non-invasive imaging 2. Enables early cancer detection 3. High sensitivity and tumor specificity 4. Captures spatial-temporal data 5. Robust multimodal integration 6. Enhanced prediction accuracy 7. Benefits of decision-level fusion 8. Achieves high classification accuracy 9. No overfitting or underfitting observed	1. Scarcity of THz datasets 2. Expensive THz imaging systems 3. Difficulty in tissue sample acquisition 4. Challenges in manual diagnosis 5. THz is still under development 6. Restricted spatial resolution	1. Limited Deep Learning use in THz/IR imaging 2. THz datasets are scarce 3. Lacks THz post-processing frameworks 4. Data management and interoperability require improvement 5. Multimodal validation needed
[227]	1.High accuracy 2. Efficient tuning 3. Strong fusion 4. Good generalization 5. Minimal overfitting	1. Large model size 2. Limited multi-scale capability 3. 2D-only segmentation	1. No native 3D support 2. Weak multi-scale modeling 3. High storage demand

Table continues on next page

Table continued from previous page

Key literature work	Pros	Cons	Research Gap
[249]	1. Concurrent prediction 2. Efficient computation 3. Correlation modeling 4. Interpretability 5. Robustness 6. Rich feature representation	1. Imbalanced data 2. Dependence on clinical records 3. Fixed prediction time-frame	1. Exploration of chronic disease interrelations 2. Temporal comorbidities, early risk factors in youth 3. Long-term prediction from shorter histories 4. Data imbalance mitigation 5. Annual disease incidence forecasting

5.4 Methodological quality, validation practices, and bias assessment

In the three areas of DL in Healthcare 5.0 Disease Prediction and Early Diagnosis, Medical Imaging Analysis and Radiology and MMDL validation practices display significant heterogeneity and heterogeneity that affects the methodological soundness, with the majority of studies using internal schemes e.g. k-fold cross-validation or hold-out splits of individual institutions or datasets e.g. ADNI of neuroimaging and small Kaggle-style COVID-19 X-ray collections, with very few studies employing external multi-center cohorts or temporal separations [130]. The nature of datasets differs wildly: small datasets (thousands of patients with rare outcomes) are used in early-diagnosis EHR models, tiny (<200 originals) sets are used in imaging studies, medium-sized multi-center MRI/PET datasets have class imbalances (fewer MCI converters), and multimodal works are large EHR-centric datasets to moderate single-center fusions (such as COPD X-ray+Pulmonary Function Test (PFTs)). There are widespread data leakage risks, such as augmentation or patch extraction prior to splitting in imaging and multimodal pipelines, per-visit (as opposed to per-patient) splits in EHR prediction, and slice-level CV in neuroimaging, which interfere with clinical belief in Healthcare 5.0 [247, 226]. These dimensions are specifically recorded in the accompanying representative studies summary table, which in turn allows organization and cross-domain comparison of the studies, but also reveals that patient-level external validation, leakage audits, balanced reporting, and standardized XAI are required to promote methodological rigor [150, 141].Table 19 presents a critical analysis across the three domains of Deep Learning in Healthcare 5.0, detailing validation type, Metrics used, Data balance , Data leakage risks. XAI used.

5.5 Quantitative aggregation of results

Table 18 presents a quantitative aggregation of reported classification accuracies for different disease categories using deep learning based models across multiple benchmark datasets. For each disease, accuracy values reported in the literature were statistically aggregated to compute the *mean accuracy* and the *sample variance*, providing a consolidated measure of overall model performance as well as performance consistency across heterogeneous datasets.

Let $x_1, x_2, \ldots, x_n$ denote the reported accuracy values (in %) for a given disease across n studies or datasets. The mean accuracy (μ) is calculated as the arithmetic average of the reported accuracies:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3}$$

The sample variance (s^2) is computed to quantify the dispersion of accuracy values around the mean, reflecting the robustness and stability of model performance across different datasets:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{4}$$

Table 19: Critical evaluation of Deep Learning Studies in healthcare

Key literature	Validation type	Metrics used	Data balance	Data leakage risks	XAI used
1. Disease Prediction and Early Diagnosis					

Table continues on next page

Table continued from previous page

Key literature	Validation type	Metrics used	Data balance	Data leakage risks	XAI used
[187]	Subject-level k-fold CV on ADNI using AAL-parcellated GM patches, with region-wise unsupervised DBNs combined via ensemble voting and SVM fusion	Accuracy, sensitivity, specificity, ROC-AUC	Moderate class imbalance, notably for converters, with many patches derived from a limited set of subjects	Subject-level CV limits patch leakage, but pre-CV preprocessing and no external test set may still yield dataset-specific optimism	Interpretability is limited to AAL-based importance maps, with no model-agnostic XAI, keeping explanations task-specific.
[49]	Train/test evaluation on a single health-system EHR, comparing RNN-GRU with several baselines, with varying observation windows but no external validation	AUROC is the main metric, with the RNN outperforming baselines at both 12- and 18-month windows, plus limited sensitivity/specificity reporting	Highly imbalanced cohort (12% cases) from a single health system, with class balance kept realistic rather than synthetically adjusted	Patient-level, temporally aligned splits limit leakage, but same-system tuning without external validation may yield optimistic performance	None
[3]	10-fold CV conducted at the beat level on MIT-BIH, using both original and synthetically balanced data, without strict patient-level separation	Per-class and overall accuracy, sensitivity, specificity, positive predictive value	Original classes highly imbalanced; synthetic augmentation equalized counts, increasing total beats	Beat-level 10-fold CV allows patient overlap and pre-split synthetic beats, potentially inflating metrics	None
[116]	Pretrained CNN fine-tuned on mammography; train/val/test splits used, though patient overlap may exist	Accuracy and ROC-AUC reported, with qualitative claims of competitive performance	Class imbalance between benign and malignant masses addressed with augmentation and sampling	Pre-split augmentation may cause lesion overlap; patient- and center-level separation unclear, risking inflated performance	None
[18]	Train/validation/test at scan or subject level with grid search; external datasets not used	Accuracy, AD detection rate, false alarm rate; ROC-based metrics	Scan-level splits inflate N but mix subject scans; subject-level splits reflect more realistic performance	Random scan-level split allows subject overlap, causing leakage; accuracy drops with true subject-level split	None; focus on architecture and hyperparameter impacts; no XAI
[67]	Subject-level 10-fold CV with separate MRI and PET 3D-CNNs, followed by FSBi-LSTM on 3D features and comparison of multiple fusion stages	Accuracy, sensitivity, specificity, F1, balanced accuracy, ROC-AUC	Moderate class imbalance (pMCI and sMCI smaller than NC and AD) with per-subject data, many learned features, and all sourced from ADNI	Subject-level folds prevent classic leakage, but pre-CV preprocessing and tuning on the same dataset, without external testing, leave generalization to other sites unverified	Interpretability is limited to ROI masking–based analyses, with no model-agnostic XAI, keeping explanations custom and task-specific.
[69]	Random 85/15 split for train/validation; 10-fold CV within SVM training on features; no independent external validation	For ternary classification: accuracy up to 95.74%; for binary tasks, accuracy, sensitivity, specificity, AUC (e.g., NC vs AD accuracy 99.1, AUC 0.999)	Multiple scans per subject; moderate class imbalance; single-source imaging with uniform acquisition	Preprocessing done pre-split; scan-level random split likely causes subject-level leakage and inflated metrics	None
[158]	Train/validation/test split; 80% for train+validation, 20% test; no k-fold CV; no external cohort	Accuracy, precision, recall, F1-score	Data balanced via augmentation from highly skewed classes, potentially causing overfitting	Pre-split augmentation and image-based splitting risk leakage; no clinical external validation	Yes; MRI decision regions interpreted via saliency maps and Grad-CAM across models
[276]	Single setup using CloudSim on Framingham and Statlog; train/test split lacks patient or temporal details; no external cohorts	Accuracy, precision/recall	Moderate-size, imbalanced CVD datasets; detailed class counts unavailable	Preprocessing and tuning done pre-split; no subject-level separation, risking leakage and inflated accuracy	None
[202]	Single protocol on a large clinical dataset with Synthetic Minority Over-sampling Technique (SMOTE); train/test split only, no k-fold CV or external validation	Accuracy, recall (sensitivity), precision, specificity, F1	SMOTE balances imbalanced data, increasing sample size but altering real-world prevalence	Pre-split SMOTE and CNN feature learning risk test-set leakage and inflated performance; no temporal or center-level separation reported	Limited; highly interpretable results when combined with clinical indicators

Table continues on next page

Table continued from previous page

Key literature	Validation type	Metrics used	Data balance	Data leakage risks	XAI used
[159]	Train–test split for each network; ablation but no k-fold or external validation	Accuracy, loss; separate train/test accuracy; ablation results	class balance not fully detailed; potential sampling bias toward specific clinics	Risk of leakage arises from optimizing scaling, model design, or fusion before splitting, with no explicit assurance of patient-level separation	LIME and SHAP used for visualizations of feature and region contributions
2. Medical Imaging Analysis and Radiology					
[63]	Multiple CV schemes were used, 10-fold, 4-fold, and 2-fold cross-validation, with the best results obtained from 2-fold CV using neural-network fusion over classical classifiers.	Best 2-fold CV achieved high sensitivity, specificity, accuracy, and Dice, with other CV schemes slightly lower	Balanced binary outcome (death vs. recovery) with 200 subjects and single X-ray modality	Subject-level CV minimizes image-level leakage, but tuning on the same 200-case dataset without external validation may inflate performance	No interpretability; method relies on stochastic texture features and shallow neural-network fusion without XAI tools
[99]	Hold-out validation on two public X-ray datasets with augmentations; no external clinical data; compared 16 DL models across multiple DA settings	Class imbalance exists and is modified through augmentation (e.g., CROP+ROT+HF)	InceptionV3+CROP+ROT+HF top accuracy/sens/spec/F1-score	Originally imbalanced and then augmented, with patient counts not reported	Pre-split augmentation can duplicate images across splits, and tuning is limited to public data with no external clinical validation.
[246]	Internal train–validation splitting used, comparing sequential, functional, and transfer learning models with dual Adam/SGD optimization against pretrained networks	acc, sens, spec, prec, F1-score; lower params/compute than baselines	Multi-class imbalance persists; segmentation and augmentation increase sample count but not patient diversity	Pre-split segmentation/augmentation may leak data, and no external cohorts are used beyond public sets	Limited qualitative insights; lacks formal CAM or saliency analysis.
[121]	k-fold CV on public ICH dataset	Acc, sens per subtype (typical)	subtype imbalance persists at the slice level	Slice-level cross-validation permits patient leakage due to the absence of a group-wise split	GradCAM / saliency for bleed localization
[178]	Hold-out + k-fold on NIH/CheXpert; XAI-integrated	Acc, AUC; XAI fidelity high	multi-label imbalanced	Public datasets exhibit label noise and potential multi-label leakage	LIME / SHAP / GradCAM for multi-disease heatmaps
[44]	Train/test split upon public dataset	Acc, sens	Small-moderate balanced	Aug pre-split	CNN attention/GradCAM
3. Multimodal Deep Learning in Healthcare					
[130]	Validated only on a single clinical cohort with no external hospital test set	Accuracy, per-stage sensitivity / specificity,macro-F1	Approximately balanced across stages after grouping	Single-center data with no patient-level external split; one-time preprocessing may favor cohort-specific overfitting	Limited interpretability; offers modality-level cues but lacks systematic CAM/SHAP for joint outputs
[73]	Small-sample dataset validated via k-fold or leave-one-subject-out CV	Accuracy, sensitivity, specificity, ROC-AUC	Very small sample size with frequent class imbalance	High overfitting risk; patch-based training may cause specimen-level leakage without patient-level folds	None
[249]	Patient-level train / val / test split on a single health-system EHR, with occasional k-fold sensitivity analysis	AUROC, AUPRC, micro/macro-F1, calibration	Large retrospective cohort with structured EHR and other modalities; disease labels are highly imbalanced	Encounter-level splitting risks leakage; patient-level splits are reported but temporal separation may be lacking	Limited interpretability; shows attention and feature importance, but lacks standardized SHAP/LIME across tasks
[62]	Temporal train/val/test split with potential internal–external hospital validation	AUROC, AUPRC, Brier score, calibration, decision-curve metrics	Large multi-year hospital cohort with notes, labs, and imaging; long-term outcomes are highly imbalanced	Imperfect temporal cuts or multiple admissions per patient can cause outcome leakage; large model size complicates detection	Prompt- and attention-based explanations with saliency on text/images and occasional SHAP for tabular data

Table continues on next page

Table continued from previous page

Key literature	Validation type	Metrics used	Data balance	Data leakage risks	XAI used
[141]	Internal k-fold or hold-out validation on single or multiple hospital datasets	AUROC, F1, and accuracy reported, with occasional per-modality ablation metrics	Visit-level splits risk leakage; fusion features may be pre-computed on the full cohort	Visit-level splits and precomputed fusion features may risk data leakage	Mostly modality-level interpretability; lacks fine-grained XAI like per-feature SHAP
[155]	Train/val/test split on hospital cohort, occasionally patient-level k-fold	AUROC, accuracy, sensitivity, specificity	Single- or few-center dataset with mammograms, prescriptions, and labs; moderate size with class imbalance	Multiple exams per patient or pre-split feature aggregation may cause leakage	Limited feature-level interpretation and occasional CAM; no unified XAI framework

6 Future Research Directions

Addressing the identified limitations requires continued research efforts and technological advancements, which are discussed in the following future research directions. Future research in healthcare plays a vital role in advancing medical knowledge and enhancing patient care. With the increasing incorporation of technology into healthcare systems, there is a pressing need to investigate novel methodologies and tools aimed at improving diagnostic precision, therapeutic effectiveness, and patient involvement. Emphasizing these areas can help overcome existing challenges and contribute to the development of more efficient and accessible healthcare solutions.

6.1 Future Research Direction for Disease Prediction and Early Diagnosis

The application of ML and DL models has demonstrated substantial potential in improving the accuracy and promptness of disease detection; however, these approaches still encounter various challenges and limitations. As the healthcare domain continues to evolve, future research is anticipated to explore advanced methodologies, with a detailed overview of key directions provided in the following section.
Data preprocessing represents a fundamental phase in data analysis and ML, serving to transform raw data into a suitable format for subsequent analysis. This process is particularly vital for enhancing model performance and ensuring data quality, especially when working with complex datasets such as those derived from medical imaging or genomic studies. Future advancements in data preprocessing are expected to focus on improving standardization protocols [69], incorporating sophisticated techniques capable of managing increasingly large and heterogeneous datasets [18].

