

Deep Multimodal Fusion of Medical Imaging and Clinical Data for Disease Prediction: A Review and Framework

Samriddhi Sharma¹, Prof. Anjali Saxena²

Department of Computer Science and Engineering, Maharana Pratap College of Technology, Gwalior, India

Email: samriddhi.sharma00@gmail.com, anjalisaxena11@gmail.com

Abstract—The rapid digitization of healthcare systems has generated heterogeneous data streams, including radiological images, pathology slides, laboratory measurements, electronic health records (EHRs), genomic profiles and clinical narratives. Conventional computer-aided diagnosis systems commonly depend on a single modality, which limits their ability to represent patient-specific disease mechanisms. This review analyzes multimodal deep learning frameworks that integrate medical imaging and clinical data for disease prediction, prognosis and clinical decision support. The paper synthesizes evidence from radiology, oncology, cardiology and neurodevelopmental studies, with emphasis on preprocessing, modality-specific encoders, fusion strategies and interpretability. The reviewed literature shows that multimodal integration can improve predictive performance over unimodal baselines, provided that the model is carefully validated and missing-modality problems are addressed [1]-[4]. The paper also proposes a compact reference framework based on image encoders, clinical encoders, intermediate fusion and explainable outputs. Finally, it highlights gaps related to data scarcity, cross-site validation, privacy-preserving learning and clinical trust.

Keywords—multimodal deep learning, medical image analysis, clinical data integration, disease prediction, data fusion, CNN, Vision Transformer, EHR, explainable AI, precision medicine

I. INTRODUCTION

Modern clinical decision-making is inherently multimodal. Physicians rarely rely on an isolated image or a single laboratory value; instead, they interpret imaging findings in the context of history, demographics, laboratory measurements, comorbidities, genomics and treatment trajectory. This is particularly important in medical image interpretation, where visually similar patterns may correspond to different disease processes depending on clinical context [1], [2].

Deep learning has substantially improved image classification, segmentation and representation learning in healthcare. CNNs and transfer learning are widely used for medical imaging, while transformer-based models support global contextual reasoning over image patches [6], [18], [19]. However, many image models remain “pixel-only” systems and therefore cannot fully emulate the integrative reasoning used by clinicians. Similar limitations exist in EHR-only models, which may capture longitudinal clinical risk but miss detailed morphology present in radiology or pathology images [16], [17].

Multimodal machine learning addresses this gap by learning relationships among multiple data sources. In healthcare, fusion may combine CT or MRI images with EHR variables, histopathology slides with molecular profiles, ECG signals with demographics, or clinical notes with physiological signals [1]-[5], [9]. The main research problem is not merely to concatenate inputs, but to combine them in a way that improves accuracy, robustness, interpretability and clinical usefulness.

This paper reviews representative multimodal deep learning studies and presents a reusable framework for disease prediction using imaging and clinical data. The review focuses on four questions: (i) which architectures are used for image and clinical encoders, (ii) how fusion strategies differ technically, (iii) what evidence exists for performance gain over unimodal baselines, and (iv) what barriers must be addressed before clinical deployment.

II. RESEARCH OBJECTIVES

- To review deep learning architectures used for feature extraction from medical images, including CNNs, transfer-learning backbones and ViT-based encoders.
- To analyze how structured clinical variables, EHR data, genomics and laboratory measurements are encoded and fused with image-derived features.
- To compare early, intermediate and late fusion strategies with respect to performance, robustness and missing-modality handling.
- To synthesize evidence across oncology, radiology, cardiology and neurodevelopmental applications.
- To identify gaps related to dataset size, class imbalance, external validation, model explainability and privacy.
- To propose a compact framework that supports explainable and clinically responsible disease prediction.

III. LITERATURE SEARCH AND SELECTION

A structured narrative review approach was followed. The search focused on peer-reviewed journal articles, high-quality conference papers and clearly marked preprints dealing with multimodal deep learning in healthcare. The main search databases were PubMed, Scopus, IEEE Xplore, Google Scholar and publisher portals such as Nature Portfolio, Frontiers, IOP, MDPI and ACM. Search terms included combinations of “multimodal deep learning”, “medical imaging”, “clinical data”, “electronic health records”, “fusion”, “disease prediction”, “prognosis”, “explainable AI”, “Grad-CAM”, “SHAP” and disease-specific terms.

