Research and Analysis of Intelligent Diagnosis Technology in Medical Imaging

Xingyue Chen [1, *], Pinxi Lin [2] and Yuhang Liu [3]

[1] School of Electrical and Automation Jiangsu Ocean University, Lianyungang, China

[2] Fuzhou Gezhi Middle School, Fuzhou, China

[3] College of Publishing, Printing, and Art Design, University of Shanghai for Science and Technology, Shanghai, China

* Corresponding Author Email: XingyueChen6688@gmail.com

Abstract. As a core basis for clinical diagnosis, medical imaging holds crucial information for disease screening, classification, and treatment efficacy evaluation. However, traditional manual image interpretation faces three major bottlenecks: low efficiency, limited capability, and uneven resource distribution. With the popularization of precision medicine and exponential growth of imaging data, intelligent medical imaging diagnosis has become vital to optimizing the diagnostic process. For example, the ViT-VGG16 hybrid architecture enables collaborative modeling of "local details-global correlations" in small-sample scenarios; BoTNet balances multi-scale complex lesion modeling with hardware compatibility; and the CNN-dictionary representation hybrid builds a "global-local" dual feature system to improve diagnostic accuracy. Due to new diseases and more complex medical needs, this field is evolving, with intelligent technologies addressing challenges like small-sample learning, data privacy, and workflow disconnection. This review compiles relevant literature with two key goals: first, systematically sort out and analyze mainstream intelligent technologies, their principles, and performance; second, examine technical value, significance, dataset characteristics, and propose clinical implementation solutions. These analyses not only guide researchers in technological innovation but also provide evidence-based references for clinicians to promote intelligent technology in frontline practice, while serving as an empirical basis for governments to formulate medical AI development strategies.

Keywords: Medical imaging, Convolutional Neural Network, Brain tumor, BoTNet, Artificial Intelligence.

1. Introduction

As a core basis for clinical diagnosis, medical imaging encompasses multimodal data such as CT, MRI, and pathological sections. The lesion morphology, tissue texture, and dynamic functional information it carries are crucial for disease screening, classification, and treatment efficacy evaluation. In recent years, with the popularization of the concept of precision medicine and the increased penetration of imaging equipment, the global volume of medical imaging data has experienced exponential growth. According to the Report on Artificial Intelligence + Intelligent Medical Imaging Analysis, the annual increase in imaging cases at Xinxiang Central Hospital in China can reach hundreds of thousands. However, the traditional manual image interpretation model is facing three significant and unavoidable bottlenecks. First, an efficiency bottleneck—radiologists need to analyze hundreds of images per day on average, leading to prominent missed diagnoses caused by fatigue; second, a capability bottleneck—diagnosing complex diseases requires capturing cross-regional correlations, and it is difficult for humans to accurately model such long-range dependencies; third, a resource bottleneck—due to insufficient physician experience in primary medical institutions, the diagnostic accuracy for some diseases is lower than that in tertiary hospitals, exacerbating contradictions in the equity of medical services.

Early research focused on Convolutional Neural Networks (CNNs) as the core. Models such as VGG16 and ResNet, relying on the "multi-layer convolution + pooling" architecture, can accurately capture fine-grained structures in images (e.g., the edge of hemorrhage in stroke CT scans). However,

their limitations are notable: global long-range dependencies can only be achieved indirectly by stacking dozens of convolutional layers, which results in high computational cost and prone to feature loss—for example, it is difficult to correlate distant lesions in whole-body PET-CT scans. Additionally, these models rely on manual annotation; annotating a single brain tumor MRI scan takes more than 25 minutes [1].

Vision Transformers (ViTs), leveraging the self-attention mechanism, can directly model dependencies between arbitrary image regions, addressing the global limitation of CNNs. Nevertheless, ViTs have obvious drawbacks: they lack the "inductive bias" of CNNs, making them prone to missing subtle changes in early pancreatic cancer; moreover, they require millions of pre-trained samples [2], which conflicts with the reality that most diseases in the medical field have only thousands of samples.

Through qualitative analysis of existing literature, this review systematically sorts out the principles and performance of mainstream technologies in intelligent medical imaging diagnosis, analyzes their value, significance, and dataset characteristics, summarizes clinical application effects, and proposes solutions. It aims to provide references for technological innovation and clinical translation, facilitate the implementation of these technologies in frontline diagnosis and treatment, and improve diagnostic accuracy and medical equity.

2. Overview of Mainstream Methods

2.1. ViT and VGG16

Vision Transformer (ViT) lacks the inductive bias (e.g., translation invariance) of Convolutional Neural Networks (CNNs), leading to significant performance degradation on small datasets—even data augmentation techniques fail to improve its performance. In contrast, the convolutional layers of the Visual Geometry Group 16-layer Convolutional Neural Network (VGG16) inherently possess visual prior knowledge, enabling efficient learning of local feature patterns from limited data.