- *Standardization:* Standardizing MRI data is vital for ensuring voxel-wise anatomical consistency across images [69]. This includes spatial normalization using tools like SPM12, along with essential preprocessing steps such as skull stripping and grey/white matter segmentation to improve data quality [69].
- *Incorporation of sophisticated methodologies:* Advanced preprocessing methods, including tools like FreeSurfer for MRI scans [187] and AI-based approaches such as CNNs for data preprocessing [6], have demonstrated potential in enhancing model performance and data preparation efficiency, suggesting promising avenues for further improvement.

The quality and complexity of data are pivotal in shaping the development and effectiveness of DL models. Maintaining high data quality necessitates the resolution of issues such as missing values, outliers, and inconsistencies, all of which can adversely affect model accuracy and reliability. Increasing data complexity introduces further challenges, demanding innovative strategies to sustain optimal model performance. Future directions in this domain include data quality enhancement [244, 223], effectively managing data complexity, and fostering methodological innovations [159].

- *Data quality enhancement:* Data augmentation techniques play a pivotal role in mitigating the limitations of datasets with insufficient images, which can distort model performance. By artificially expanding the dataset, these methods help achieve balance and improve image quality, ultimately enhancing model accuracy [158]. Preprocessing methods such as normalization and data cleaning are essential for improving data quality, ensuring that DL models receive accurate and relevant input, thereby boosting their reliability

Table 18: Quantitative aggregation of results using Mean and sample Variance

Method Type	Task Type	Method	Dataset	Mean	Variance
Deep Learning [116, 263, 223, 97, 235, 115, 125, 101, 125, 47]	Cancer diagnosis (Breast, Lung, Blood, Skin, Oral, etc.)	CNN (ResNet, AlexNet, Inception, VGG, EfficientNet, etc.)	DDSM, MIAS, Kaggle SLC, ISIC archive dataset, C-NMC, etc.	0.908	114.28
Deep Learning [187, 18, 67, 158, 159, 184, 137, 257, 141], Hybrid Model [69]	Alzheimer's disease diagnosis	CNN (VGG16, ResNet, Inception, VGG19, DenseNet169, DenseNet201 etc.) , RNN, CNN-SVM	ISIC, ADNI, Kaggle etc.	0.864	127.53
Deep Learning [3, 176, 276, 47]	Heart disease	CNN , RNN (Bi-LSTM , Bi-GRU etc.)	Sutter PAMF,PhysioBank MIT-BIH arrhythmia database, UCI, NIH Biospecimen Information Coordination Center and Data Warehouse, Kaggle	0.966	10.80
Deep Learning [6, 190, 63, 178, 222, 99, 246, 130]	Lung disease (Nodules, tuberculosis, pneumonia, COVID-19, pulmonary edema, COPD)	CNN (VGG16, ResNet50, InceptionV3, DenseNet121 etc.)	Kaggle, SLC, Cohen's COVID-19 data, CORD-19, University of Louisville and Man soura University, COVID-CT, COVID NET, TCVGH etc.	0.958	5.45
Deep Learning [244]	Diabetes	DNN	Frankfurt hospital, Pabna diabetes hospital, PIMA Indian diabetes dataset	0.963	5.57
Deep Learning [287, 143], Hybrid MMDL [204]	Eye disease (glaucoma, AMD, PM, DR, cataracts etc.)	CNN (ResNet), GAN, attention mechanisms	DRISHTI-GS, RIM ONE v3, Ichallenge AMD, Ichallenge PM, EyePACS, Kaggle, Fundus FFA etc.	0.933	19.75

and performance[244]. Ethical consideration in data quality requires ensuring unbiased model outcomes. Future research should explore dataset biases with guidance from ethics and healthcare experts to enhance ethical standards [223]. The integration of real-time data from wearable health-monitoring devices further enhances data quality by reducing dependence on potentially inaccurate user-entered data, thus increasing system reliability, particularly for users less experienced with tracking or inputting health information [159].

- *Managing data complexity:* Enhancing the robustness of DL models requires expanding dataset size and diversity, particularly by incorporating varied demographics and disease stages to ensure broader generalizability [223]. Future efforts should prioritize inclusive data collection across age, ethnicity, and geography to strengthen real-world applicability [159]. Moreover, optimizing hyperparameters through advanced techniques can significantly improve performance on complex datasets [223]. Dimensionality reduction methods, such as PCA, aid in managing high-dimensional data by isolating the most informative features [244]. The adoption of advanced model architectures, including deep and Convolutional Neural Networks, enables effective extraction of hierarchical features from complex data [113, 202]. Moreover, ensuring scalability and efficiency in models—through lightweight classifiers like LightGBM and XGBoost—remains a key research direction for handling large-scale datasets without compromising accuracy [202].

- *Incorporation of advanced technologies:* The integration of wearable health-monitoring devices enhances data quality and system reliability by providing real-time, automated health metrics, minimizing dependence on potentially inaccurate user input [159]. Moreover, adapting models for varied environments through preprocessing techniques—such as histogram equalization and brightness normalization—can improve predictive performance across diverse imaging conditions [159].

With the growing reliance on data-driven healthcare, ensuring the availability of high-quality data is vital. This requires overcoming challenges in data collection, storage, and sharing, while upholding ethical and privacy standards. The following sections outline key future research directions concerning data availability in healthcare.

- *Augmenting data and ensuring class balance:* Data augmentation is essential for improving the size and balance of limited datasets, enhancing model performance through techniques such as flipping, rotation,

and zooming [158]. Future research should aim to optimize these methods to better represent minority classes, thereby reducing bias and increasing model robustness [158].

- *Multi-source data integration:* Integrating multi-modal data—such as genetic, demographic, and longitudinal information—can enhance diagnostic accuracy and early detection [158]. Additionally, employing advanced feature extraction techniques, including modern computer vision methods, may improve model performance by capturing complex data patterns and preserving structural details [67].
- *Resource-efficient model adaptation:* Optimizing models for resource-constrained environments involves adapting them for low-power devices through techniques like quantization and pruning. Additionally, enabling offline functionality can enhance accessibility in regions with limited internet connectivity. [159].

Deep Learning models typically depend on extensive and diverse datasets for optimal performance. However, in fields like medical imaging and diagnostics, acquiring such data is challenging due to privacy constraints and the requirement for expert annotations. Future research is directed toward addressing these limitations related to data scarcity and lack of diversity.

- *Expanding dataset scale and heterogeneity:* Data augmentation enhances dataset size and balance by creating altered versions of existing images, helping improve model accuracy, especially for underrepresented classes [202, 18, 158].
- *Pretrained model utilization:* Hyperparameter optimization [223] leverages pretrained models on large datasets and adapts them to smaller [159], domain-specific data, improving model performance in data-limited settings. Investigating various pretrained models and their combinations offers potential for enhancing diagnostic accuracy and represents a valuable direction for future research [158].
- *Ensemble Learning:* Employing Ensemble methods enhances model accuracy by integrating the strengths of multiple DL models, effectively reducing overfitting in data-limited contexts [187].

Data imbalance occurs when some classes are significantly underrepresented within a dataset, which can result in biased and unreliable model predictions. Effectively addressing this issue is essential for developing robust and accurate predictive models. Future directions in this domain include the implementation of synthetic data generation approaches [244], adoption of resampling methods [244, 159], and design of cost-sensitive learning strategies [18, 202].

- *Implementation of synthetic data generation approaches:* The Implementation of synthetic data generation using models like GANs offers a promising approach to addressing class imbalance by augmenting minority classes with realistic, data-driven samples[244].
- *Adoption of resampling methods:* Future research may consider applying resampling techniques, such as oversampling minority classes or undersampling majority ones, to achieve class balance, with methods like SMOTE proving effective in improving model training and performance [244, 159].
- *Design of cost-sensitive learning strategies:* Exploring cost-sensitive learning methods that assign varying misclassification costs can enhance model focus on minority classes, improving prediction sensitivity and precision in imbalanced datasets [18, 202].

Methodological challenges, including model complexity, interpretability, feature extraction, and the integration of diverse data modalities, must be addressed to enhance the effectiveness and applicability of these models. The following sections highlight key future research directions related to these constraints in healthcare.

- *Improving model interpretability:* Future research should aim to enhance XAI methods to improve the transparency and interpretability of DL models, thereby fostering trust and supporting their adoption in clinical practice [158, 223].
- *High-level feature engineering techniques:* Incorporating multi-modal data and optimizing feature integration can enhance diagnostic accuracy, highlighting the need for improved dynamic feature weighting in Deep Learning models[158, 159].
- *Unsupervised Feature Learning:* Unsupervised Deep Learning enables representative feature extraction without reliance on ground truth, benefiting scenarios with unreliable labels [187].

- *Data fusion techniques:* Integrating imaging, genetic, and clinical data enhances predictive accuracy. Future work should explore end-to-end models combining autoencoders with advanced integration methods beyond simple concatenation, enabling multi-level data fusion to improve disease understanding and outcome prediction [159].

- *Strengthening Frameworks for liability and accountability:* In order to facilitate the safe, ethical, and trustful incorporation of AI into Healthcare 5.0, future research should create elaborate liability and accountability schemes specific to AI-facilitating clinical settings. This involves the definition of the sharing of responsibility among the developers, healthcare providers, and institutions especially on high stakes decisions. Transparency and incident attribution can be ensured with help of innovative practice, including robust audit trails, explainable AI models, and standardized reporting mechanisms. The studies should also address technical methods of improving human supervision, reducing automation bias, and incorporation of fail-safe measures into AI processes. To harmonize the liability regulations worldwide, cross-jurisdictional studies are required to cover the differences between the regions in terms of legal frameworks and policies on healthcare. Also, longitudinal research assessing effects of liability on AI adoption, clinical outcomes and patient trust will be used to develop balanced policies that are up to the task of protecting patients without stifling innovation. Such a multi-disciplinary collaboration between legal, technical, and clinical expertise is needed to support the safe and fair implementation of AI in Healthcare 5.0[253, 254].

DL models in healthcare face challenges like high computational demands, limited real-time capability, scalability, and demographic generalizability. The following sections outline key future research directions to address these issues.

- *Significant resource demands and parameter optimization:* Complex models, like 3D-CNNs, demand high computational resources, limiting deployment in resource-constrained settings [67]. Lightweight models, such as MobileNet, offer reduced resource requirements but may sacrifice accuracy [159]. Models like MOD-2D-CNN balance efficiency and performance [159]. Minimizing the number of trainable parameters improves convergence speed and minimizes memory usage for practical applications [67].

- *Real-time model optimization:* Optimizing models for low-power or mobile devices is essential for real-time use, particularly in low-connectivity environments. Techniques like model quantization and pruning reduce computational overhead, improving usability in such scenarios[159].

  Within the low-resource or developing healthcare system, deep learning methods can greatly increase the diagnostic and decision-support functions despite the infrastructural constraints [268, 282]. Lightweight and optimized models may be deployed on a mobile or edge device, allowing prediction of disease and analysis of an image without accessing high-end computational capabilities . Cloud-based and edge computing platforms will be used to evaluate remote data and deploy models to enhance access to diagnostic services in underserved regions. Furthermore, transfer learning enables the application of the existing models to the medical data of a region, which reduces the scope of the datasets and training data requirements, e.g. brain tumour (glioma) segmentation models based on deep learning have been trained on large datasets and fine-tuned to operate on lower quality or even limited MRI images in Sub-Saharan Africa, allowing diagnostic results where other radiology facilities might be limited [191]. All of these methods contribute to making deep learning a feasible and scalable solution to enhance the quality and accessibility of healthcare in resource-constrained settings [96].

- *Cross-platform scalability:* Integrating AI with portable imaging devices offers scalable diagnostic solutions, expanding healthcare access [7]. Ensuring model efficiency across diverse hardware is crucial for widespread adoption [159].

- *Comprehensive data acquisition:* Expanding datasets to cover diverse demographics is vital for improving model generalizability and accuracy [159]. Addressing biases and ensuring representation of underrepresented groups enhances model reliability in real-world applications [7].

Clinical validation is imperative to ascertain that predictive models demonstrate consistent performance and reliability within real-world healthcare settings. But such models' clinical adoption requires addressing ethical and validation challenges. The following sections outline future research directions to tackle these issues.

- *Ethical and equitable deployment:* Ensuring fairness is vital to mitigate bias and prevent discriminatory outcomes in healthcare [223]. Enhancing model transparency through explainable AI is key to clinical trust and ethical use [244]. Additionally, adapting models to varied global healthcare environments is necessary to support equitable access and practical deployment [223].
- *Improved protocols for clinical validation:* Clinical validation and integration into healthcare workflows are crucial to ensure the safety, efficacy, and practical utility of such models in real-world medical settings. Meaningful collaboration with healthcare professionals is essential to support the design and execution of clinical studies that assess the safety, efficacy, and practical utility of these models in routine medical practice [223].
- *Ensuring the privacy and integrity of patient data:* Maintaining patient confidentiality and regulatory compliance necessitates the implementation of robust data protection strategies, encompassing secure data management and transparent usage policies to foster patient trust [223].
- *Role of clinical trials and prospective validation:* The need to bring deep learning models to actual Healthcare 5.0 systems implies that such models should be clinically validated. A variety of potential studies and clinical trials have revealed the performance of AI models deployed in the conditions of normal clinical practice. As an example, the IDx-DR system, used to detect diabetic retinopathy and cleared by the FDA, was put in a prospective, multisite clinical trial and demonstrated autonomous diagnostic performance in the primary care setting [2]. On the same note, deep learning stroke detection systems like Viz.ai LVO system have been tested in clinical practice and demonstrated to save time to treatment. The diabetic retinopathy deep learning model developed by Google was also confirmed in the large-scale prospective implementation trial at Thai hospitals and demonstrated high generalizability but also indicated operative difficulties [25]. Future studies in the critical care, including the evaluation of deep learning models against early sepsis detection, state that real-time clinical outcomes can vary compared to the retrospective ones owing to data drift and patient heterogeneity. The critical role of clinical evidence, workflow integration, and ongoing observation, as highlighted in these studies, is to make sure that deep learning systems are reliable, safe, and effective in the world of Healthcare 5.0 settings.
- *Involvement of key stakeholders:* Engaging diverse stakeholders—such as patients, clinicians, and ethicists—in model development promotes alignment with practical needs and ethical standards [223].
- *Approaches addressing real-time constraints:* Recent evidence of edge intelligence in the healthcare domain demonstrates that distributed computing to edge or fog devices near sensors can potentially save a lot of latency and bandwidth. Techniques such as model compression, pruning, quantization and lightweight architectures (e.g., MobileNet, EfficientNet-lite, compact CNNs) are being implemented to execute deep models on low-powered hardware at sufficiently diagnostic accuracy and response times suitable to support ongoing monitoring and anomaly detection [98, 243]. Research on arrhythmia detection and sepsis prediction also shows that patient safety can be affected by delays or unstable inference pipelines, further explaining the importance of streamlined architecture and well-developed real-time data processing infrastructures [94, 88]. Frameworks based on edge-cloud collaboration and architecture with 5G/URRLC are also suggested to ensure the trade of accuracy and real-time guarantees to applications such as remote vital-sign prediction and real-time health surveillance [15, 239, 200].

