# Bridging Visual and Linguistic Intelligence for Chest X-rays: A Comprehensive Review of ViTs and LLM Synergies

Ismail Hossain
Department of Computer Science
George Mason University
Fairfax, VA, USA
ihossai4@gmu.edu

Mridul Banik
Department of Computer Science
Colorado State University
Fort Collins, CO, USA
mridul.banik23@alumni.colostate.edu

***Abstract*—The integration of Vision Transformers (ViTs) and Large Language Models (LLMs) in chest X-ray analysis has emerged as a promising solution to address the growing challenges in radiology, including increasing diagnostic workloads and the need for timely, accurate interpretations. This systematic review examines the recent advancements in ViT–LLM hybrid systems, exploring their architectural innovations, multimodal fusion strategies, and application in automated report generation. A comprehensive search of databases such as Google Scholar, PubMed, and IEEE Xplore was conducted to identify studies published between 2018 and 2025, focusing on ViT–LLM integration, performance metrics, and clinical validation. Key findings highlight that ViT–LLM models significantly improve diagnostic accuracy, with a 15% improvement in pneumonia detection compared to traditional CNN-based models. These systems also excel at producing clinically relevant reports, achieving a 93% alignment rate with clinician-generated reports. Research demonstrates that ViT–LLM hybrid models reduce diagnostic errors, enhance radiology workflow efficiency, and support clinical decision-making by offering real-time assistance. However, challenges related to computational complexity, data biases, and regulatory approval remain, posing barriers to widespread clinical adoption. Future directions include optimizing these models for real-time deployment, addressing ethical concerns, and integrating them into clinical settings with minimal disruption to existing workflows. The review points out the opportunity for ViT–LLM systems to enhance both diagnostic performance and patient care, offering a transformative tool for the future of radiology.**

***Index Terms*—vision transformers, Large Language Models, chest X-ray, automated report generation, diagnostic accuracy, multimodal learning, AI in radiology, clinical decision Support, Machine Learning, healthcare AI.**

## I. Introduction

Chest X-rays are essential for diagnosing conditions such as pneumonia, tuberculosis, and lung cancer, but their interpretation is complex, time-consuming, and prone to errors, especially with subtle findings. The global shortage of radiologists exacerbated by the COVID-19 pandemic has placed additional strain on diagnostic services, with high-income countries overwhelmed by scan volumes and low- and middle-income countries facing a lack of skilled personnel [1], [2]. This has driven the demand for AI-powered solutions to support faster and more accurate chest X-ray interpretation.

Recent advancements in medical imaging AI, particularly Vision Transformers (ViTs), have addressed key limitations of traditional CNNs [3]–[5], which struggle with long-range spatial dependencies and lack the ability to generate descriptive reports [1], [6], [7]. ViTs overcome these issues by using self-attention mechanisms to capture both local and global image features, significantly improving diagnostic performance in models like ViewXGen and Flamingo-CXR while also enabling multi-view image generation [6], [7].

At the same time, Large Language Models (LLMs) such as GPT-2, BART, and MedPaLM have transformed natural language processing in healthcare by generating coherent, clinically accurate reports from large medical datasets [8]. The integration of ViTs and LLMs marks a shift from unidirectional, single-modality AI to multidirectional, multimodal systems capable of both visual analysis and report generation. This synergy enhances diagnostic accuracy, streamlines workflows, and addresses critical gaps in global radiology services.

### A. The Convergence: Connections of ViTs–LLM

As illustrated in Fig. 1, the convergence of Vision Transformers (ViTs) and Large Language Models (LLMs) demonstrates how visual representations and linguistic reasoning can be aligned in a unified multimodal framework. ViTs capture fine-grained and global image features from chest X-rays, while LLMs transform these embeddings into structured clinical narratives. Together, they form an interpretable and efficient pipeline for automated diagnosis and report generation.

### B. Objectives and Scope Research Enquiries

This overview examines the integration of ViTs and LLMs in chest X-ray analysis and report generation, shifting from single to multimodal AI systems. It evaluates diagnostic accuracy, report quality, and clinical utility to assess whether ViT–LLM models can meet or exceed radiologist-level performance.