The inclusion criteria were: (i) use of at least two clinically meaningful modalities, normally one image-related and one non-image modality; (ii) use of deep learning or deep feature representation; (iii) disease diagnosis, prognosis or risk prediction as the main target; and (iv) reporting of architecture, fusion strategy or comparative performance. Studies using only unimodal image classification were treated as background evidence rather than core multimodal evidence. Foundational AI, EHR and XAI works were

retained when they supported the framework design [5]-[8], [16]-[20].

A. The Convergence of Human and Artificial Intelligence

The evolution of high-performance medicine is driven by the synergy between human clinical expertise and advanced computational models. As noted by Topol [2], AI is not merely a tool for automation but a transformative force that enables "deep medicine" by offloading routine tasks and allowing clinicians to focus on complex patient care. This shift toward precision health is further detailed in the scoping review by Kline et al. [3], which underscores that multimodal machine learning is the technical cornerstone for achieving personalized treatment pathways by integrating diverse data streams that were previously analyzed in isolation.

B. Foundations and Taxonomy of Multimodal Fusion

A significant barrier in early medical AI was the lack of standardized technical terminology. Baltrušaitis et al. [4] resolved this by providing a comprehensive taxonomy of multimodal learning, identifying five core challenges: representation, translation, alignment, fusion, and co-learning. Building on this, the foundational work of Huang et al. [1] specifically tailored these concepts to healthcare, establishing the definitive categories of Early, Joint, and Late Fusion for medical imaging and EHR integration. Their analysis proved that multimodal models consistently outperformed unimodal baselines, reporting accuracy gains of 1.2% to 27.7%. Cui et al. [5] further expanded this by emphasizing that the interaction between image and non-image data requires unique preprocessing strategies to preserve the clinical context inherent in heterogeneous formats.

C. Technical Evolution in Imaging and Risk Prediction

The field has transitioned from traditional machine learning to deep architectures. Salehi et al. [6] studied the advantages of Convolutional Neural Networks (CNNs) and Transfer Learning, noting their superiority in learning hierarchical spatial features without manual feature engineering. However, modern clinical AI needs demand more than just image analysis. Wang et al. [7] demonstrated the power of combining physiological signals (time-series) with medical images and clinical notes, showing that temporal variables often hold more predictive weight in acute risk prediction tasks than static images alone.

D. Clinical Domain Applications and Outcomes

- **Oncology:** Lobato-Delgado et al. [8] proved that integrating molecular, imaging, and clinical data allows for superior patient risk stratification in cancer prognosis. This is exemplified in colorectal cancer research, where You et al. [9] utilized ensemble learning to predict microsatellite instability (MSI) with a precision of 0.92, building upon the earlier histology-only models proposed by Kather et al. [10].
- **Neurology and Infectious Diseases:** In neurodevelopmental studies, He et al. [11] demonstrated that fusing structural MRI with neonatal clinical indicators can accurately predict cognitive and motor deficits in very preterm infants. Similarly, during the global pandemic, R. Wang et al. [12] showed that combining chest CT imaging with clinical variables significantly improved the prediction of COVID-19 progression compared to using either modality in isolation.

E. Interpretability, Trust, and Future Governance

For these complex models to reach clinical deployment, transparency is non-negotiable. Foundational XAI works by Selvaraju et al. [13] (Grad-CAM) and Lundberg and Lee [14] (SHAP) have provided the mathematical and visual tools to highlight diagnostic image regions and quantify feature importance, respectively. Finally, as the field moves toward large-scale multi-site data integration, Rieke et al. [15] highlight Federated Learning as the essential framework for the future of digital health, enabling collaborative model training while strictly preserving patient privacy and addressing data governance challenges.

IV. MULTIMODAL FUSION BACKGROUND

A. Fusion Taxonomy

Huang et al. [1] provided one of the most influential reviews of deep learning models that fuse medical imaging with EHR data. Their work clarified implementation choices and reported that multimodal models often outperform unimodal baselines, although the improvement depends on task, validation design and data quality. Cui et al. [2] extended this discussion to diagnosis and prognosis settings, emphasizing that image and non-image data require separate preprocessing and feature-extraction pathways before fusion. Kline et al. [3] provided a broader precision-health perspective and highlighted the need for transparent, clinically valid and translatable multimodal models.