6.2 Future Research Direction for Medical Imaging Analysis and Radiology

The future of medical image analysis and radiology is set for substantial progress propelled by innovative technologies and techniques. Primary focal points are the incorporation of AI, DL, and radiomics, which are expected to improve diagnostic precision and treatment customization. The subsequent sections delineate the principal research trajectories expected in this domain. Data limitations in medical imaging—such as scarcity, variability, and artifacts—pose significant challenges but can be mitigated through innovative approaches that improve the effectiveness of DL models.

- *Integration of public datasets and clinical partnerships:* The use of open-source datasets offers a practical solution to data scarcity by supporting model training in data-limited domains. Additionally, collaborations with healthcare institutions enable access to diverse, well-annotated datasets, essential for developing robust models [246].

- *Data expansion strategies:* Data augmentation methods—such as rotation, scaling, and flipping—enhance the variability of training datasets, thereby improving model performance under limited data conditions [246].
- *Preprocessing strategies and attention mechanisms:* Applying preprocessing techniques like normalization and augmentation enhances data quality and model performance by standardizing input characteristics. Additionally, integrating attention mechanisms into CNNs improves both interpretability and accuracy by enabling the model to prioritize the most relevant image regions, effectively addressing dataset heterogeneity [121].
- *Advanced segmentation strategies:* Advanced segmentation techniques like Mask R-CNN improve classification accuracy by isolating relevant features and reducing the influence of artifacts in medical images [246].
- *Quality control protocols:* Applying quality control during data acquisition and preprocessing ensures high-quality input data, essential for reliable performance [121].

DL has significantly advanced diagnostic accuracy in medical imaging, yet it faces key challenges such as limited interpretability, algorithmic inefficiencies, classification errors, and generalizability issues. Subsequent investigations are focused on mitigating these limitations.

- *AI interpretability frameworks:* Explainable AI frameworks are essential in healthcare to enhance the transparency and trustworthiness of DL models. It needs to be transitioned away in passive visualization to decision-oriented explanations that project model signals into clinical reasoning. Future work should include: (a) creating clinically aligned explanations (e.g., connecting heatmaps or token attributions to diagnostic criteria), (b) integrating uncertainty estimates with explanation (highlight low-confidence cases), and (c) integrate the output of the explanations into the workflow used by clinicians (audit logs, justification notes, counterfactual suggestions). Model explanations that are verified by the domain experts can increase the levels of clinician trust, allow analysis of errors more precisely, and increase the speed of regulatory acceptance-enhancing both the accuracy of early prediction and diagnostic safety [207, 12, 178].
- *Predictive output transparency:* Ensuring the interpretability of model predictions is essential for fostering trust among healthcare professionals. This requires enhancing the transparency of the AI model's decision-making processes to support informed clinical adoption [246].
- *Algorithmic optimization techniques:* Employing dual optimization algorithms, such as Adam and stochastic gradient descent (SGD), can enhance model performance by effectively reducing training parameters and computational overhead [246].
- *Model Evaluation and Assessment:* Thorough testing and assessment of models with varied datasets are crucial to reduce incorrect classification results and enhance diagnostic precision [121].
- *Integration of multiple imaging modalities:* The integration of multiple imaging modalities, such as CT scans and X-rays, can augment the diagnostic effectiveness of AI models by offering a more holistic representation of medical conditions [246].
- *Equitable model performance:* Ensuring fairness and mitigating bias in algorithmic design is fundamental to the ethical implementation of AI in healthcare. This requires training models on diverse and representative datasets to prevent discriminatory or skewed outcomes [178].
- *Generalizability assessment:* External validation in real-world clinical contexts is essential to confirm the generalizability of AI models across varied populations and healthcare settings [121].

DL models in healthcare encounter challenges in real-world clinical integration, including radiologist workload and complexities in manual interpretation. The subsequent sections highlight key future research directions aimed at addressing these concerns.

- *Automated diagnostic imaging systems:* DL models offer automated medical image screening, effectively reducing radiologists' workload by accurately detecting and classifying chronic diseases such as lung diseases [246]. This enables radiologists to concentrate on complex cases, thereby minimizing fatigue-related diagnostic errors [131].

- *Enhanced diagnostic precision:* AI systems, when trained on extensive datasets, can achieve diagnostic performance comparable to or surpassing that of expert radiologists. For example, by delivering objective, quantitative metrics—such as the extent of lung involvement—they promote greater consistency and reliability in clinical decision-making [131].

- *Advanced medical image interpretation:* Different DL algorithms, particularly CNNs, have significantly advanced medical image analysis by extracting hierarchical features that enhance diagnostic accuracy. They enable automated detection of conditions such as cerebral hemorrhages, thereby reducing manual interpretation and lowering the risk of human error [121].

- *Integration into clinical workflows and model refinement:* Workflow integration is a crucial prerequisite of DL in Healthcare 5.0, but it is one of the least developed fields of current research. In practice, a large number of AI systems have a difficult time in clinical environments due to lack of alignment of their output with clinical information flows, their inability to integrate with hospital IT ecosystems, and the fact that they add cognitive and operational overhead to final users. Thus, despite some models demonstrating good performance in the experiment, they fail to serve the purpose of clinical impact [12].

  Future research should put in place technically strict frameworks of integrating deep learning tools into end to end clinical processes. Among the requirements are standard data interfaces, compatibility with EHR, Picture Archiving and Communication System (PACS), Laboratory Information System (LIS), and Fast Healthcare Interoperability Resources (FHIR)-based systems as well as mechanisms that enable constant ingestion of multimodal data provided by IoT devices, wearables, and remote monitoring platforms. The microservice-based architecture, API-based orchestration, and ontology-based knowledge graph can be used to overcome interoperability barriers and provide data exchange among different healthcare systems on a semantically consistent basis [167, 229]. The modeling of the dynamic workflow behavior is another significant direction. AI tools can be programmed to adapt their suggestions depending on the patient acuity, workload of the staff, and changes in resource availability using adaptive workflow engines, multi-agent coordination systems, digital twins, and context-aware decision-support modules. In the real world, reinforcement learning can also help optimize triage, task allocation and time-sensitive clinical processes. There should also be the issues of scalability and robustness. Federated learning, decentralized integration pipelines, and privacy preserving computation will play a crucial role in the implementation of AI systems in multiple institutions without violation of security and regulatory requirements. Longitudinal evaluation systems, which will include drift detection, real-time performance checks, and feedback mechanisms, are required to make deployed models stable and safe in the course of their life [194] The AI systems should also be incorporated in the clinical workflows in a manner that assists in real-time diagnostics and decision-making. This involves enhancing the interpretability of the model, its reliability, and operational transparency to allow the utilization of outputs at the point of care. On-going updating of models with different datasets and taking into consideration the real-life use feedbacks will be crucial to the preservation of accuracy of the diagnostic in different clinical environments. To make sure that deep learning applications support clinical priorities and in fact decrease cognitive load, cooperation among developers, workflow engineers, and healthcare professionals is paramount to achieve this goal, as well as to make sure that the tools are aligned with clinical priorities [97].

- *Advancing accuracy in clinical diagnosis:* AI systems can be embedded into clinical workflows to support real-time diagnostics, thereby improving care quality. To ensure effective clinical adoption, future research should prioritize enhancing model interpretability and reliability. Continuous training with diverse datasets and close collaboration between developers and healthcare professionals are vital for maintaining diagnostic accuracy and aligning AI tools with clinical needs[97, 121].

- *Minimizing diagnostic inaccuracies:* DL enhances CADx by minimizing errors linked to manual interpretation. Models like the SDAE outperform traditional methods by autonomously extracting and leveraging meaningful features from medical images[47].

- *Optimizing clinical efficiency:* AI-powered diagnostic tools can improve hospital efficiency by decreasing clinicians' workload and enabling greater focus on patient care, especially in rural settings with limited medical personnel [246].

- *Physician acceptance and trust:* The acceptance of deep learning in Healthcare 5.0 depends on the acceptance of physicians. The clinicians tend to be reluctant to use AI systems when the reasoning of a model is not clear and where the outputs fail to respond to the clinical judgment. Trust can be enhanced by ensuring systems have understandable explanations, uncertainty estimates, and actionable insights as opposed to opaque predictions[39]. Human-centered design needs to be considered in the future, with the clinicians involved in the model development, consideration of the explanation form, and cognitive workflow requirements. Further research is also required to quantify the interaction of clinicians with AI tools in practice, such as overriding tendencies and responsibility or automation bias issues [53]. Enhancement of these trust issues will be critical to any meaningful use of deep learning in daily clinical practice.

- *Real world deployment:* A major translational gap in deep learning research for Healthcare 5.0 is the relative scarcity of real-world, large-scale deployments that extend beyond controlled academic prototypes. However this trend is changing very fast. As an illustration, Advocate Health has implemented more than 40 AI solutions, which are ready to production, such as Microsoft Dragon Copilot in ambient documentation and imaging skills like Aidoc and Rad AI, within the scope of its clinical enterprise. Kaiser Permanente implemented Abridge's generative documentation system in more than 40 hospitals and 600+ medical offices, the largest generative AI deployment in clinical operations to date, demonstrating that scalable, field-tested tools can dramatically reduce documentation burden while integrating with EHR ecosystems. Predictive sepsis and respiratory failure models, digital twin models, and closed-loop ventilator management have been tested and implemented in various ICUs around the world using frameworks such as the Learn-Predict-Monitor-Detect-Correct (LPMDC) architecture, which have demonstrated significant reductions in mortality, ICU length-of-stay and cognitively load on clinicians. Also, existing medical imaging triage AI systems like Aidoc are already used in the US and European hospitals on a regular basis to rank radiology workflows and identify critical results to be reviewed faster [36]. Nevertheless, the uptake of deep learning systems is projected to be very low (estimated at 22% of health organizations around the world) due to the complexity of integration, compliance mandated by the government, and workforce adaptation issues. Therefore, the future development is to create more robust deployment models, scalable interoperability models, and large-scale longitudinal studies that can help close the ongoing research-to-practice gap based on theoretical experimentation and sustainable and pragmatic deployment in various and complicated healthcare settings [279].

6.3 Future research direction for Multimodal Deep Learning in Healthcare

The field of MMDL in healthcare will be a topic of future research to resolve the existing limitations and widen its functions to produce exact, steady, and clinical combination diagnostic schemes. The most important areas involve model optimization on large and diverse datasets, adding new types of data, easing model explainability, and potentially exploiting novel training approaches such as federated learning to alleviate data privacy and sharing issues.

6.3.1 Data Limitations Solutions

- *Expansion and diversification of datasets:* An initial issue to address is the relatively small sample size of many medical datasets, e.g., the ADNI dataset to study AD, and limited purely publicly accessible THz imaging datasets to study cancer. The future work will be focused on testing models on increased, ultra-rich, diverse datasets to enhance generalizability and resiliency [257, 73].

- *Approaches to missing data:* The potential problem with missing data can be observed in clinical data, especially data related to a certain test, such as the gene mutation status. It is planned that more advanced imputation techniques will be investigated in the future to deal with such an unavoidable issue [141].

- *Normalization:* Normalization of clinical data entails rescaling numerical variables to a defined range and transforming categorical variables into binary representations, commonly through approaches such as one-hot encoding [257].

- *Alternative training datasets:* Future work could find and utilize subsets of already well-known and complementary modalities (such as IR imaging for THz imaging) as an alternative training dataset to address the problem of scarcity of training data with newer and more unproven imaging schemes [73].

- *Optimization on varying datasets:* Future work will be to optimize different models, such as DenseNet, over larger and more diverse data to increase their power to predict [134].

6.3.2 Model Architecture and Performance Improvement

- *Improved fusion techniques:* The early and late fusion techniques are advanced techniques and oversimplistic. The future research will be aimed at more sophisticated fusion strategies, including models utilizing attention capabilities and graph neural networks, which describe complex relationships between the types of data in a better way [204].
- *Fusion of new modalities:* Multimodal data can be integrated to get an improved result. E.g., an investigation can be undertaken into the integration of additional data sources, such as cognitive assessments and genetic markers, to develop more robust and effective predictive models [134].
- *End-to-end training:* A future direction is to support end-to-end training of multimodal models: integrating feature extraction steps (such as auto-encoders) and integration and classification, instead of training them separately [257].

6.3.3 Clinical Integration and Explainability

- *Increased explainability:* One of the largest impediments to clinical uptake is the *black box* nature of most DL models. Future work will focus on the enhancement of explainability, with methods such as Grad-CAM and SHAP to make it more amenable to clinician trust and practices [204].
- *Advancing clinical DL through emerging methods:* Emerging methods like foundation models, large multimodal AI models, and federated learning offer tangible ways to address major drawbacks of the existing clinical deep learning. Large-scale medical and general-domain trained foundation models can offer powerful generalization, can support low-resource tasks, and lower the need for task-specific labeled datasets [173]. Large multimodal models combine imaging, text, genomics and signals into a unified architecture to allow clinical reasoning in a more holistic manner and to decrease errors from single-modality blind spot [261]. The federated learning will play a critical role in overcoming the issue of data security and privacy. This can enable training models on decentralized data across multiple institutions, without sharing raw patient data, and so provide access to larger, more diverse datasets to train high-precision, multimodal diagnostic systems [119, 144, 204]. All these new approaches offer tangible and practical improvements that can enhance generalizability, decrease bias, reinforce multimodal integration, and render clinical realization.
- *Increased scope:* Models can be set to the behavior of unusual or complex cases and can be scaled to be able to diagnose more diseases, and hence that increases their clinical value[204].
- *Regulatory and compliance requirements* Companies that regulate the field have started to codify the specifications of AI-powered medical tools. The U.S. Food and Drug Administration (FDA) [28] has published frameworks including the Software as a Medical Device (SaMD) guidelines and the regulatory approach proposal of AI/ML-based adaptive algorithms which focus on transparency, real-world performance monitoring, and risk control. Likewise, the European Medicines Agency (EMA) offers a guideline in its AI reflection paper, which identifies data quality, model validation, explainability and patient safety requirements [124, 267]. These frameworks facilitate the standardization of the development and implementation of DL applications to make emerging Healthcare 5.0 systems to address the clinical, ethical, and safety expectations.