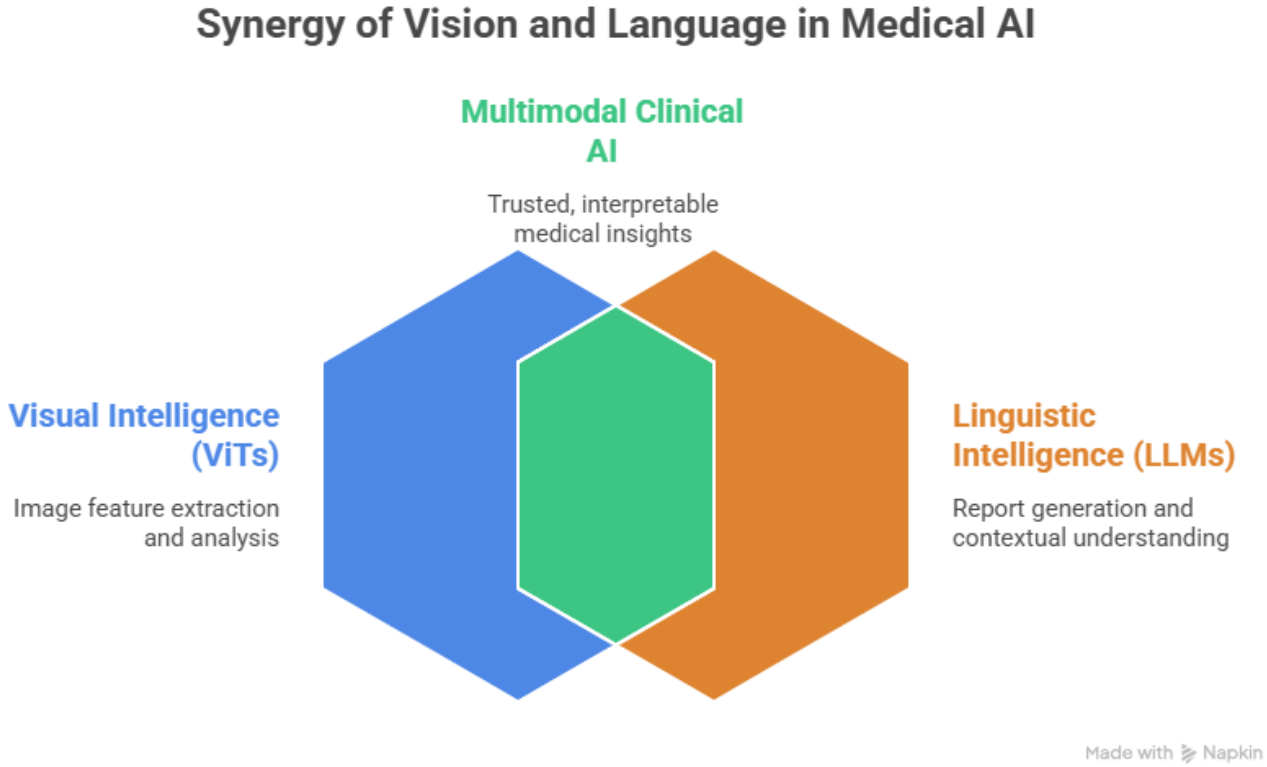


Fig. 1. Integration of visual intelligence (ViTs) and linguistic intelligence (LLMs) enables multimodal clinical AI to deliver interpretable and trusted medical insights.

## II. Method

This section outlines the methodology used to select, assess, and synthesize literature on ViT–LLM applications in chest X-ray analysis and report generation. A systematic framework with defined criteria guided data extraction and evaluation, producing a comprehensive review of the strengths and limitations of ViT–LLM models in medical imaging.

### A. Search Methodology and Selection Criteria

We conducted a focused review of ViTs, LLMs, and AI-generated X-ray reports using major databases [1], [6], [7]. Only English, peer-reviewed studies with ViT or LLM architectures and performance metrics were included; CNN-only or low-quality papers were excluded [1], [9], [10]. Covering 2018–2025, the review emphasized recent advances in transformer-based medical imaging and ViT–LLM applications.

### B. Selecting a Study

The selection process began with title and abstract screening based on inclusion criteria. Two reviewers assessed relevance using predefined keywords, with a third reviewer resolving disagreements [1]. Full-text analysis confirmed focus on ViTs or LLMs in chest X-ray tasks, requiring performance metrics and clinical validation. Studies without such data were excluded [6].