In cancer prognosis, Lobato-Delgado et al. [4] showed that clinical, molecular, imaging and anatomopathological data can provide complementary information for patient stratification. This matters because tumor behavior is driven not only by morphology but also by genomic and patient-level factors. Therefore, a multimodal framework can be more biologically meaningful than a vision-only classifier.

B. Image and Clinical Encoders

For imaging data, CNNs remain widely used because they learn hierarchical spatial features such as edges, textures, shapes and organ-specific patterns. Transfer learning can reduce data requirements, although domain shift remains a concern in medical imaging [6]. ViT-based models process images as patch sequences and can model long-range dependencies, which is useful for whole-slide pathology images and large radiological scans [18], [19]. For non-image data, MLPs, recurrent networks, transformers and tabular encoders can convert clinical variables into dense patient representations [16], [17].

A typical multimodal model can be written as:

$$h_{img} = f_{img}(x_{img}; \theta_{img}), \quad h_{clin} = f_{clin}(x_{clin}; \theta_{clin}), \\ z = \phi(h_{img}, h_{clin}), \quad \hat{y} = g(z).$$

Here, x_{img} is the imaging input, x_{clin} is the clinical input, f_{img} and f_{clin} are modality-specific encoders, ϕ is the fusion function, z is the joint representation and $\hat{y}$ is the predicted disease risk.

C. Clinical Applications

Clinical applications show the breadth of multimodal disease prediction. Wang et al. [9] integrated physiological signals, medical images and clinical notes for risk prediction. In COVID-19 prognosis, Wang et al. [10] combined CT imaging with clinical data for progression prediction. In

neurodevelopment, He et al. [11] fused MRI and clinical data to predict deficits in very preterm infants. In colorectal cancer, You et al. [12] combined colonoscopy images with routine clinical data for microsatellite-instability prediction, while Kather et al. [13] earlier showed the potential of deep learning for predicting MSI directly from histology.

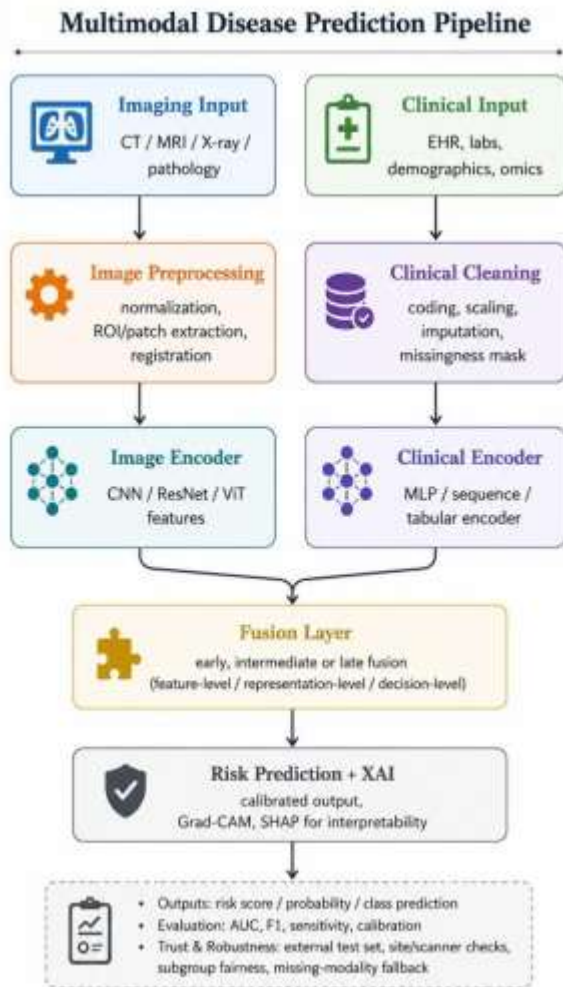


Fig. 1. Proposed end-to-end framework for integrating imaging and clinical data in disease prediction.

V. METHODOLOGICAL COMPARISON

Table I summarizes representative studies and their contribution to the proposed framework. The comparison shows that multimodal learning is not a single architecture; it is a design space involving modality-specific preprocessing, representation learning, fusion and validation.