7 Conclusions

The integration of DL into smart healthcare systems has opened new frontiers in disease prediction, diagnostics, medical image analysis, and radiology. This survey systematically addresses the research questions outlined at the beginning. In relation to RQ1, our review has demonstrated that DL serves as a cornerstone for advancing Healthcare 5.0 by enabling automation, personalization, and precision, thereby surpassing the capabilities of traditional ML approaches, as discussed in the introduction. For RQ2, we have presented a detailed examination

of the most widely utilized DL algorithms, emphasizing their architectures, strengths, and clinical suitability, which are elaborated in the background section. With respect to RQ3, we have synthesized findings from approximately 20 studies in each of the three primary application domains, including disease prediction and early diagnosis, medical imaging and radiology, and MMDL. These studies are analyzed regarding their methodologies, outcomes, and clinical implications. Our survey further connects diseases, predictive models, accuracy levels, and dataset characteristics, demonstrating how factors such as dataset size, modality, and diversity affect model performance and generalizability, as reported in the DL in healthcare 5.0 section. Addressing RQ4, we have identified major limitations, including a lack of interpretability, scarcity of annotated data, heterogeneity of data sources, high computational requirements, and ethical/privacy concerns, as outlined in the challenges and limitations section. In response to RQ5, we have highlighted promising research directions such as the development of XAI frameworks, integration of multimodal data, privacy-preserving learning techniques, and lightweight architectures for real-time clinical applications, as presented in the future research directions section. Responding to RQ6, we found that the majority of studies are based on the standard validation methods and performance measurement procedures, yet they do not consider the external validation, data leakage, bias evaluation, and reproducibility, and standardized assessment protocols and open reporting practices are necessary. Lastly, to answer RQ7, we highlighted the role of interpretability, fairness, and regulatory compliance with the XAI approaches, bias-conscious assessment, and compliance with regulatory standards being essential to fostering clinical trust and enabling the safe use of deep learning systems.

This study systematically reviews deep learning methodologies in healthcare, highlighting how innovations such as generative models and attention-based architectures are being applied to tackle complex medical challenges. Our bibliometric analysis further highlighted the intellectual landscape of this interdisciplinary field, underscoring emerging research hotspots and collaborative trends. Despite the remarkable progress, significant hurdles persist, including data scarcity, ethical concerns, lack of interpretability, and the difficulty in deploying models in real-world clinical settings. Addressing these challenges requires a balanced focus on algorithmic innovation, regulatory frameworks, and cross-disciplinary collaboration. The review provides a highly empirical and objectively distinguished viewpoint through the comprehensive comparative analysis of DL studies on Healthcare 4.0 and 5.0 and combining reported findings. It further points out a lack of rigor in evaluation and pinpoints interpretability, fairness, and regulatory compliance as the needs of reliable and deployable Healthcare 5.0 systems. As the healthcare sector increasingly embraces AI-driven solutions, DL stands at the core of this transformation, holding great promise for building resilient, accessible, and personalized healthcare systems for the future.

Declarations

Ethical Approval:

No ethical approval is required.

Consent to Participate:

We consent to participate in this Study.

Consent to Publish:

We authorize the use of my data for research publication purposes.

Funding:

There is no funding involved for this work.

Author's Contribution:

Paramita Kundu Maji, Afifa Sadiq, and Sanjay Chakraborty: Writing, review & editing, writing original draft, visualization, validation, methodology, formal analysis, data curation, and conceptualization. **Sanjay Chakraborty, Krishnendu Ghosh**: Supervision, formal analysis, conceptualization, writing, review

& editing, visualization, validation, and project administration. **Saikat Basu:** Writing, review & editing, visualization, validation, and formal analysis.

Competing Interests:

There are no competing interests.

Bibliography

[1] Abdullah A Abdullah, Masoud M Hassan, and Yaseen T Mustafa. A review on bayesian deep learning in healthcare: Applications and challenges. *IEEe Access*, 10:36538–36562, 2022.

[2] Michael D Abràmoff, Philip T Lavin, Michele Birch, Nilay Shah, and James C Folk. Pivotal trial of an autonomous ai-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ digital medicine*, 1(1):39, 2018.

[3] U Rajendra Acharya, Shu Lih Oh, Yuki Hagiwara, Jen Hong Tan, Muhammad Adam, Arkadiusz Gertych, and Ru San Tan. A deep convolutional neural network model to classify heartbeats. *Computers in biology and medicine*, 89:389–396, 2017.

[4] Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3529–3530, 2020.

[5] Mugahed A Al-Antari, Seung-Moo Han, and Tae-Seong Kim. Evaluation of deep learning detection and classification towards computer-aided diagnosis of breast lesions in digital x-ray mammograms. *Computer methods and programs in biomedicine*, 196:105584, 2020.

[6] Moutaz Alazab, Albara Awajan, Abdelwadood Mesleh, and Salah Alhyari. Covid-19 prediction and detection using deep learning. *International Journal of Computer Information Systems and Industrial Management Applications*, 12:14–14, 2020.

[7] Bader Aldughayfiq, Farzeen Ashfaq, NZ Jhanjhi, and Mamoona Humayun. Explainable ai for retinoblastoma diagnosis: interpreting deep learning models with lime and shap. *Diagnostics*, 13(11):1932, 2023.

[8] Hazrat Ali, Farida Mohsen, and Zubair Shah. Improving diagnosis and prognosis of lung cancer using vision transformers: a scoping review. *BMC Medical Imaging*, 23(1):129, 2023.

[9] Hazrat Ali, Zubair Shah, Tanvir Alam, Priyantha Wijayatunga, and Eyad Elyan. Recent advances in multimodal artificial intelligence for disease diagnosis, prognosis, and prevention. *Frontiers in radiology*, 3:1349830, 2024.

[10] Abeer Aljohani and Nawaf Alharbe. Generating synthetic images for healthcare with novel deep pix2pix gan. *Electronics*, 11(21):3470, 2022.

[11] Laith Alzubaidi, Jinglan Zhang, Amjad J Humaidi, Ayad Al-Dujaili, Ye Duan, Omran Al-Shamma, José Santamaría, Mohammed A Fadhel, Muthana Al-Amidie, and Laith Farhan. Review of deep learning: concepts, cnn architectures, challenges, applications, future directions. *Journal of big Data*, 8(1):53, 2021.

[12] Julia Amann, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I Madai, and Precise4Q Consortium. Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, 20:1–9, 2020.

[13] Rashid Amin, Mohammed A Al Ghamdi, Sultan H Almotiri, Meshrif Alruily, et al. Healthcare techniques through deep learning: issues, challenges and opportunities. *IEEE Access*, 9:98523–98541, 2021.

[14] Syed Muhammad Anwar, Muhammad Majid, Adnan Qayyum, Muhammad Awais, Majdi Alnowami, and Muhammad Khurram Khan. Medical image analysis using convolutional neural networks: a review. *Journal of medical systems*, 42:1–13, 2018.

[15] Narenthirakumar Appavu. Analysing the effect of edge-optimized deep learning models on improving low-powered iot devices real-time object detection. In *2025 9th International Conference on Inventive Systems and Control (ICISC)*, pages 1663–1669. IEEE, 2025.

[16] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, 7(6):756–779, 2023.

[17] Yahia Baashar, Gamal Alkawsi, Hitham Alhussian, Luiz Fernando Capretz, Ayed Alwadain, Ammar Ahmed Alkahtani, and Malek Almomani. Effectiveness of artificial intelligence models for cardiovascular disease prediction: Network meta-analysis. *Computational intelligence and neuroscience*, 2022(1):5849995, 2022.

[18] Karl Bäckström, Mahmood Nazari, Irene Yu-Hua Gu, and Asgeir Store Jakola. An efficient 3d deep convolutional network for alzheimer's disease diagnosis using mr images. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 149–153. IEEE, 2018.

[19] Mohammed Badawy, Nagy Ramadan, and Hesham Ahmed Hefny. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, 10(1):40, 2023.

[20] SM Saiful Islam Badhon, Serdar Bozdag, Mohammad Adibuzzaman, Ana D Cleveland, Junhua Ding, and KSM Tozammel Hossain. Temporal concept tracing: Making deep learning predictions interpretable and actionable for icu acute kidney injury prevention. In *Proceedings of the AAAI Symposium Series*, volume 7, pages 448–455, 2025.

[21] Tian Bai, Shanshan Zhang, Brian L Egleston, and Slobodan Vucetic. Interpretable representation learning for healthcare via capturing disease progression through time. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 43–51, 2018.

[22] Shahab S Band, Atefeh Yarahmadi, Chung-Chian Hsu, Meghdad Biyari, Mehdi Sookhak, Rasoul Ameri, Iman Dehzangi, Anthony Theodore Chronopoulos, and Huey-Wen Liang. Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods. *Informatics in Medicine Unlocked*, 40:101286, 2023.

[23] SH Shabbeer Basha, Shiv Ram Dubey, Viswanath Pulabaigari, and Snehasis Mukherjee. Impact of fully connected layers on performance of convolutional neural networks for image classification. *Neurocomputing*, 378:112–119, 2020.

[24] Arash Bateni, Matthew C Chan, and Ray Eitel-Porter. Ai fairness: from principles to practice. *arXiv preprint arXiv:2207.09833*, 2022.

[25] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–12, 2020.

[26] Fatemeh Behrad and Mohammad Saniee Abadeh. An overview of deep learning methods for multimodal medical data mining. *Expert Systems with Applications*, 200:117006, 2022.

[27] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. Ai fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4–1, 2019.

[28] Stan Benjamens, Pranavsingh Dhunnoo, and Bertalan Meskó. The state of artificial intelligence-based fda-approved medical devices and algorithms: an online database. *NPJ digital medicine*, 3(1):118, 2020.

[29] Markus Bertl, Yngve Lamo, Martin Leucker, Tiziana Margaria, Esfandiar Mohammadi, Suresh Kumar Mukhiya, Ludwig Pechmann, Gunnar Piho, and Fazle Rabbi. Challenges for ai in healthcare systems. In *International Conference on Bridging the Gap between AI and Reality*, pages 165–186. Springer Nature Switzerland Cham, 2023.

[30] S Bhavya and Anitha S Pillai. Prediction models in healthcare using deep learning. In *International Conference on Soft Computing and Pattern Recognition*, pages 195–204. Springer, 2019.

[31] Filippo Maria Bianchi, Enrico Maiorino, Michael C Kampffmeyer, Antonello Rizzi, and Robert Jenssen. Recurrent neural network architectures. In *Recurrent Neural Networks for Short-Term Load Forecasting: An Overview and Comparative Analysis*, pages 23–29. Springer, 2017.

[32] Nadine Bienefeld, Jens Michael Boss, Rahel Lüthy, Dominique Brodbeck, Jan Azzati, Mirco Blaser, Jan Willms, and Emanuela Keller. Solving the explainable ai conundrum by bridging clinicians' needs and developers' goals. *npj digital medicine*, 6(1):94, 2023.

[33] Hazrat Bilal, Yibin Tian, Ahmad Ali, Yar Muhammad, Abid Yahya, Basem Abu Izneid, and Inam Ullah. An intelligent approach for early and accurate predication of cardiac disease using hybrid artificial intelligence techniques. *Bioengineering*, 11(12):1290, 2024.

[34] Maneet Kaur Bohmrah and Harjot Kaur. Advanced hybridization and optimization of dnns for medical imaging: A survey on disease detection techniques. *Artificial Intelligence Review*, 58(4):122, 2025.

[35] Katarzyna Borys, Yasmin Alyssa Schmitt, Meike Nauta, Christin Seifert, Nicole Krämer, Christoph M Friedrich, and Felix Nensa. Explainable ai in medical imaging: An overview for clinical practitioners–beyond saliency-based xai approaches. *European journal of radiology*, 162:110786, 2023.

[36] Hanene Boussi Rahmouni, Nesrine Ben El Hadj Hassine, Mariem Chouchen, Halil İbrahim Ceylan, Raul Ioan Muntean, Nicola Luigi Bragazzi, and Ismail Dergaa. Healthcare 5.0-driven clinical intelligence: The learn-predict-monitor-detect-correct framework for systematic artificial intelligence integration in critical care. In *Healthcare*, volume 13, page 2553. MDPI, 2025.

[37] Umar Ali Bukar, Md Shohel Sayeed, Siti Fatimah Abdul Razak, Sumendra Yogarayan, Oluwatosin Ahmed Amodu, and Raja Azlina Raja Mahmood. A method for analyzing text using vosviewer. *MethodsX*, 11:102339, 2023.

[38] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[39] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1721–1730, 2015.

[40] Isabella Castiglioni, Leonardo Rundo, Marina Codari, Giovanni Di Leo, Christian Salvatore, Matteo Interlenghi, Francesca Gallivanone, Andrea Cozzi, Natascha Claudia D'Amico, and Francesco Sardanelli. Ai applications to medical images: From machine learning to deep learning. *Physica medica*, 83:9–24, 2021.

[41] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. Survey of explainable ai techniques in healthcare. *Sensors*, 23(2):634, 2023.

[42] Chiranjib Chakraborty, Manojit Bhattacharya, Soumen Pal, and Sang-Soo Lee. From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Current Research in Biotechnology*, 7:100164, 2024.

[43] Heang-Ping Chan, Lubomir M Hadjiiski, and Ravi K Samala. Computer-aided diagnosis in the era of deep learning. *Medical physics*, 47(5):e218–e227, 2020.

[44] Arkapravo Chattopadhyay and Mausumi Maitra. Mri-based brain tumour image detection using cnn based deep learning method. *Neuroscience informatics*, 2(4):100060, 2022.

[45] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021.

[46] Tsai-Min Chen, Chih-Han Huang, Edward SC Shih, Yu-Feng Hu, and Ming-Jing Hwang. Detection and classification of cardiac arrhythmias by a challenge-best deep learning neural network model. *Iscience*, 23(3), 2020.