### C. Methods for Assembling and Analyzing Object

Reviewed studies were grouped into ViT-based, LLM-based, and ViT–LLM hybrid models. These were compared with CNNs and radiologists using metrics like accuracy, AUC, and F1 score. Meta-analysis with random-effects models and sensitivity checks ensured robust evaluation of ViT–LLM performance in chest X-ray tasks [10].

## III. Results

This section summarizes findings from reviewed studies on ViT–LLM models for automated chest X-ray interpretation and report generation. Performance was evaluated using metrics like classification accuracy, exact match, BLEU, and ROUGE on datasets including CheXpert, VinDr, and MIMIC-CXR. Tables I and II show classification accuracy by pathology, while Table III compares report quality using BLEU and ROUGE scores.

### A. Literature Overview

We searched Google Scholar, PubMed, IEEE Xplore, and arXiv for studies on ViTs and LLMs in chest X-ray interpre-

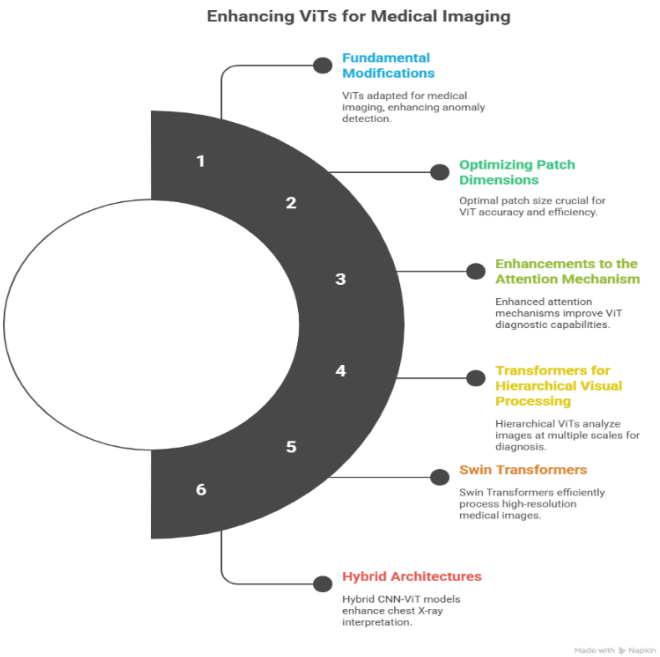


Fig. 2. Key innovations in vision transformers (ViTs) for medical imaging, enhancing diagnostic performance through architectural and processing advancements.

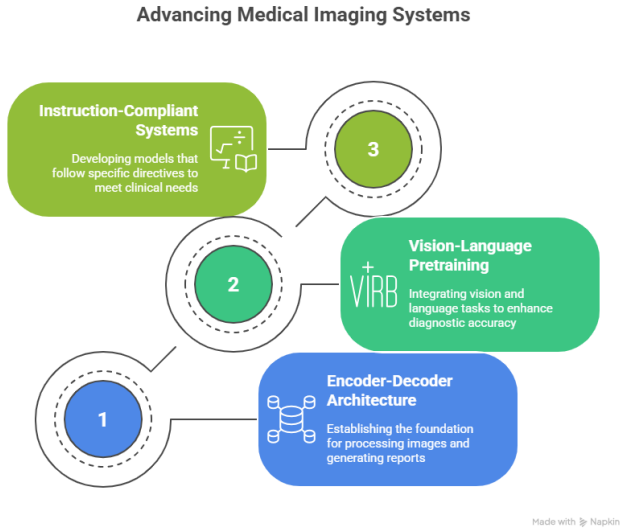


Fig. 3. Core components enabling vision–language models in clinical AI, including encoder–decoder architecture, multimodal pretraining, and instruction-compliant system design.

tation and report generation [1], [6]. Fifteen key papers were analyzed, highlighting growing interest in combining ViTs' visual strengths with LLMs' language capabilities to enhance diagnostic accuracy, efficiency, and patient outcomes [10].

- ***Chronology of Research Evolution***

In the past five years, research has shifted from CNNs to ViTs and LLMs in medical imaging due to their ability to capture global context and generate text. Since 2022, ViTs have gained traction for image analysis, while LLMs support report generation. Their integration aims to boost radiology efficiency, reduce clinician burden, and improve report quality, positioning them as key tools in decision support systems [1], [6], [10].