The core principle of the ViT-VGG16 hybrid architecture is to achieve collaborative modeling of "local details-global correlations" in medical images through functional complementarity [3]. Serving as the front-end feature extraction module, VGG16 adopts a hierarchical architecture with multi-layer 3×3 convolution and pooling operations. It captures local features such as lesion edges and textures via sliding-window convolution—for example, accurately extracting calcified textures in bone tumor CT scans and identifying the morphology of hemorrhage points in stroke images. Its convolutional layer parameter sharing mechanism and pooling operations effectively preserve key spatial details, laying a foundation for subsequent analysis.

The hybrid architecture adopts a "CNN feature pre-extraction + lightweight ViT modeling" pattern. It uses the feature selection results of VGG16 as input constraints for ViT, significantly reducing dependency on annotated data. In bone tumor diagnosis, this architecture achieves excellent performance with only 1,244 cases, whereas a pure ViT typically requires millions of samples to unleash its potential. This advantage makes it particularly suitable for small-sample scenarios such as rare diseases. The secondary transfer learning strategy further enhances this characteristic: by reusing pre-trained features, it reduces the risk of model overfitting on small datasets.

2.2. Hybrid CNN and Dictionary Representation

In cardiac CT image analysis, the CNN encoder—with the mature advantages of U-Net in the field of medical image segmentation—serves as the core preprocessing module. U-Net features a symmetric architecture of "encoder down sampling for feature extraction + decoder up sampling for resolution restoration," combined with skip connections to retain shallow spatial details [4]. This enables it to accurately capture the hierarchical features of cardiac structures: it can not only extract macroscopic information such as the overall morphology of the myocardium and the contour of heart chambers but also identify suspected lesion clues like abnormal coronary artery courses and uneven myocardial density. The final output high-level semantic feature maps can automatically frame

multiple high-risk regions, achieving a localization accuracy of over 89% and narrowing the scope for subsequent detailed analysis [5].

The dictionary module focuses on mining microscopic features of suspected regions. First, based on a dedicated cardiac CT dataset, it uses the K-SVD algorithm to learn sparse atoms of local features (e.g., vascular wall textures and myocardial density gradients), constructing a dedicated dictionary library [6]. Next, it performs sparse coding on 64×64-pixel image patches in the regions localized by the CNN, compressing high-dimensional image information into low-dimensional sparse feature vectors. These vectors can accurately characterize subtle features such as the lipid core proportion of atherosclerotic plaques and the endothelial roughness of blood vessels, compensating for the CNN's insufficiency in capturing microstructures.

Finally, through feature concatenation or attention-weighted fusion strategies, the global structural features of the CNN and the local sparse features of the dictionary are integrated and input into a classifier to determine lesion nature. This improves overall diagnostic accuracy and reduces the false positive rate, providing clinicians with more comprehensive imaging evidence for formulating interventional treatment plans.

2.3. BoTNet

The core method of Bottleneck Transformer Network (BoTNet) is to retain the early convolutional layers of ResNet while replacing the 3×3 convolutions in its last three bottleneck blocks with the Multi-Head Self-Attention (MHSA) mechanism [7].

In the task of brain tumor segmentation, the model first performs local feature extraction on multimodal MRI images through its CNN layers, constructing a multi-scale feature pyramid [8]. Then, in the deep network, it calculates the key-value pair relationships between all spatial positions in the feature maps and generates attention weight maps. These weight maps can automatically learn and highlight regions more relevant to the target. For feature maps of different resolutions: MHSA in low-resolution feature maps focuses on capturing the overall shape and spatial information of the tumor; in high-resolution feature maps, it focuses on finely identifying subtle texture changes at the tumor edge, the tumor's internal structure, and its interaction with surrounding tissues.

3. Value and Significance

3.1. Optimizing Global Modeling Capability for Complex Lesions

The core challenge in diagnosing complex lesions lies in the scattered spatial distribution of lesions, complex correlations among multimodal features, and blurred boundaries between lesions and normal tissues. Traditional Convolutional Neural Networks (CNNs) rely on stacked local convolutions to indirectly capture long-range dependencies, which suffer from computational redundancy and low accuracy in global correlation modeling. Optimizing the global modeling capability for complex lesions—by developing innovative architectures to accurately capture global lesion correlations while efficiently controlling computational consumption and holds both significant clinical value and technical breakthrough significance.