[47] Jie-Zhi Cheng, Dong Ni, Yi-Hong Chou, Jing Qin, Chui-Mei Tiu, Yeun-Chung Chang, Chiun-Sheng Huang, Dinggang Shen, and Chung-Ming Chen. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in ct scans. *Scientific reports*, 6(1):24454, 2016.

[48] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[49] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[50] Ahmet Çinar and Muhammed Yildirim. Detection of tumors on brain mri images using the hybrid convolutional neural network architecture. *Medical hypotheses*, 139:109684, 2020.

[51] Joseph Paul Cohen, Tianshi Cao, Joseph D Viviano, Chin-Wei Huang, Michael Fralick, Marzyeh Ghassemi, Muhammad Mamdani, Russell Greiner, and Yoshua Bengio. Problems in the deployment of machine-learned models in health care. *Cmaj*, 193(35):E1391–E1394, 2021.

[52] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A Bharath. Generative adversarial networks: An overview. *IEEE signal processing magazine*, 35(1):53–65, 2018.

[53] Elliott Crigger, Karen Reinbold, Chelsea Hanson, Audiey Kao, Kathleen Blake, and Mira Irons. Trustworthy augmented intelligence in health care. *Journal of Medical Systems*, 46(2):12, 2022.

[54] Can Cui, Haichun Yang, Yaohong Wang, Shilin Zhao, Zuhayr Asad, Lori A Coburn, Keith T Wilson, Bennett A Landman, and Yuankai Huo. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022001, 2023.

[55] Matteus Vargas Simão da Silva, Rodrigo Reis Arrais, Jhessica Victoria Santos da Silva, Felipe Souza Tânios, Mateus Antonio Chinelatto, Natalia Backhaus Pereira, Renata De Paris, Lucas Cesar Ferreira Domingos, Rodrigo Dória Villaça, Vitor Lopes Fabris, et al. explainable artificial intelligence on medical images: A survey. *arXiv preprint arXiv:2305.07511*, 2023.

[56] Thomas Davenport and Ravi Kalakota. The potential for artificial intelligence in healthcare. *Future healthcare journal*, 6(2):94–98, 2019.

[57] Meha Desai and Manan Shah. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). *Clinical eHealth*, 4:1–11, 2021.

[58] Jun-En Ding, Phan Nguyen Minh Thao, Wen-Chih Peng, Jian-Zhe Wang, Chun-Cheng Chug, Min-Chen Hsieh, Yun-Chien Tseng, Ling Chen, Dongsheng Luo, Chenwei Wu, et al. Large language multimodal models for new-onset type 2 diabetes prediction using five-year cohort electronic health records. *Scientific Reports*, 14(1):20774, 2024.

[59] W Dm. Evaluation: From precision recall and f-factor to roc informedness markedness correlation. *Journal of Machine Learning Technologies*, 2:6, 2011.

[60] Guanliang Dong, Zhangquan Wang, Yourong Chen, Yuliang Sun, Hongbo Song, Liyuan Liu, and Haidong Cui. An efficient segment anything model for the segmentation of medical images. *Scientific Reports*, 14(1):19425, 2024.

[61] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[62] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[63] Mohamed Elsharkawy, Ahmed Sharafeldeen, Fatma Taher, Ahmed Shalaby, Ahmed Soliman, Ali Mahmoud, Mohammed Ghazal, Ashraf Khalil, Norah Saleh Alghamdi, Ahmed Abdel Khalek Abdel Razek, et al. Early assessment of lung function in coronavirus patients using invariant markers from chest x-rays images. *Scientific reports*, 11(1):12095, 2021.

[64] Mohammad Ennab and Hamid Mcheick. Enhancing interpretability and accuracy of ai models in healthcare: a comprehensive review on challenges and future directions. *Frontiers in Robotics and AI*, 11:1444763, 2024.

[65] Mengjie Fang, Zipei Wang, Sitian Pan, Xin Feng, Yunpeng Zhao, Dongzhi Hou, Ling Wu, Xuebin Xie, Xu-Yao Zhang, Jie Tian, et al. Large models in medical imaging: Advances and prospects. *Chinese Medical Journal*, pages 10–1097, 2025.

[66] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[67] Chiyu Feng, Ahmed Elazab, Peng Yang, Tianfu Wang, Feng Zhou, Huoyou Hu, Xiaohua Xiao, and Baiying Lei. Deep learning framework for alzheimer's disease diagnosis via 3d-cnn and fsbi-lstm. *IEEE Access*, 7:63605–63618, 2019.

[68] Qizhang Feng, Mengnan Du, Na Zou, and Xia Hu. Fair machine learning in healthcare: A survey. *IEEE Transactions on Artificial Intelligence*, 2024.

[69] Wei Feng, Nicholas Van Halm-Lutterodt, Hao Tang, Andrew Mecum, Mohamed Kamal Mesregah, Yuan Ma, Haibin Li, Feng Zhang, Zhiyuan Wu, Erlin Yao, et al. Automated mri-based deep learning model for detection of alzheimer's disease process. *International Journal of Neural Systems*, 30(06):2050032, 2020.

[70] Maayan Frid-Adar, Idit Diamant, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. Gan-based synthetic medical image augmentation for increased cnn performance in liver lesion classification. *Neurocomputing*, 321:321–331, 2018.

[71] Alessandro Gambetti, Qiwei Han, Hong Shen, and Cláudia Soares. A survey on human-centered evaluation of explainable ai methods in clinical decision support systems. *arXiv preprint arXiv:2502.09849*, 2025.

[72] Yunhe Gao, Mu Zhou, and Dimitris N Metaxas. Utnet: a hybrid transformer architecture for medical image segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 61–71. Springer, 2021.

[73] Mavis Gezimati and Ghanshyam Singh. Deep learning for multimodal breast cancer characterization with emergence of terahertz and infrared imaging. *IEEE Transactions on Instrumentation and Measurement*, 2025.

[74] Nafiseh Ghaffar Nia, Erkan Kaplanoglu, and Ahad Nasab. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3(1):5, 2023.

[75] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings*, 2020:191, 2020.

[76] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

[77] Prashant Gohel, Priyanka Singh, and Manoranjan Mohanty. Explainable ai: current status and future directions. *arXiv preprint arXiv:2107.07045*, 2021.

[78] Lovedeep Gondara. Medical image denoising using convolutional denoising autoencoders. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*, pages 241–246. IEEE, 2016.

[79] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[80] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[81] Valerio Guarrasi, Fatih Aksu, Camillo Maria Caruso, Francesco Di Feola, Aurora Rofena, Filippo Ruffini, and Paolo Soda. A systematic review of intermediate fusion in multimodal deep learning for biomedical applications. *Image and Vision Computing*, page 105509, 2025.

[82] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *jama*, 316(22):2402–2410, 2016.

[83] Odd Erik Gundersen and Sigbjørn Kjensmo. State of the art: Reproducibility in artificial intelligence. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[84] Aditya Gupta and Amritpal Singh. Healthcare 4.0: recent advancements and futuristic research directions. *Wireless Personal Communications*, 129(2):933–952, 2023.

[85] Rajesh Gupta, Pronaya Bhattacharya, Sudeep Tanwar, Neeraj Kumar, and Sherali Zeadally. Garuda: A blockchain-based delivery scheme using drones for healthcare 5.0 applications. *IEEE Internet of Things Magazine*, 4(4):60–66, 2022.

[86] Abid Haleem, Mohd Javaid, and Ibrahim Haleem Khan. Current status and applications of artificial intelligence (ai) in medical field: An overview. *Current Medicine Research and Practice*, 9(6):231–237, 2019.

[87] Abid Haleem, Mohd Javaid, Ravi Pratap Singh, and Rajiv Suman. Medical 4.0 technologies for healthcare: Features, capabilities, and applications. *Internet of Things and Cyber-Physical Systems*, 2:12–30, 2022.

[88] Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

[89] Yixue Hao, Mohd Usama, Jun Yang, M Shamim Hossain, and Ahmed Ghoneim. Recurrent convolutional neural network based multimodal disease risk prediction. *Future Generation Computer Systems*, 92:76–83, 2019.

[90] Andrew Hard, Kanishka Rao, Rajiv Mathews, Swaroop Ramaswamy, Françoise Beaufays, Sean Augenstein, Hubert Eichner, Chloé Kiddon, and Daniel Ramage. Federated learning for mobile keyboard prediction. *arXiv preprint arXiv:1811.03604*, 2018.

[91] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[92] Vahideh Hayyolalam, Moayad Aloqaily, Öznur Özkasap, and Mohsen Guizani. Edge intelligence for empowering iot-based healthcare systems. *IEEE Wireless Communications*, 28(3):6–14, 2021.

[93] Huiguang He, Hongwei Wen, Dai Dai, and Jieqiong Wang. Computer-aided prognosis: Accurate prediction of patients with neurologic and psychiatric diseases via multi-modal mri analysis. In *Artificial Intelligence in Decision Support Systems for Diagnosis in Medical Imaging*, pages 225–265. Springer, 2018.

[94] Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

[95] Robert Hertel and Rachid Benlamri. Deep learning techniques for covid-19 diagnosis and prognosis based on radiological imaging. *ACM Computing Surveys*, 55(12):1–39, 2023.

[96] Zahra Hoodbhoy, Babar Hasan, and Khan Siddiqui. Does artificial intelligence have any role in healthcare in low resource settings? *Journal of Medical Artificial Intelligence*, 2, 2019.

[97] Yoshimasa Horie, Toshiyuki Yoshio, Kazuharu Aoyama, Shoichi Yoshimizu, Yusuke Horiuchi, Akiyoshi Ishiyama, Toshiaki Hirasawa, Tomohiro Tsuchida, Tsuyoshi Ozawa, Soichiro Ishihara, et al. Diagnostic outcomes of esophageal cancer by artificial intelligence using convolutional neural networks. *Gastrointestinal endoscopy*, 89(1):25–32, 2019.

[98] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

[99] Nussair Adel Hroub, Ali Nader Alsannaa, Maad Alowaifeer, Motaz Alfarraj, and Emmanuel Okafor. Explainable deep learning diagnostic system for prediction of lung disease from medical images. *Computers in Biology and Medicine*, 170:108012, 2024.

[100] Tzu-Ming Harry Hsu, Wei-Hung Weng, Willie Boag, Matthew McDermott, and Peter Szolovits. Unsupervised multimodal representation learning across medical images and reports. *arXiv preprint arXiv:1811.08615*, 2018.

[101] Kai-Lung Hua, Che-Hao Hsu, Shintami Chusnul Hidayati, Wen-Huang Cheng, and Yu-Jen Chen. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. *OncoTargets and therapy*, pages 2015–2022, 2015.

[102] Tim Hulsen. Explainable artificial intelligence (xai): concepts and challenges in healthcare. *AI*, 4(3):652–666, 2023.

[103] Dildar Hussain, Mohammed A Al-Masni, Muhammad Aslam, Abolghasem Sadeghi-Niaraki, Jamil Hussain, Yeong Hyeon Gu, and Rizwan Ali Naqvi. Revolutionizing tumor detection and classification in multimodality imaging based on deep learning approaches: Methods, applications and limitations. *Journal of X-Ray Science and Technology*, 32(4):857–911, 2024.

[104] Matthew Hutson. Artificial intelligence faces reproducibility crisis, 2018.

[105] Vo Trong Quang Huy and Chih-Min Lin. An improved densenet deep neural network model for tuberculosis detection using chest x-ray images. *IEEE Access*, 11:42839–42849, 2023.

[106] Jhon A Intriago, Pablo A Estevez, Jose A Cortes-Briones, Cecilia A Okuma, Fernando A Henriquez, Patricia Lillo, and Andrea Z Slachevsky. Detecting early risk of alzheimer's disease using self-supervised multimodal representation learning. In *2023 IEEE Conference on Artificial Intelligence (CAI)*, pages 158–160. IEEE, 2023.

[107] MD Samiul Islam, Haider Muhamed Umran, Samir M Umran, and Mohammed Karim. Intelligent healthcare platform: cardiovascular disease risk factors prediction using attention module based lstm. In *2019 2nd international conference on artificial intelligence and big data (ICAIBD)*, pages 167–175. IEEE, 2019.

[108] Tanzir Ul Islam, Reza Ghasemi, and Noman Mohammed. Privacy-preserving federated learning model for healthcare data. In *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, pages 0281–0287. IEEE, 2022.

[109] Elif Izci, Mehmet Akif Ozdemir, Murside Degirmenci, and Aydin Akan. Cardiac arrhythmia detection from 2d ecg images by using deep learning technique. In *2019 medical technologies congress (TIPTEKNO)*, pages 1–4. IEEE, 2019.

[110] Abhyuday N Jagannatha and Hong Yu. Bidirectional rnn for medical event detection in electronic health records. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2016, page 473, 2016.

[111] Hashaam Jamil, Waleed Tariq, Muhammad Atif Ameer, Muhammad Sohaib Asghar, Hamid Mahmood, Muhammad Junaid Tahir, and Zohaib Yousaf. Interventional radiology in low-and middle-income countries. *Annals of Medicine and Surgery*, 77, 2022.

[112] Ebenezer Jangam, Aaron Antonio Dias Barreto, and Chandra Sekhara Rao Annavarapu. Automatic detection of covid-19 from chest ct scan and chest x-rays images using deep learning, transfer learning and stacking. *Applied Intelligence*, 52(2):2243–2259, 2022.

[113] RR Janghel and YK Rathore. Deep convolution neural network based system for early diagnosis of alzheimer's disease. *Irbm*, 42(4):258–267, 2021.

[114] Mohd Javaid, Abid Haleem, Ravi Pratap Singh, Rajiv Suman, and Shanay Rab. Significance of machine learning in healthcare: Features, pillars and applications. *International Journal of Intelligent Networks*, 3:58–73, 2022.

[115] Pandia Rajan Jeyaraj and Edward Rajan Samuel Nadar. Computer-assisted medical image classification for early diagnosis of oral cancer employing deep learning algorithm. *Journal of cancer research and clinical oncology*, 145:829–837, 2019.

[116] Zhicheng Jiao, Xinbo Gao, Ying Wang, and Jie Li. A parasitic metric learning net for breast mass classification based on mammography. *Pattern Recognition*, 75:292–301, 2018.

[117] Yuliana Jiménez-Gaona, María José Rodríguez-Álvarez, and Vasudevan Lakshminarayanan. Deep-learning-based computer-aided systems for breast cancer imaging: a critical review. *Applied Sciences*, 10(22):8298, 2020.