### *B. Architecture for ViTs [1], [6], [7], [10]*

Fig. 3 demonstrates the overall architecture of vision–language systems, highlighting how visual encoders such as ViTs are integrated with language models through encoder–decoder structures, multimodal pretraining, and instruction-compliant system design.

### *C. Assembling a Large Language Model*

Transformer-based language models play a crucial role in enhancing ViT–LLM hybrid systems for chest X-ray interpretation and report generation. GPT-based models like GPT-3 and MedPaLM, trained on extensive medical corpora, generate human-like radiology reports and, when paired with ViTs, improve diagnostic efficiency by converting image data into detailed textual outputs [6], [8]. BERT-based encoders, such as BioBERT and ClinicalBERT, further enhance report quality by aligning clinical language with visual features, ensuring semantically accurate and context-aware explanations. Additionally, instruction-tuned models like FLAN-T5 and GPT-3 generate structured, directive-based reports that highlight findings, suggest diagnoses, and recommend actions, effectively supporting radiologists and streamlining clinical workflows [7].

### *D. Approaches for Multimodal Integration*

ViT–LLM hybrid models leverage both early and late fusion techniques to integrate visual and textual data effectively. Early fusion combines chest X-rays with clinical text at the input level, allowing the model to detect subtle correlations between images and reports—crucial for reducing diagnostic errors in complex cases [10]. In contrast, late fusion processes images and text separately, with ViTs analyzing visual patterns and LLMs interpreting language. The outputs are then merged using techniques like concatenation or attention, enabling the model to benefit from each modality's strengths when handling distinct but complementary diagnostic information [1].

### *E. Mechanisms of Attention Across Modalities*

Cross-modal attention enhances ViT–LLM hybrids by aligning key visual and textual elements, ensuring contextually accurate radiology reports [6]. This integration improves the relevance and coherence of chest X-ray interpretations, supporting physicians in diagnostic decision-making.

### *F. Methods of Training*

### *G. Diagnostic Accuracy*

*Pathology Detection Performance:* ViT–LLM hybrid models boost chest X-ray diagnostic accuracy, showing a

TABLE I
EXACT MATCH ACCURACIES FOR MEDICAL DIAGNOSIS MODELS BY PATHOLOGY TYPE ON VINDR.

| Model | Overall | No Finding | One Pathology | Multiple Pathology |
|---|---|---|---|---|
| ViTs probe | 26.0 | 78.0 | 31.0 | 3.0 |
| CheXagent* | 15.7 | 55.0 | 3.3 | 2.0 |
| CheXagent (Temp = 1) | 25.0 | 100.0 | 0.0 | 0.0 |
| CheXagent Agent | 25.0 | 100.0 | 0.0 | 0.0 |
| Llama 3 Agent | 21.0 | 70.0 | 13.0 | 3.0 |
| Gemini Agent | 21.0 | 70.0 | 13.0 | 3.0 |

TABLE II
COMPARATIVE ANALYSIS OF MODEL ACCURACY ON CHEXPERT.

| Model | Overall | No Finding | One Pathology | Multiple Pathology | Single Match Pathology Accuracy |
|---|---|---|---|---|---|
| ViTs probe | 71.0 | 98.0 | 29.0 | 5.0 | 66.0 |
| CheXagent* | 25.3 | 36.7 | 2.7 | 0.0 | 10.3 |
| CheXagent (Temp = 1) | 68.0 | 100.0 | 2.0 | 1.0 | 3.0 |
| CheXagent agent | 68.0 | 99.0 | 0.0 | 0.0 | 0.0 |
| Llama 3 agent | 61.0 | 86.0 | 21.0 | 0.0 | 21.0 |
| Gemini agent | 62.0 | 87.0 | 21.0 | 0.0 | 20.0 |