From a clinical application perspective, this optimization can substantially improve the accuracy and practicality of complex lesion diagnosis. For example, in the recognition of multi-regional correlations among the core area, edema area, and necrotic cavity in gliomas, the BoTNet-based hybrid architecture retains the local detail extraction capability of CNNs while directly modeling the spatial dependencies between lesion regions via a deep Multi-Head Self-Attention (MHSA) module. This enhances the detection rate of glioma sub-lesions and avoids missed diagnoses of tiny sub-lesions caused by limited local field of view.

For the evaluation of functional correlations between coronary artery stenosis and myocardial blood supply regions, a hierarchical attention design is adopted: the low-resolution layer captures the global mapping relationship between stenotic blood vessels and blood-supplying myocardium [9], while the high-resolution layer focuses on detailed features of stenosis edges. This not only improves

the accuracy of stenosis severity assessment but also, through lightweight optimization, reduces the inference time of the model on NVIDIA Jetson AGX Orin devices (commonly used in primary medical institutions), meeting the needs of real-time clinical diagnosis and addressing the issue of pure ViT models being difficult to implement in practice.

3.2. Improving Imaging Feature Modeling Dimensions

The core value of the hybrid CNN and dictionary representation lies in constructing a "global-local" dual feature capture system, which compensates for the inherent shortcomings of single models in medical imaging feature extraction. Leveraging its hierarchical convolutional architecture, CNNs excel at capturing the overall structural features of medical images from a macroscopic perspective—such as the overall morphology of the heart and the general contour of tumors—providing a global framework for lesion localization.

In contrast, dictionary representation, through sparse coding of local image patches [10], can accurately mine subtle features from a microscopic perspective, such as texture changes in blood vessel walls and differences in myocardial tissue density gradients. This collaborative "global control + local focus" mode enables the model to not only grasp the overall structural logic of images but also capture easily overlooked microscopic lesion clues, forming a more comprehensive feature description system. This addresses the issues where a single CNN lacks sensitivity to microscopic details and a single dictionary model struggles to establish global correlations.

3.3. Enhancing Compatibility and Efficiency of Technology Implementation

BoTNet is modified based on the mature ResNet architecture rather than being a fully reconstructed model framework, a design that endows it with significant technical compatibility advantages. On one hand, it can directly reuse the pre-trained weights and technical accumulations of ResNet in the field of medical imaging, avoiding model training from scratch and greatly shortening the R&D cycle. On the other hand, its architectural modifications are concentrated in the deep network, and the retention of shallow convolutions makes it more computationally efficient than pure Transformer models— it can operate without relying on hardware devices with extremely high configurations.

This balance between compatibility and efficiency lowers the R&D threshold and deployment costs of medical imaging AI technology. It is particularly suitable for scenarios with limited hardware resources, such as primary medical institutions, and provides convenience for the large-scale promotion and clinical implementation of the technology.

4. Dataset characteristics

4.1. Adapt to small sample training scenarios

The ViT and VGG16 hybrid architecture dataset is currently mainly based on the "Brain Stroke CT Image Dataset", which provides standardized training samples for stroke CT image analysis. Although the sample size of the data set does not reach the scale of one million, the sample distribution is consistent with the true proportion of clinical stroke incidence, including both typical cases and cases prone to missed diagnosis such as early mild symptoms and atypical symptoms. This scale and distribution characteristics just fit the small sample generalization advantage of the mixed architecture of Vit and VGG16- relying on the inductive bias of VGG16 to reduce the dependence on massive data, without relying on millions of pre training samples like pure ViT, the stable learning and generalization of the model in the task of stroke diagnosis can be realized.

4.2. Multi pathology standardization

ACDC (MICCAI 2017) is an important benchmark data set for cardiac cine MRI image analysis. It contains data of 150 multicenter patients, covering five major cardiac pathological types (normal,

anterior myocardial infarction, dilated cardiomyopathy, hypertrophic cardiomyopathy, and right ventricular abnormalities). The data set provides a complete cardiac cycle sequence, and the end diastolic and end systolic frames are accurately labeled by professional physicians, with rich clinical metadata. All images are preprocessed by spatial registration, uniform resolution, gray normalization and other standardization, which significantly reduces the data heterogeneity. This design makes ACDC particularly suitable for the study of dictionary encoded texture changes - the characteristic texture differences presented by different cardiomyopathy can be accurately captured by the dictionary. Combined with CNN's ability to understand the global structure, it can effectively adapt to multiple pathological modes and significantly improve the generalization performance of the model in cross center and cross device scenarios.