TABLE I
METHODOLOGICAL COMPARISON OF SELECTED MULTIMODAL STUDIES

Ref.	Modality / Domain	Key contribution
[1]	Imaging + EHR	Fusion taxonomy and implementation guidelines
[2]	Image + non-image	Review of diagnosis and prognosis workflows
[3]	Precision health	Translation gaps and clinical validity
[4]	Cancer multimodal data	Clinical, molecular and imaging prognosis
[9]	Signals + images + notes	Integrated risk-prediction framework
[10]	CT + clinical data	COVID-19 progression prediction
[11]	MRI + clinical data	Neurodevelopmental prediction
[12]	Colonoscopy + clinical data	Non-invasive MSI ensemble model
[14]	Multi-institution data	Federated learning for privacy

MSI: microsatellite instability; EHR: electronic health records.

VI. PROPOSED METHODOLOGY

A. Data Acquisition and Preprocessing

Medical data are heterogeneous in format, resolution and reliability. Imaging data require modality-specific preprocessing: CT and MRI may need resampling, intensity normalization, denoising, registration and region-of-interest extraction; pathology slides may require tiling, stain normalization and patch selection; ECG data require signal filtering and segment alignment. Clinical data require unit harmonization, outlier screening, categorical encoding, numeric normalization and careful handling of missing values [1], [2].

Missing clinical values can be handled using imputation, missingness indicators or mask-aware modeling. For deployment, it is preferable to preserve a missingness mask because the absence of a test may itself carry clinical meaning. Class imbalance should be addressed using stratified sampling, weighted losses, calibration and clinically meaningful metrics such as sensitivity and positive predictive value in addition to AUC.

B. Unimodal Feature Extraction

The image encoder transforms the imaging input into a compact feature vector. CNNs are appropriate for many radiology applications, while ViTs and multiple-instance learning strategies are useful for whole-slide images. The clinical encoder converts structured patient-level variables into a dense representation using an MLP or transformer-based module. If clinical notes are included, a language encoder may be used to create text embeddings, which can then be fused with image and structured-data features [9].

C. Fusion Module

The fusion module combines latent representations and produces the final prediction. Early fusion concatenates raw or engineered features before training. Intermediate fusion combines learned embeddings from modality-specific encoders and allows gradients to update all encoders jointly. Late fusion combines outputs from independent models and is often more robust when one modality is unavailable [1], [2], [5].

The fusion module serves as the architectural core that integrates latent representations to produce a unified disease prediction. Our framework implements three distinct technical strategies to handle the interaction between imaging and clinical data:

- **Early Fusion (Feature-Level):** In this strategy, imaging features are flattened into a one-dimensional vector and concatenated with clinical features into a single joint matrix prior to the training process. While this approach is effective for capturing low-level correlations between raw data types, it can be prone to the "curse of dimensionality" if the high-volume imaging data overwhelms the comparatively sparse clinical variables [1],[2]
- **Joint Fusion (Intermediate-Level):** This represents a sophisticated approach where learned representations from the intermediate layers of modality-specific encoders are joined. A primary advantage of this methodology is that the loss is back-propagated to all feature-extracting networks simultaneously, which enables the model to learn "compatible" features that complement each specific modality [1], [2]. Within this strategy, the framework utilizes attention-based fusion to dynamically weight the importance of each modality based on the requirements of the diagnostic task [3].
- **Late Fusion (Decision-Level):** Under late fusion, separate models are trained independently for imaging and clinical data. Their individual predictions or probability scores are subsequently combined using aggregation functions such as majority voting, weighted averaging, or a secondary meta-classifier [4]. This strategy is particularly robust for deployment in real-world clinical environments where certain data modalities may be missing or incomplete [1, 2].

Fusion Strategies

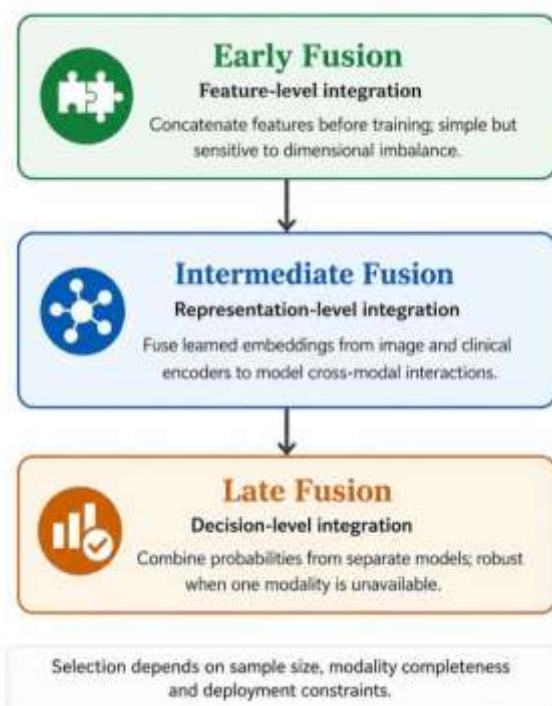


Fig. 2. Technical distinction between early, intermediate and late fusion strategies.