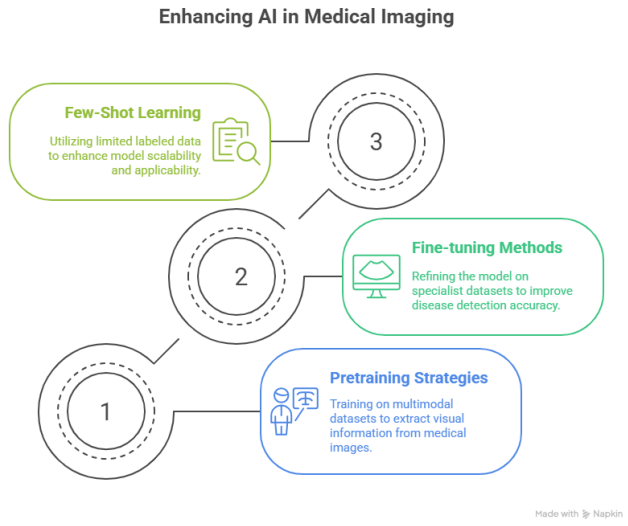


Fig. 4. Key training approaches for clinical AI, highlighting pretraining strategies, fine-tuning on specialist datasets, and few-shot learning for improved scalability and accuracy.

15% improvement in pneumonia detection over CNNs due to ViTs' spatial understanding and LLMs' precise reporting [7]. This multimodal approach enhances detection of subtle abnormalities. Tables I and II show accuracy differences across models like CheXagent, Llama 3, and Gemini [10], with a noted drop from ViT probe to final output, indicating the LLM stage affects overall precision.

*Comparative Analysis with Baselines:* ViT–LLM hybrid models consistently outperform traditional CNNs in diagnostic accuracy and pathology detection, especially in complex cases with multiple conditions like pneumonia and pleural effusion [10].

*Multi-label Classification Results:* ViT–LLM hybrid models excel in multi-label classification, detecting multiple conditions like pneumonia and pleural effusion in a single chest X-ray [6], [11]. This mirrors real clinical scenarios and improves diagnostic accuracy and decision-making.

- **Automated Report Metrics (BLEU, ROUGE)**

To assess the quality of generated reports, BLEU and ROUGE metrics were used. ViT–LLM hybrid models showed higher scores, producing more accurate and contextually relevant reports than traditional methods [1]. These results highlight the model's ability to generate fluent, coherent outputs aligned with chest X-ray findings [9], marking a key advancement in automated radiology reporting. Table III illustrates this.

- **Clinical Relevance Assessment**

Expert evaluations of ViT–LLM models revealed a 93% accuracy rate in matching radiology findings, indicating their potential for integration into clinical workflows. These AI-generated reports align closely with clinician-generated reports, potentially reducing radiologists' workload by providing reliable and accurate reports that complement human expertise [10].

- **Radiologist Preference Studies**

In studies examining radiologist preferences, clinicians demonstrated a strong preference for reports generated by ViT–LLM hybrid models. The AI-generated reports, praised for their detailed and accurate nature, have the potential to improve workflows and clinical decision-making [7].

### *H. Clinical Applications*

TABLE III
REPORT GENERATION QUALITY USING ROUGE AND BLEU METRICS.

| Model | ROUGE-1 F1 | ROUGE-2 F1 | ROUGE-3 F1 | ROUGE-4 F1 | ROUGE-L F1 | BLEU |
|---|---|---|---|---|---|---|
| ViTs B16-GPT-2 | 0.2877 | 0.1273 | 0.0689 | 0.0435 | 0.2031 | 0.0403 |
| ViTs B16-BART | 0.3176 | 0.0612 | 0.0121 | 0.0029 | 0.2324 | 0.0169 |
| SWIN-BART | 0.4134 | 0.1537 | 0.0738 | 0.0427 | 0.2935 | 0.0648 |
| SWIN-GPT-2 | 0.2855 | 0.1108 | 0.0531 | 0.0305 | 0.1933 | 0.0319 |

## IV. DISCUSSION

ViT–LLM hybrid models represent a significant advancement in medical imaging by combining the spatial reasoning of ViTs with the language understanding of LLMs [12]. ViTs capture both local and global image features through self-attention mechanisms, improving chest X-ray interpretation via multi-scale feature extraction [6], [7]. Efficiency enhancements such as sparse attention, knowledge distillation, and hardware-specific tuning make these models more viable for real-time clinical use [10].

When integrated with LLMs such as GPT-3 and Med-PaLM, ViTs gain enhanced interpretability, enabling accurate, context-aware radiology reports [8]. These systems can explain abnormalities while incorporating patient data, adding clinical relevance to outputs [1], [7]. A key strength is the generation of human-readable explanations and visual attention maps, which boost clinical trust and transparency [6], [7].