[118] Weina Jin, Xiaoxiao Li, Mostafa Fatehi, and Ghassan Hamarneh. Guidelines and evaluation of clinical explainable ai in medical image analysis. *Medical image analysis*, 84:102684, 2023.

[119] Georgios A Kaissis, Marcus R Makowski, Daniel Rückert, and Rickmer F Braren. Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6):305–311, 2020.

[120] Erkan Kaplanoglu, A Nasab, et al. Evaluation of artificial intelligence techniques in disease diagnosis and prediction. *Discover Artificial Intelligence*, 3(1), 2023.

[121] Naresh Kumar Kar, S Jana, Abdur Rahman, Patil Rahul Ashokrao, G Indhumathi, and R Alarmelu Mangai. Automated intracranial hemorrhage detection using deep learning in medical image analysis. In *2024 International Conference on Data Science and Network Security (ICDSNS)*, pages 1–6. IEEE, 2024.

[122] Chiranjeevi Karri, Lalit Garg, Vijay Prakash, and Bhushan Dinkar Pawar. Healthcare 5.0 opportunities and challenges: A literature review. *Intelligent Biomedical Technologies and Applications for Healthcare 5.0*, pages 133–146, 2025.

[123] Maxime Kayser, Bayar Menzat, Cornelius Emde, Bogdan Bercean, Alex Novak, Abdala Espinosa, Bartlomiej W Papiez, Susanne Gaube, Thomas Lukasiewicz, and Oana-Maria Camburu. Fool me once? contrasting textual and visual explanations in a clinical decision-support setting. *arXiv preprint arXiv:2410.12284*, 2024.

[124] Christopher J Kelly, Alan Karthikesalingam, Mustafa Suleyman, Greg Corrado, and Dominic King. Key challenges for delivering clinical impact with artificial intelligence. *BMC medicine*, 17:1–9, 2019.

[125] Aditya Khamparia, Prakash Kumar Singh, Poonam Rani, Debabrata Samanta, Ashish Khanna, and Bharat Bhushan. An internet of health things-driven deep learning framework for detection and classification of skin cancer using transfer learning. *Transactions on Emerging Telecommunications Technologies*, 32(7):e3963, 2021.

[126] Md Saikat Islam Khan, Anichur Rahman, Tanoy Debnath, Md Razaul Karim, Mostofa Kamal Nasir, Shahab S Band, Amir Mosavi, and Iman Dehzangi. Accurate brain tumor detection using deep convolutional neural network. *Computational and structural biotechnology journal*, 20:4733–4745, 2022.

[127] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[128] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[129] Roshan R Kotkondawar, Sanjay R Sutar, Arvind W Kiwelekar, and Vinod J Kadam. Integrating transformer-based language model for drug discovery. In *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*, pages 1096–1101. IEEE, 2024.

[130] A Krishnaveni, G Vinoth Rajkumar, J Relin Francis Raj, R Santhana Krishnan, S Murali, and Imran Javeed Settu. Multimodnet: A multimodal deep learning model for copd staging based on chest x-ray and pulmonary function tests. In *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, pages 1683–1690. IEEE, 2024.

[131] Maheshwar Kuchana, Amritesh Srivastava, Ronald Das, Justin Mathew, Atul Mishra, and Kiran Khatter. Ai aiding in diagnosing, tracking recovery of covid-19 using deep learning on chest ct scans. *Multimedia tools and applications*, 80:9161–9175, 2021.

[132] Casimir A Kulikowski. Beginnings of artificial intelligence in medicine (aim): computational artifice assisting scientific inquiry and clinical art–with reflections on present aim challenges. *Yearbook of medical informatics*, 28(01):249–256, 2019.

[133] A Kumar and J Kaur. Machine learning and deep learning based healthcare system: A review. *Clin Case Rep Stud*, 5(6):1–5, 2024.

[134] M Sirish Kumar, Gangineni Charmi, Yeragadinda Chandana, Dasari Venkata Jaiyesh, and Mallela Hari Kartheek. Advanced multimodal deep learning for predicting cognitive decline in alzheimer's disease. In *2025 Fourth International Conference on Smart Technologies, Communication and Robotics (STCR)*, pages 1–6. IEEE, 2025.

[135] Mohit Kumar, Ashwani Kumar, Sahil Verma, Pronaya Bhattacharya, Deepak Ghimire, Seong-heum Kim, and ASM Sanwar Hosen. Healthcare internet of things (h-iot): Current trends, future prospects, applications, challenges, and security issues. *Electronics*, 12(9):2050, 2023.

[136] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

[137] Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung-Ah Sohn, and Dokyoon Kim. Predicting alzheimer's disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1):1952, 2019.

[138] Yeong Chan Lee, Jiho Cha, Injeong Shim, Woong-Yang Park, Se Woong Kang, Dong Hui Lim, and Hong-Hee Won. Multimodal deep learning of fundus abnormalities and traditional risk factors for cardiovascular risk prediction. *npj Digital Medicine*, 6(1):14, 2023.

[139] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023.

[140] Jingshan Li and Pascale Carayon. Health care 4.0: A vision for smart and connected health care. *IISE Transactions on Healthcare Systems Engineering*, 11(3):171–180, 2021.

[141] Kangshun Li, Can Chen, Wuteng Cao, Hui Wang, Shuai Han, Renjie Wang, Zaisheng Ye, Zhijie Wu, Wenxiang Wang, Leng Cai, et al. Deaf: a multimodal deep learning framework for disease prediction. *Computers in Biology and Medicine*, 156:106715, 2023.

[142] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE signal processing magazine*, 37(3):50–60, 2020.

[143] Xiaomeng Li, Mengyu Jia, Md Tauhidul Islam, Lequan Yu, and Lei Xing. Self-supervised feature learning via exploiting multi-modal data for retinal disease diagnosis. *IEEE Transactions on Medical Imaging*, 39(12):4023–4033, 2020.

[144] Xiaoxiao Li, Yufeng Gu, Nicha Dvornek, Lawrence H Staib, Pamela Ventola, and James S Duncan. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: Abide results. *Medical image analysis*, 65:101765, 2020.

[145] Yanchun Li, Qiuzhen Wang, Jie Zhang, Lingzhi Hu, and Wanli Ouyang. The theoretical research of generative adversarial networks: an overview. *Neurocomputing*, 435:26–41, 2021.

[146] Yikuan Li, Mohammad Mamouei, Gholamreza Salimi-Khorshidi, Shishir Rao, Abdelaali Hassaine, Dexter Canoy, Thomas Lukasiewicz, and Kazem Rahimi. Hi-behrt: hierarchical transformer-based model for accurate prediction of clinical events using multimodal longitudinal electronic health records. *IEEE journal of biomedical and health informatics*, 27(2):1106–1117, 2022.

[147] Zhixi Li, Yifan He, Stuart Keel, Wei Meng, Robert T Chang, and Mingguang He. Efficacy of a deep learning system for detecting glaucomatous optic neuropathy based on color fundus photographs. *Ophthalmology*, 125(8):1199–1206, 2018.

[148] Yuxuan Liang, Hanqing Chao, Jiajin Zhang, Ge Wang, and Pingkun Yan. Unbiasing fairness evaluation of radiology ai model. *Meta-radiology*, 2(3):100084, 2024.

[149] Qika Lin, Yifan Zhu, Xin Mei, Ling Huang, Jingying Ma, Kai He, Zhen Peng, Erik Cambria, and Mengling Feng. Has multimodal learning delivered universal intelligence in healthcare? a comprehensive survey. *Information Fusion*, 116:102795, 2025.

[150] Siqi Liu, Sidong Liu, Weidong Cai, Hangyu Che, Sonia Pujol, Ron Kikinis, Dagan Feng, Michael J Fulham, et al. Multimodal neuroimaging feature learning for multiclass diagnosis of alzheimer's disease. *IEEE transactions on biomedical engineering*, 62(4):1132–1140, 2014.

[151] Xiaoxuan Liu, Samantha Cruz Rivera, David Moher, Melanie J Calvert, Alastair K Denniston, Hutan Ashrafian, Andrew L Beam, An-Wen Chan, Gary S Collins, Ara DarziJonathan J Deeks, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the consort-ai extension. *The Lancet Digital Health*, 2(10):e537–e548, 2020.

[152] Kosmia Loizidou, Rafaella Elia, and Costas Pitris. Computer-aided breast cancer detection and classification in mammography: A comprehensive review. *Computers in Biology and Medicine*, 153:106554, 2023.

[153] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[154] Htet Myet Lynn, Sung Bum Pan, and Pankoo Kim. A deep bidirectional gru network model for biometric electrocardiogram classification based on recurrent neural networks. *Ieee Access*, 7:145395–145405, 2019.

[155] M Mahalakshmi, Gangisetty Raj Charan, and Geetanj Sharma. Integrative breast cancer detection: A deep learning approach with multi-modal data fusion of mammograms, prescription and blood reports. In *2024 International Conference on Communication, Computing and Internet of Things (IC3IoT)*, pages 1–8. IEEE, 2024.

[156] Md Ishtyaq Mahmud, Muntasir Mamun, and Ahmed Abdelgawad. A deep analysis of brain tumor detection from mr images using deep learning networks. *Algorithms*, 16(4):176, 2023.

[157] SM Mahmud, Md Mamun Ali, Mohammad Fahim Shahriar, Fahad Ahmed Al-Zahrani, Kawsar Ahmed, Dip Nandi, and Francis M Bui. Detection of different stages of alzheimer's disease using cnn classifier. *Computers, Materials and Continua*, 76(3):3933–3948, 2023.

[158] Tanjim Mahmud, Koushick Barua, Sultana Umme Habiba, Nahed Sharmen, Mohammad Shahadat Hossain, and Karl Andersson. An explainable ai paradigm for alzheimer's diagnosis using deep transfer learning. *Diagnostics*, 14(3):345, 2024.

[159] Paramita Kundu Maji, Soubhik Acharya, Priti Paul, Sanjay Chakraborty, and Saikat Basu. Deep learning inspired game-based cognitive assessment for early dementia detection. *Engineering Applications of Artificial Intelligence*, 142:109901, 2025.

[160] Samir Malakar, Soumya Deep Roy, Soham Das, Swaraj Sen, Juan D Velasquez, and Ram Sarkar. Computer based diagnosis of some chronic diseases: a medical journey of the last two decades. *Archives of Computational Methods in Engineering*, 29(7):5525–5567, 2022.

[161] Manjula Mandava et al. Mdensnet201-idrsrnet: Efficient cardiovascular disease prediction system using hybrid deep learning. *Biomedical Signal Processing and Control*, 93:106147, 2024.

[162] Mireya Martínez-García and Enrique Hernández-Lemus. Data integration challenges for machine learning in precision medicine. *Frontiers in medicine*, 8:784455, 2022.

[163] Balduíno César Mateus, Mateus Mendes, José Torres Farinha, Rui Assis, and António Marques Cardoso. Comparing lstm and gru models to predict the condition of a pulp paper press. *Energies*, 14(21):6958, 2021.

[164] Maciej A Mazurowski, Haoyu Dong, Hanxue Gu, Jichen Yang, Nicholas Konz, and Yixin Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023.

[165] Elliot Mbunge, Benhildah Muchemwa, Sipho'esihle Jiyane, and John Batani. Sensors and healthcare 5.0: transformative shift in virtual care through emerging digital health technologies. *Global Health Journal*, 5(4):169–177, 2021.

[166] GA Meijer, JA Beliën, PJ Van Diest, and JP Baak. Origins of... image analysis in clinical pathology. *Journal of clinical pathology*, 50(5):365, 1997.

[167] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2018.

[168] Sushruta Mishra, Anuttam Dash, and Lambodar Jena. Use of deep learning for disease detection and diagnosis. In *Bio-inspired neurocomputing*, pages 181–201. Springer, 2020.

[169] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 220–229, 2019.

[170] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *arXiv preprint arXiv:1811.07867*, 2018.

[171] Shentong Mo and Paul Pu Liang. Multimed: Massively multimodal and multitask medical understanding. *arXiv preprint arXiv:2408.12682*, 2024.

[172] Mana Moassefi, Pouria Rouzrokh, Gian Marco Conte, Sanaz Vahdati, Tianyuan Fu, Aylin Tahmasebi, Mira Younis, Keyvan Farahani, Amilcare Gentili, Timothy Kline, et al. Reproducibility of deep learning algorithms developed for medical imaging analysis: A systematic review. *Journal of digital imaging*, 36(5):2306–2312, 2023.

[173] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[174] Ashok Kumar Munnangi, Satheeshwaran UdhayaKumar, Vinayakumar Ravi, Ramesh Sekaran, and Suthendran Kannan. Survival study on deep learning techniques for iot enabled smart healthcare system. *Health and Technology*, 13(2):215–228, 2023.

[175] Nafeesa Yousuf Murad, Mohd Hilmi Hasan, Muhammad Hamza Azam, Nadia Yousuf, and Jameel Shehu Yalli. Unraveling the black box: A review of explainable deep learning healthcare techniques. *IEEE Access*, 12:66556–66568, 2024.

[176] A Angel Nancy, Dakshanamoorthy Ravindran, PM Durai Raj Vincent, Kathiravan Srinivasan, and Daniel Gutierrez Reina. Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. *Electronics*, 11(15):2292, 2022.

[177] Ali Nawaz, Shehroz S Khan, and Amir Ahmad. Ensemble of autoencoders for anomaly detection in biomedical data: A narrative review. *IEEE Access*, 12:17273–17289, 2024.

[178] Zubaira Naz, Muhammad Usman Ghani Khan, Tanzila Saba, Amjad Rehman, Haitham Nobanee, and Saeed Ali Bahaj. An explainable ai-enabled framework for interpreting pulmonary diseases from chest radiographs. *Cancers*, 15(1):314, 2023.

[179] Asifa Nazir, Ahsan Hussain, Mandeep Singh, and Assif Assad. Deep learning in medicine: advancing healthcare with intelligent solutions and the future of holography imaging in early diagnosis. *Multimedia Tools and Applications*, 84(17):17677–17740, 2025.

[180] Minh Nguyen, Nanbo Sun, Daniel C Alexander, Jiashi Feng, and BT Thomas Yeo. Modeling alzheimer's disease progression using deep recurrent neural networks. In *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, pages 1–4. IEEE, 2018.

[181] Dur-E-Maknoon Nisar, Rashid Amin, Noor-Ul-Huda Shah, Mohammed A. Al Ghamdi, Sultan H. Almotiri, and Meshrif Alruily. Healthcare techniques through deep learning: Issues, challenges and opportunities. *IEEE Access*, 9:98523–98541, 2021.