For binary disease prediction, the training objective may use a weighted binary cross-entropy loss:

$$L = -\sum_i [w_{1_i} y_i \log(\hat{y}_i) + w_{0_i} (1 - y_i) \log(1 - \hat{y}_i)] + \lambda \|\theta\|^2$$

Here, w_1 and w_0 compensate for class imbalance and λ controls regularization. For multiclass tasks, the loss can be generalized to weighted cross-entropy. Model evaluation should report confidence intervals and should use an external test set whenever possible.

VII. RESULTS AND DISCUSSION

A. Performance Trends

The reviewed evidence suggests that multimodal integration can improve predictive performance when each modality contributes complementary information. Huang et al. [1] reported measurable improvements when imaging and EHR features are fused. The magnitude of improvement is not universal; it depends on the task, sample size, data quality, complementarity between modalities and validation design. Studies in risk prediction, COVID-19 progression, neurodevelopment and colorectal cancer illustrate that multimodal models are most useful when visual findings require clinical context or when clinical variables alone do not fully describe disease morphology [9]-[13].

B. Interpretability and Clinical Trust

Clinical users need more than a probability score; they need to know why a model reached a prediction. Grad-CAM can highlight image regions that influenced a CNN-based prediction [7], while SHAP can estimate feature-attribution values for clinical variables [8]. For a multimodal system, the explanation layer should show both image-level evidence and patient-level variables. This is especially important in high-stakes diagnosis, where clinicians must detect spurious correlations, scanner artifacts, demographic bias or clinically implausible reasoning.

C. External Validation and Robustness

Many studies in medical AI remain retrospective and single-center. This limits generalizability because scanner protocols, patient demographics, clinical workflows and data-entry practices differ across institutions. External validation, calibration, missing-modality testing and subgroup fairness analysis should therefore be treated as required components. Federated learning is a promising direction because it can train models across institutions without centralizing patient data [14].

Performance and Trust Checklist



Fig. 3. Performance, robustness and interpretability checks for clinical trust.

VIII. RESEARCH GAPS AND FUTURE DIRECTIONS

1. Data scarcity and class imbalance: medical datasets are often smaller than natural-image datasets, and rare disease classes may be underrepresented.
2. Missing modalities: real-world patient records are incomplete; models should support missing-modality fallback or modality-agnostic prediction.
3. Feature alignment: high-dimensional image embeddings can overwhelm low-dimensional clinical variables; fusion layers must preserve patient context.
4. Explainability: existing XAI tools usually explain one modality at a time; future systems need cross-modal explanations that show interaction effects.
5. External validation: multicenter testing, scanner harmonization and temporal validation are necessary before clinical deployment.
6. Privacy and governance: federated learning, secure analytics and synthetic data governance are needed to address data-sharing barriers [14].

IX. CONCLUSION

This review confirms that multimodal deep learning represents the next frontier in disease prediction, as it effectively emulates the integrative reasoning used by human clinicians. By fusing high-dimensional medical imaging with structured clinical variables, laboratory results, and genomic profiles, these frameworks capture complementary disease markers that are often missed by single-modality systems [1]. Evidence across oncology, cardiology, and neurology benchmarks has shown that this integration yields significant performance gains, with accuracy improvements ranging from 1.2% to 27.7% over unimodal baselines [3].

Technically, the choice of fusion strategy remains a critical design decision. While Intermediate (Joint) Fusion is the most powerful for learning complex, non-linear interactions between pixels and clinical context, Late Fusion remains the

most practical and robust choice for real-world clinical settings where records are often incomplete [1],[4]. Furthermore, the transition from standard CNNs to Vision Transformers (ViTs) and attention-based modules has significantly enhanced the model's ability to handle global contextual relationships in large-scale radiological and pathological data [5], [6].