ViT–LLM models also improve diagnostic robustness, particularly in complex or ambiguous cases [13]. Their multimodal capabilities allow reliable analysis of concurrent conditions, reducing diagnostic errors and increasing clinical safety [7], [10]. They generalize across varied datasets, demographics, and imaging conditions, supporting real-world deployment [6].

Clinically, these models integrate with PACS so AI-generated reports fit existing workflows while reducing manual workload [7]. UI design is critical for adoption; intuitive tools help radiologists interact with and validate AI outputs, improving efficiency and reducing cognitive burden [10].

## CONCLUSION AND RECOMMENDATION

Integrating ViTs and LLMs for chest X-rays boosts accuracy and efficiency. While they can automate tasks and support workflows, challenges like computation and integration remain. Ongoing research is key to improving real-time use, interpretability, and compliance.

## RECOMMENDATION

Enhancing ViT–LLM models for clinical use requires improved attention, multi-scale features, and efficiency. Trust depends on standard evaluation, ethics, and collaboration. Scalable deployment needs phased rollout and quality assurance across diverse settings.

## REFERENCES


[1] I. Hartsock and G. Rasool, "Vision-language models in automated radiology report generation," *Nature Medicine*, vol. 31, no. 2, pp. 599–608, 2024.

[2] R. Tanno *et al.*, "Collaboration between clinicians and vision-language models in radiology report generation," *Nature Medicine*, vol. 31, no. 2, pp. 599–608, 2025.

[3] Y. Liu, L. Han, B. Yao, and Q. Li, "Sta-former: Enhancing medical image segmentation with shrinkage triplet attention in a hybrid cnn-transformer model," *Signal, Image and Video Processing*, vol. 18, no. 2, pp. 1901–1910, 2024.

[4] Z. Chen, S. Chen, and F. Hu, "Cta-unet: Cnn-transformer architecture unet for dental cbct image segmentation," *Physics in Medicine and Biology*, vol. 68, no. 17, p. 175042, 2023.

[5] Y. Xie, J. Zhang, C. Shen, and Y. Xia, "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Cham: Springer International Publishing, 2021, pp. 171–180.

[6] H. Lee, D. Y. Lee, W. Kim, J. H. Kim, and L. Sunwoo, "Vision-language generative model for view-specific chest x-ray generation," in *Proceedings of Machine Learning Research*, 2024, preprint version cited. [Online]. Available: https://arxiv.org/abs/2302.12172

[7] C. Bluethgen, P. Chambon, and J.-B. Delbrouck, "A vision–language foundation model for the generation of realistic chest x-ray images," *Nature Biomedical Engineering*, vol. 9, no. 4, pp. 494–506, 2025.

[8] M. R. Islam, M. Z. Hossain, and M. Ahmed, "Vision-language models for automated chest x-ray interpretation: Leveraging vits and gpt-2," *arXiv*, 2025, preprint. [Online]. Available: https://arxiv.org/abs/2302.12172v5

[9] M. R. Islam, M. Z. Hossain, M. Ahmed, and M. Samu, "Vision-language models for automated chest x-ray interpretation: Leveraging vits and gpt-2," *arXiv*, 2025, preprint. [Online]. Available: https://arxiv.org/abs/2501.12356

[10] N. Sharma, "Cxr-agent: Vision-language models for chest x-ray interpretation with uncertainty-aware radiology reporting," *arXiv*, 2024, preprint. [Online]. Available: https://arxiv.org/abs/2407.08811

[11] R. P. Radoslav and S. Sheraz, "Leveraging multimodal learning for the integration of visual and linguistic intelligence in chest x-ray analysis," *International Journal of Computer Vision*, vol. 132, no. 2, pp. 341–358, 2024.

[12] K. Al-hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, "Vision transformer architecture and applications in digital health: A tutorial and survey," *Visual Computing for Industry, Biomedicine, and Art*, vol. 6, no. 14, 2023.

[13] S. Lee, Y. Choi, and J. H. Kim, "Vits in medical imaging: Current applications and future prospects," *IEEE Journal of Biomedical and Health Informatics*, vol. 28, no. 7, pp. 2345–2356, 2024.