[182] Syed Muhammad Hayyan Nishat, Ammar Shahid Tanweer, Bashayer Alshamsi, Majd H Shaheen, Ariba Shahid Tanveer, Aroob Nishat, Yaman Alharbat, Ahmad Alaboud, Mahra Almazrouei, and Raghad A Ali-Mohamed. Artificial intelligence: A new frontier in rare disease early diagnosis. *Cureus*, 17(2), 2025.

[183] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

[184] Modupe Odusami, Rytis Maskeliūnas, Robertas Damaševičius, and Sanjay Misra. Explainable deep-learning-based diagnosis of alzheimer's disease using multimodal input fusion of pet and mri images. *Journal of Medical and Biological Engineering*, 43(3):291–302, 2023.

[185] Shu Lih Oh, Eddie YK Ng, Ru San Tan, and U Rajendra Acharya. Automated diagnosis of arrhythmia using combination of cnn and lstm techniques with variable length heart beats. *Computers in biology and medicine*, 102:278–287, 2018.

[186] World Health Organization et al. *Ethics and governance of artificial intelligence for health: WHO guidance: executive summary*. World Health Organization, 2021.

[187] Andres Ortiz, Jorge Munilla, Juan M Gorriz, and Javier Ramirez. Ensembles of deep learning architectures for the early diagnosis of the alzheimer's disease. *International journal of neural systems*, 26(07):1650025, 2016.

[188] Ishak Pacal. Maxcervixt: A novel lightweight vision transformer-based approach for precise cervical cancer detection. *Knowledge-Based Systems*, 289:111482, 2024.

[189] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[190] Mohammad Khalid Pandit and Shoaib Amin Banday. Sars n-cov2-19 detection from chest x-ray images using deep neural networks. *International Journal of Pervasive Computing and Communications*, 16(5):419–427, 2020.

[191] Abhijeet Parida, Daniel Capellán-Martín, Zhifan Jiang, Austin Tapp, Xinyang Liu, Syed Muhammad Anwar, María J Ledesma-Carbayo, and Marius George Linguraru. Adult glioma segmentation in sub-saharan africa using transfer learning on stratified finetuning data. *arXiv preprint arXiv:2412.04111*, 2024.

[192] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021.

[193] Fernando Roberto Pereira, João Mario Clementin De Andrade, Dante Luiz Escuissato, and Lucas Ferrari De Oliveira. Classifier ensemble based on computed tomography attenuation patterns for computer-aided detection system. *IEEE Access*, 9:123134–123145, 2021.

[194] Ramesh Pingili et al. Generative ai unlocking adaptive workflow design. *Journal of Next-Generation Research 5.0*, 2025.

[195] Vladimir V Popov, Elena V Kudryavtseva, Nirmal Kumar Katiyar, Andrei Shishkin, Stepan I Stepanov, and Saurav Goel. Industry 4.0 and digitalisation in healthcare. *Materials*, 15(6):2140, 2022.

[196] David Powers. Ailab. evaluation: From precision, recall and f-measure to roc, informedness, markedness & correlation. *J. Mach. Learn. Technol*, 2(22293981):01, 2011.

[197] Foster Provost and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking.* " O'Reilly Media, Inc.", 2013.

[198] Vijaya Lakshmi PS, Satheesh Kumar Nataraj, et al. Foundations and obstacles of deep learning in healthcare. In *Revolutionizing Data Science and Analytics for Industry Transformation*, pages 153–174. IGI Global Scientific Publishing, 2025.

[199] Anichur Rahman, Tanoy Debnath, Dipanjali Kundu, Md Saikat Islam Khan, Airin Afroj Aishi, Sadia Sazzad, Mohammad Sayduzzaman, and Shahab S Band. Machine learning and deep learning-based approach in smart healthcare: Recent advances, applications, challenges and opportunities. *AIMS Public Health*, 11(1):58, 2024.

[200] Deepika Rajagopal and Pradeep Kumar Thimma Subramanian. Ai augmented edge and fog computing for internet of health things (ioht). *PeerJ Computer Science*, 11:e2431, 2025.

[201] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

[202] K. Ramu, Sridhar Patthi, Yogendra Narayan Prajapati, Janjhyam Venkata Naga Ramesh, Sudipta Banerjee, K.B.V. Brahma Rao, Saleh I. Alzahrani, and Rajaram ayyasamy. Hybrid cnn-svm model for enhanced early detection of chronic kidney disease. *Biomedical Signal Processing and Control*, 100:107084, 2025.

[203] Sita Rani, Raman Kumar, BS Panda, Rajender Kumar, Nafaa Farhan Muften, Mayada Ahmed Abass, and Jasmina Lozanović. Machine learning-powered smart healthcare systems in the era of big data: Applications, diagnostic insights, challenges, and ethical implications. *Diagnostics*, 15(15):1914, 2025.

[204] D Ranjith and M Sakthivanitha. A novel multi-modal deep learning framework for early detection of ocular diseases. In *2025 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 1340–1346. IEEE, 2025.

[205] G. Madhukar Rao, Dharavath Ramesh, Vandana Sharma, Anurag Sinha, Md. Mehedi Hassan, and Amir H. Gandomi. Attgru-hmsi: enhancing heart disease diagnosis using hybrid deep learning approach. *Scientific Reports*, 14, 2024.

[206] Khalid Raza and Nripendra K Singh. A tour of unsupervised deep learning for medical image analysis. *Current Medical Imaging Reviews*, 17(9):1059–1077, 2021.

[207] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[208] David A Rosman, Judith Bamporiki, Rebecca Stein-Wexler, and Robert D Harris. Developing diagnostic radiology training in low resource countries. *Current Radiology Reports*, 7:1–7, 2019.

[209] Munshi Saifuzzaman and Tajkia Nuri Ananna. Toward smart healthcare: challenges and opportunities in iot and ml. *IoT and ML for information management: A smart healthcare perspective*, pages 325–355, 2024.

[210] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29, 2016.

[211] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J Anders, and Klaus-Robert Müller. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278, 2021.

[212] Iván Sánchez Fernández and Jurriaan M Peters. Machine learning and deep learning in medicine and neuroimaging. *Annals of the Child Neurology Society*, 1(2):102–122, 2023.

[213] Deepti Saraswat, Pronaya Bhattacharya, Ashwin Verma, Vivek Kumar Prasad, Sudeep Tanwar, Gulshan Sharma, Pitshou N Bokoro, and Ravi Sharma. Explainable ai for healthcare 5.0: opportunities and challenges. *IEEe Access*, 10:84486–84517, 2022.

[214] Peter Savadjiev, Jaron Chong, Anthony Dohan, Maria Vakalopoulou, Caroline Reinhold, Nikos Paragios, and Benoit Gallix. Demystification of ai-driven medical image interpretation: past, present and future. *European radiology*, 29:1616–1624, 2019.

[215] Allen Schmaltz and Andrew L Beam. Sharpening the resolution on data matters: a brief roadmap for understanding deep learning for medical data. *The Spine Journal*, 21(10):1606–1609, 2021.

[216] Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117, 2015.

[217] Daan Schouten, Giulia Nicoletti, Bas Dille, Catherine Chia, Pierpaolo Vendittelli, Megan Schuurmans, Geert Litjens, and Nadieh Khalili. Navigating the landscape of multimodal ai in medicine: a scoping review on technical challenges and clinical applications. *Medical Image Analysis*, page 103621, 2025.

[218] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[219] Mark P Sendak, Joshua D'Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. *EMJ Innov*, 10:19–00172, 2020.

[220] Shahab Shamshirband, Mahdis Fathi, Abdollah Dehzangi, Anthony Theodore Chronopoulos, and Hamid Alinejad-Rokny. A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. *Journal of Biomedical Informatics*, 113:103627, 2021.

[221] Moolchand Sharma, Bhanu Jain, Chetan Kargeti, Vinayak Gupta, and Deepak Gupta. Detection and diagnosis of skin diseases using residual neural networks (resnet). *International Journal of Image and Graphics*, 21(05):2140002, 2021.

[222] Shagun Sharma and Kalpna Guleria. A deep learning based model for the detection of pneumonia from chest x-ray images using vgg-16 and neural networks. *Procedia Computer Science*, 218:357–366, 2023.

[223] Amr I Shehta, Mona Nasr, and Alaa El Din M El Ghazali. Blood cancer prediction model based on deep learning technique. *Scientific Reports*, 15(1):1889, 2025.

[224] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19(1):221–248, 2017.

[225] Farzan Shenavarmasouleh, Farid Ghareh Mohammadi, Khaled M Rasheed, and Hamid R Arabnia. Deep learning in healthcare: An in-depth analysis. *arXiv preprint arXiv:2302.10904*, 2023.

[226] Ruey-Kai Sheu, Lun-Chi Chen, Chieh-Liang Wu, Mayuresh Sunil Pardeshi, Kai-Chih Pai, Chien-Chung Huang, Chia-Yu Chen, and Wei-Cheng Chen. Multi-modal data analysis for pneumonia status prediction using deep learning (mda-psp). *Diagnostics*, 12(7):1706, 2022.

[227] Xiaoyu Shi, Rahul Kumar Jain, Yinhao Li, Shurong Chai, Jingliang Cheng, Jie Bai, Guohua Zhao, Lanfen Lin, and Yen-Wei Chen. Multi-modal medical sam: An adaptation method of segment anything model (sam) for glioma segmentation using multi-modal mr images. *ACM Transactions on Computing for Healthcare*, 6(2):1–21, 2025.

[228] Debaditya Shome, Tejaswini Kar, Sachi Nandan Mohanty, Prayag Tiwari, Khan Muhammad, Abdullah AlTameem, Yazhou Zhang, and Abdul Khader Jilani Saudagar. Covid-transformer: Interpretable covid-19 detection using vision transformer for healthcare. *International Journal of Environmental Research and Public Health*, 18(21):11086, 2021.

[229] Edward H Shortliffe and Martin J Sepúlveda. Clinical decision support in the era of artificial intelligence. *Jama*, 320(21):2199–2200, 2018.

[230] Aisha Siam, Abdel Rahman Alsaify, Bushra Mohammad, Md Rafiul Biswas, Hazrat Ali, and Zubair Shah. Multimodal deep learning for liver cancer applications: a scoping review. *Frontiers in artificial intelligence*, 6:1247195, 2023.

[231] Sarkar Siddique and James CL Chow. Machine learning in healthcare communication. *Encyclopedia*, 1(1):220–239, 2021.

[232] Vijendra Singh, Vijayan K Asari, and Rajkumar Rajasekaran. A deep neural network for early detection and prediction of chronic kidney disease. *Diagnostics*, 12(1):116, 2022.

[233] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information processing & management*, 45(4):427–437, 2009.

[234] Nitiwat Sompawong, Jintapatee Mopan, Pakinee Pooprasert, Wanwisa Himakhun, Komsun Suwannarurk, Jarun Ngamvirojcharoen, Tee Vachiramon, and Charturong Tantibundhit. Automated pap smear cervical cancer screening using deep learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 7044–7048. IEEE, 2019.

[235] Junho Song, Young Jun Chai, Hiroo Masuoka, Sun-Won Park, Su-jin Kim, June Young Choi, Hyoun-Joong Kong, Kyu Eun Lee, Joongseek Lee, Nojun Kwak, et al. Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules. *Medicine*, 98(15):e15133, 2019.

[236] Vartika Srivastava, Ravinder Kumar, Mohmmad Younus Wani, Keven Robinson, and Aijaz Ahmad. Role of artificial intelligence in early diagnosis and treatment of infectious diseases. *Infectious Diseases*, 57(1):1–26, 2025.

[237] Demetrio Stojanoff. Index of hadamard multiplication by positive matrices. *Linear algebra and its applications*, 290(1-3):95–108, 1999.

[238] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. arxiv: 190602243 [cs], 2019.

[239] NP Subiramaniyam et al. Enhanced real-time analysis and anomaly detection in smart health monitoring systems through integration of deep learning algorithm with iot-edge computing. In *2024 International Conference on Computing and Data Science (ICCDS)*, pages 1–6. IEEE, 2024.

[240] MA Sufian, L Alsadder, W Hamzi, S Zaman, ASMS Sagar, and B Hamzi. Mitigating algorithmic bias in ai-driven cardiovascular imaging for fairer diagnostics. diagnostics 2024; 14: 2675.

[241] Zhaoyi Sun, Mingquan Lin, Qingqing Zhu, Qianqian Xie, Fei Wang, Zhiyong Lu, and Yifan Peng. A scoping review on multimodal deep learning in biomedical images and texts. *Journal of biomedical informatics*, 146:104482, 2023.

[242] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[243] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[244] Sharia Arfin Tanim, Al Rafi Aurnob, Tahmid Enam Shrestha, MD Rokon Islam Emon, M.F. Mridha, and Md Saef Ullah Miah. Explainable deep learning for diabetes diagnosis with deepnetx2. *Biomedical Signal Processing and Control*, 99:106902, 2025.

[245] Huiqian Tao, Chengfeng Wang, Hongxia Qi, Hui Li, Yane Li, Ruifei Xie, Yuzhu Dai, Qingyang Sun, Yingqiang Zhang, Xinyi Yu, et al. A real-time computer-aided diagnosis system for coronary heart disease prediction using clinical information. *Reviews in Cardiovascular Medicine*, 26(3):26204, 2025.

[246] Medhat A Tawfeek, Ibrahim Alrashdi, Madallah Alruwaili, Warda M Shaban, and Fatma M Talaat. Enhancing the efficiency of lung cancer screening: predictive models utilizing deep learning from ct scans. *Neural Computing and Applications*, pages 1–19, 2025.

[247] Kalaiselvi Thiruvenkadam, Vasanthi Ravindran, and Anitha Thiyagarajan. Deep learning with xai based multi-modal mri brain tumor image analysis using image fusion techniques. In *2024 international conference on trends in quantum computing and emerging business technologies*, pages 1–5. IEEE, 2024.

[248] Shuo Tian, Wenbo Yang, Jehane Michael Le Grange, Peng Wang, Wei Huang, and Zhewei Ye. Smart healthcare: making medical care more intelligent. *Global Health Journal*, 3(3):62–65, 2019.

[249] Hsinhan Tsai, Ta-Wei Yang, Tien-Yi Wu, Ya-Chi Tu, Cheng-Lung Chen, and Cheng-Fu Chou. Multitask learning multimodal network for chronic disease prediction. *Scientific Reports*, 15(1):15468, 2025.