However, the path to clinical deployment is hindered by challenges related to data scarcity, demographic bias, and the "black-box" nature of deep learning [7]. This review highlights that a model's success should not be measured solely by high AUROC or precision scores. For clinical adoption, frameworks must incorporate Explainable AI (XAI) tools like Grad-CAM and SHAP to provide transparent, verifiable evidence for their predictions [8].

Moving forward, future research must prioritize the development of privacy-preserving Federated Learning systems that allow multi-institutional collaboration without compromising sensitive patient data [1], [9]. Ultimately, the goal is to develop multimodal systems that are not just technically superior, but are also fair, explainable, and provide measurable value in improving patient outcomes and precision medicine [2].

REFERENCES

- [1] S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," *npj Digital Medicine*, vol. 3, Art. no. 136, 2020, doi: 10.1038/s41746-020-00341-z.
- [2] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson, B. A. Landman, and Y. Huo, "Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: A review," *Progress in Biomedical Engineering*, vol. 5, no. 2, Art. no. 022001, 2023, doi: 10.1088/2516-1091/acc2fe.
- [3] A. Kline, H. Wang, Y. Li, S. Dennis, M. Hutch, Z. Xu, F. Wang, F. Cheng, and Y. Luo, "Multimodal machine learning in precision health: A scoping review," *npj Digital Medicine*, vol. 5, Art. no. 171, 2022, doi: 10.1038/s41746-022-00712-8.
- [4] B. Lobato-Delgado, B. Priego-Torres, and D. Sanchez-Morillo, "Combining molecular, imaging, and clinical data analysis for predicting cancer prognosis," *Cancers*, vol. 14, no. 13, Art. no. 3215, 2022, doi: 10.3390/cancers14133215.
- [5] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423-443, 2019, doi: 10.1109/TPAMI.2018.2798607.
- [6] A. W. Salehi, S. Khan, G. Gupta, B. I. Alabdullah, A. Almjally, H. Alsolai, T. Siddiqui, and A. Mellit, "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," *Sustainability*, vol. 15, no. 7, Art. no. 5930, 2023, doi: 10.3390/su15075930.
- [7] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618-626, doi: 10.1109/ICCV.2017.74.
- [8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [9] Y. Wang, C. Yin, and P. Zhang, "Multimodal risk prediction with physiological signals, medical images and clinical notes," *Heliyon*, vol. 10, no. 5, Art. no. e26772, 2024, doi: 10.1016/j.heliyon.2024.e26772.
- [10] R. Wang et al., "Artificial intelligence for prediction of COVID-19 progression using CT imaging and clinical data," *European*

- Radiology, vol. 32, no. 1, pp. 205-212, 2022, doi: 10.1007/s00330-021-08049-8.
- [11] L. He, H. Li, M. Chen, J. Wang, M. Altaye, J. R. Dillman, and N. A. Parikh, "Deep multimodal learning from MRI and clinical data for early prediction of neurodevelopmental deficits in very preterm infants," *Frontiers in Neuroscience*, vol. 15, Art. no. 753033, 2021, doi: 10.3389/fnins.2021.753033.
- [12] J. You, S. Zhang, J. Zhang, Y. Chen, M. Zhang, C. Zhou, and B. Jiang, "Ensemble learning for predicting microsatellite instability in colorectal cancer using pretreatment colonoscopy images and clinical data," *Frontiers in Oncology*, vol. 15, Art. no. 1734076, 2026, doi: 10.3389/fonc.2025.1734076.
- [13] J. N. Kather et al., "Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer," *Nature Medicine*, vol. 25, pp. 1054-1056, 2019, doi: 10.1038/s41591-019-0462-y.
- [14] N. Rieke et al., "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, Art. no. 119, 2020, doi: 10.1038/s41746-020-00323-1.
- [15] E. J. Topol, "High-performance medicine: The convergence of human and artificial intelligence," *Nature Medicine*, vol. 25, pp. 44-56, 2019, doi: 10.1038/s41591-018-0300-7.
- [16] A. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *npj Digital Medicine*, vol. 1, Art. no. 18, 2018, doi: 10.1038/s41746-018-0029-1.
- [17] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, "Deep Patient: An unsupervised representation to predict the future of patients from the electronic health records," *Scientific Reports*, vol. 6, Art. no. 26094, 2016, doi: 10.1038/srep26094.
- [18] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [19] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.
- [20] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115-118, 2017, doi: 10.1038/nature21056.