[250] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[251] Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, et al. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138, 2024.

[252] SL van der Meijden, Y Wang, MS Arbous, BF Geerts, EW Steyerberg, and T Hernandez-Boussard. Navigating fairness in ai-based prediction models: Theoretical constructs and practical applications. *medRxiv*, pages 2025–03, 2025.

[253] Kicky Gerhilde van Leeuwen, Leon Doorn, and Erik Gelderblom. The ai act: Responsibilities and obligations for healthcare professionals and organizations. *Diagnostic and Interventional Radiology*, 2025.

[254] Emmanouil P Vardas, Maria Marketou, and Panos E Vardas. Medicine, healthcare and the ai act: gaps, challenges and future implications. *European Heart Journal-Digital Health*, page ztaf041, 2025.

[255] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[256] Korosh Vatanparvar, Viswam Nathan, Ebrahim Nemati, Md Mahbubur Rahman, and Jilong Kuang. A generative model for speech segmentation and obfuscation for remote health monitoring. In *2019 IEEE 16th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, pages 1–4. IEEE, 2019.

[257] Janani Venugopalan, Li Tong, Hamid Reza Hassanzadeh, and May D Wang. Multimodal deep learning models for early detection of alzheimer's disease stage. *Scientific reports*, 11(1):3254, 2021.

[258] Rohit Verma, Samrat Patel, Saviour Viyadis Minj, and Aruna Bhat. Healthcare 5.0: A study on improving personalized care. In *2022 6th international conference on intelligent computing and control systems (ICICCS)*, pages 1815–1818. IEEE, 2022.

[259] Michael Vogt. An overview of deep learning techniques. *at-Automatisierungstechnik*, 66(9):690–703, 2018.

[260] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

[261] Zhongwei Wan, Che Liu, Mi Zhang, Jie Fu, Benyou Wang, Sibo Cheng, Lei Ma, César Quilodrán-Casas, and Rossella Arcucci. Med-unic: Unifying cross-lingual medical vision-language pre-training by diminishing bias. *Advances in Neural Information Processing Systems*, 36:56186–56197, 2023.

[262] Yi Jie Wang, Wei Chong Choo, Keng Yap Ng, Ran Bi, and Peng Wei Wang. Evolution of ai enabled healthcare systems using textual data with a pretrained bert deep learning model. *Scientific Reports*, 15(1):7540, 2025.

[263] Niyaz Ahmad Wani, Ravinder Kumar, and Jatin Bedi. Deepxplainer: An interpretable deep learning based approach for lung cancer detection using explainable artificial intelligence. *Computer Methods and Programs in Biomedicine*, 243:107879, 2024.

[264] Jonathan Waring, Charlotta Lindvall, and Renato Umeton. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial intelligence in medicine*, 104:101822, 2020.

[265] Mohammad Wazid, Basudeb Bera, Ashok Kumar Das, and Devesh Pratap Singh. Iot and blockchain technology-based healthcare monitoring. In *Blockchain in Digital Healthcare*, pages 69–92. Chapman and Hall/CRC, 2021.

[266] Mohammad Wazid, Ashok Kumar Das, Noor Mohd, and Youngho Park. Healthcare 5.0 security framework: applications, issues and future research directions. *IEEE Access*, 10:129429–129442, 2022.

[267] Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340, 2019.

[268] Douglas Williams, Heiko Hornung, Adi Nadimpalli, and Ashton Peery. Deep learning and its application for healthcare delivery in low and middle income countries. *Frontiers in Artificial Intelligence*, 4:553987, 2021.

[269] Yaojue Xie, Yuansheng Zhai, and Guihua Lu. Evolution of artificial intelligence in healthcare: a 30-year bibliometric study. *Frontiers in Medicine*, 11:1505692, 2025.

[270] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[271] Wen Xu, Jing He, and Yanfeng Shu. Deephealth: Deep representation learning with autoencoders for healthcare prediction. In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 42–49. IEEE, 2020.

[272] Yiwen Xu, Tariq M Khan, Yang Song, and Erik Meijering. Edge deep learning in computer vision and medical diagnostics: a comprehensive survey. *Artificial Intelligence Review*, 58(3):93, 2025.

[273] Keyue Yan, Tengyue Li, João Alexandre Lobo Marques, Juntao Gao, and Simon James Fong. A review on multimodal machine learning in medical diagnostics. *Math. Biosci. Eng*, 20(5):8708–8726, 2023.

[274] Guang Yang, Arvind Rao, Christine Fernandez-Maloigne, Vince Calhoun, and Gloria Menegaz. Explainable ai (xai) in biomedical signal and image processing: promises and challenges. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 1531–1535. IEEE, 2022.

[275] Zhenyu Yang, Yantao Li, and Gang Zhou. Ts-gan: Time-series gan for sensor-based health data augmentation. *ACM Transactions on Computing for Healthcare*, 4(2):1–21, 2023.

[276] A Yashudas, Dinesh Gupta, GC Prashant, Amit Dua, Dokhyl AlQahtani, and A Siva Krishna Reddy. Deep-cardio: Recommendation system for cardiovascular disease prediction using iot network. *IEEE Sensors Journal*, 2024.

[277] Xin Yi, Ekta Walia, and Paul Babyn. Generative adversarial network in medical imaging: A review. *Medical image analysis*, 58:101552, 2019.

[278] Zhitao Ying, Dylan Bourgeois, Jiaxuan You, Marinka Zitnik, and Jure Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[279] Abolfazl Younesi, Elyas Oustad, Mohsen Ansari, Thomas Fahringer, and Rajkumar Buyya. Healthcare 5.0: An industry 5.0 perspective for next-generation medical systems with synergistic integration of iot, ai, and 6g. *Internet of Things*, page 101815, 2025.

[280] Zengchen Yu, Ke Wang, Zhibo Wan, Shuxuan Xie, and Zhihan Lv. Popular deep learning algorithms for disease prediction: a review. *Cluster Computing*, 26(2):1231–1251, 2023.

[281] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018.

[282] Talat Zehra, Anil Parwani, Jamshid Abdul-Ghafar, and Zubair Ahmad. A suggested way forward for adoption of ai-enabled digital pathology in low resource organizations in the developing world. *Diagnostic pathology*, 18(1):68, 2023.

[283] Xu Zhang, Yaming Wang, Liang Zhang, Bo Jin, and Hongzhe Zhang. Exploring unsupervised multivariate time series representation learning for chronic disease diagnosis. *International Journal of Data Science and Analytics*, 15(2):173–186, 2023.

[284] Yi-Lun Zhang, Wen-Tao Wang, Jia-Hui Cuan, Deepak Kumar Jain, Tian-Yang Wang, and Swalpa Kumar Roy. Mocformer: a two-stage pre-training-driven transformer for drug–target interactions prediction. *International Journal of Computational Intelligence Systems*, 17(1):165, 2024.

[285] Yiming Zhang, Ying Weng, and Jonathan Lund. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics*, 12(2):237, 2022.

[286] Xia Zhao, Limin Wang, Yufei Zhang, Xuming Han, Muhammet Deveci, and Milan Parmar. A review of convolutional neural networks in computer vision. *Artificial Intelligence Review*, 57(4):99, 2024.

[287] Julian Zilly, Joachim M Buhmann, and Dwarikanath Mahapatra. Glaucoma detection using entropy sampling and ensemble learning for automatic optic cup and disc segmentation. *Computerized Medical Imaging and Graphics*, 55:28–41, 2017.

[288] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Appendix

A Background of Foundational Deep Learning Architectures

In this section, we give an overview of the most important structures that have dominated contemporary AI: CNNs, RNNs, LSTM and GRUs.

A.1 Convolutional Neural Networks (CNNs)

CNNs are commonly used for feature extraction from pixel images and classification tasks. It was first introduced by LeCun et al. in their paper on document recognition in 1998 [136]. It lowered the reliance on manual extraction of features by sliding the convolution filters over input data [23] and captured the spatial structures. The pooling layers condense the feature maps [286] and improve computational efficiency. A fully connected layer takes the flattened feature maps as the input and produces the output [11]. Equation 5 represents the forward pass of a CNN.

$$y_{k,i,j} = \sigma\left(b_k + \sum_{c=1}^{C_{in}} \sum_{m=0}^{K_h-1} \sum_{n=0}^{K_w-1} W_{k,c,m,n} X_{c,i+m,j+n}\right) \tag{5}$$

where:

$y_{k,i,j}amp; =$ Output for channel k at position (i, j),
$\sigma amp; =$ Activation function,
$b_k amp; =$ Bias for channel k,
$C_{in} amp; =$ Total input channels,
$K_h, K_w amp; =$ Width and height of Kernel K,
$W_{k,c,m,n} amp; =$ For input channel C and output channel K, the weights at (m, n) position in the filter,
$X_{c,i+m,j+n} amp; =$ Pixel value for channel c, after moving m rows and n columns from the start.

A.2 Recurrent Neural Networks (RNNs)

RNN has three layers: input, hidden, and output layers, in which the outer layers are feed forward but the hidden layers are recurrent, in which the output of the present state is fed to the next, which is crucial for understanding sequences [31]. The hidden state contains the short-term memory. RNNs are slow and fail to handle long sequences and suffer from vanishing and exploding gradient problems. The hidden state computation is represented in equation 6. RNNs are especially well applied to clinical problems that deal with sequential or time-dependent or variable data like patient history, physiology, and clinical text. Their memory, dependency modeling, and future health outcome forecasting capabilities render them essential in time-sensitive medical activities such as modeling the course of diseases, real-time monitoring, and predictive analytics. For example, they are used in alzheimer's detection [180], detecting medical events [110], etc.

$$h_t = \phi\left(W_{xh} x_t + W_{hh} h_{t-1} + b_h\right) \tag{6}$$

where:

$h_t =$ Hidden state or internal memory of RNN,
$\phi =$ Activation function,
$x_t =$ Input vector for present state,
$W_{xh} =$ Weight matrix from current state input to hidden state,
$W_{hh} =$ Weight matrix from previous hidden state to present hidden state,
$b_h =$ Bias for the hidden layer.

A.3 Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs)

LSTM introduces another cell state, which is the long-term memory state, along with the short-term memory present in traditional RNNs. It consists of three gates discussed below and represented in equation 12 [31].

The equations represent the LSTM operations, which include forget gate, input gate, candidate cell state, cell state update, output gate, and hidden state update computations.

1. Forget gate: to decide what to discard from the previous state.
2. Input gate: decides the amount of new information to be added to the state.
3. Output gate: decides what to show as output.

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \tag{7}$$
$$i_t = \sigma\left(W_i[h_{t-1}, x_t] + b_i\right) \tag{8}$$
$$\tilde{c}_t = \tanh\left(W_c[h_{t-1}, x_t] + b_c\right) \tag{9}$$
$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{10}$$
$$o_t = \sigma\left(W_o[h_{t-1}, x_t] + b_o\right) \tag{11}$$
$$h_t = o_t \odot \tanh(c_t) \tag{12}$$

Where:

x_t = Input vector at time t,
h_t = Short-term memory – hidden state at t,
c_t = Long-term memory – cell state at t,
$\tilde{c}_t$ = Candidate cell state,
f_t, i_t, o_t = Activation for forget gate, input gate, and output gate,
W_f, W_i, W_c, W_o = Weight matrices for respective gates,
b_f, b_i, b_c, b_o = Bias vectors for respective gates,
σ = Sigmoid activation function,
$\odot$ = Hadamard multiplication [237],
$[h_{t-1}, x_t]$ = Current input and previous hidden state concatenation.

GRU was proposed by Cho et al. in 2014 [48]. It has two gates [31]:

1. Update gate: How much information must be retained from the past state?
2. Reset gate: Decides what is to be forgotten during state calculation.

The calculations are represented in equation 16

$$z_t = \sigma\left(W_z[h_{t-1}, x_t] + b_z\right) \tag{13}$$
$$r_t = \sigma\left(W_r[h_{t-1}, x_t] + b_r\right) \tag{14}$$
$$\tilde{h}_t = \tanh\left(W_h\left[r_t \odot h_{t-1}, x_t\right] + b_h\right) \tag{15}$$
$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \tilde{h}_t \tag{16}$$

Where:

x_t = Input vector at time t,
h_t = Hidden state at t,
z_t = Update gate,
r_t = Reset gate,
$\tilde{h}_t$ = Candidate hidden state,
W_z, W_r, W_h = Weight matrices for the respective gates and candidate state,
b_z, b_r, b_h = Bias vectors for the gates and candidate state,
σ = Sigmoid activation function,
$\odot$ = Hadamard multiplication,
$[h_{t-1}, x_t]$ = Current input and previous hidden state concatenation.

A.4 Autoencoders

We train the autoencoder to minimize differences between the input and the reconstructed output. Although autoencoders are unsupervised and do not require labeled data, they use input data as the target to reconstruct. Thus have a self-supervised training. Errors are calculated by taking the differences between the input and the output, usually calculating the mean squared error, to deal with positive values only. During training, the latent space becomes organized by learning the representation, and results converge, mostly forming clusters based on similarity. The encoder and decoder equations have been represented in equation 17 and 18. Due to their data compression ability, autoencoders are used in dimensionality reduction. They are used to reduce noise from data and also to detect anomalies.

$$h_j = \sigma\left(b_j^{(h)} + \sum_{i=1}^{n_x} W_{ji}^{(h)} x_i\right) \tag{17}$$

$$\hat{x}_i = \sigma\left(b_i^{(o)} + \sum_{j=1}^{n_h} W_{ij}^{(o)} h_j\right) \tag{18}$$

Where:

$$\begin{aligned} x_i &= \text{The i-th input feature} \\ h_j &= \text{The j-th latent feature} \\ \hat{x}_i &= \text{Reconstructed i-th feature} \\ W^{(h)}, W^{(o)} &= \text{Encoder and decoder weight matrices} \\ b^{(h)}, b^{(o)} &= \text{Biases} \\ \sigma &= \text{Activation function} \\ n_x &= \text{Number of input features} \\ n_h &= \text{Number of neurons in the hidden layer} \end{aligned}$$

In healthcare-related fields, autoencoders are used learning representations from unlabeled data, as medical data is extensively available but labeled data is limited [271]. They have also been used in detecting anomalies in biomedical data [177], noise removal [78], imputation, multimodal integration, and unsupervised feature learning. Their ability to handle scarce and heterogeneous data makes them highly valuable for disease detection, medical imaging, EHR modeling, and privacy-preserving healthcare solutions.