4.3. Multimodal data fusion

Brain tumor segmentation (BraTS) challenge dataset is the gold standard dataset to promote the development and evaluation of advanced models including BoTNet in the field of brain tumor segmentation. The data set provides brain MRI images of four core modes: T1 weighted, T1 enhanced, T2 weighted and FLAIR. Each mode corresponds to different pathological characteristics of the tumor - for example, T1 enhanced mode clearly shows the enhancement boundary of the tumor, FLAIR mode highlights the edema area, and T2 weighted mode presents the whole structure of the tumor. Multimodal data can provide comprehensive information such as tumor location, boundary, infiltration range and surrounding tissue impact. BotNet does not need to rely on massive data of single mode to make up for the lack of features. It can not only reduce the burden of storage and training caused by excessive single mode data, but also provide more accurate feature input for the model and improve the accuracy of tumor segmentation.

5. Challenges and Solutions

5.1. Disjointed workflow and insufficient stability

When the intelligent imaging diagnosis technology is launched, it often faces the problems of disconnection from the clinical workflow and poor performance stability. Through the development of embedded AI interface, it can realize the seamless connection with the hospital PACS system and electronic medical record system, so that the AI analysis results can be automatically pushed to the doctor's film reading interface, without the need for manual secondary import of data, which greatly improves the efficiency of the diagnosis process [11]; At the same time, we carried out multi center large sample clinical verification, covering different levels of medical institutions such as tertiary hospitals and county-level hospitals, and included the data generated by different brands and models of imaging equipment such as GE, Siemens and Lianying, to comprehensively verify the performance stability of the model in a variety of clinical scenarios, avoid fluctuations in the accuracy of the model due to equipment differences and different diagnosis and treatment environments, and lay a reliable foundation for the clinical application of technology.

5.2. Data privacy disclosure

If the training data is unbalanced in population distribution and disease coverage, it is easy to cause bias in the algorithm, such as the low diagnostic accuracy of elderly patients and rare diseases; And when the AI diagnosis is wrong, the responsibilities among developers, hospitals and physicians have not been clearly defined. To solve these problems, technical research team can use differential privacy, homomorphic encryption and other technologies to protect data privacy, build a balanced and diverse training data set, and supplement the samples of vulnerable groups and rare diseases to reduce algorithm bias. At the same time, researchers can work with industry associations to formulate the responsibility division norms and ethical guidelines for AI diagnosis, and delimit the security boundary for technology development.

6. Conclusion

As the core path to break through the bottleneck of traditional manual film reading and promote precision medicine, the research of medical imaging intelligent diagnosis technology has made significant progress in technical architecture innovation, data set support and clinical adaptability exploration, and also defined the future breakthrough direction, providing a clear context for technology iteration and clinical transformation.

From the perspective of technological innovation, the current medical AI imaging diagnosis technology realizes "developing strengths and circumventing weaknesses" through architecture fusion - the hybrid architecture of Vit and VGG16 uses the inductive bias of VGG16 to solve the massive data dependence of pure ViT with local feature pre extraction and lightweight Global Association modeling, and the accuracy of 1000 samples can reach the accuracy of 100 million samples of pure ViT, and reduces the computing power to adapt to the basic hardware; Hybrid CNN and dictionary representation to build a global local feature system, U-Net to locate lesions, and K-SVD to mine micro textures to make up for the lack of fine lesions captured by a single model; BoTNet is based on RESNET transformation, reusing pre training resources to shorten the research and development cycle, and the deep self-attention module strengthens the multi-scale modeling of complex lesions to balance compatibility and efficiency. These innovations break the traditional CNN global correlation short board and pure transformer sample dependence dilemma.

Remarkable achievements have been made in data set construction: Brain Stroke CT Image Dataset adapts to small sample learning with clinical fitting sample distribution; ACDC data set covers 5 types of heart diseases, and data heterogeneity is resolved after preprocessing; The BraTS dataset is filled with multimodal MRI to fill the single modal gap. The three not only solve the pain points of data scarcity and low quality, but also build a unified evaluation standard to promote the comparison and optimization of results.

The technology still has limitations in clinical application: the diagnostic accuracy of rare diseases fluctuates, and multimodal fusion does not reach pathological depth correlation; The decision-making trust of the model and the medical system needs to be more interpretable; Cross center data collaboration needs to balance privacy and security.

In the future, it can be deepened from three aspects. First, it is to integrate federal learning and small sample learning, build a data alliance and mine the value of unmarked data through supervision; The second is to strengthen the interpretability of the model, integrate medical prior knowledge, and enhance trust through visualization and natural language; The third is to promote the standardization of human-computer cooperation of "AI preliminary screening physician final review", and establish a dynamic monitoring mechanism to optimize the stability of the model.

To sum up, medical imaging intelligent diagnosis technology has moved from technical verification to clinical adaptation. Its core value is not only to improve the accuracy and efficiency, but also to narrow the medical gap through technology. With the breakthrough of bottleneck and the deepening of clinical docking, this technology will become an important support for precision medicine, and provide guarantee for early intervention, efficacy evaluation and prognosis improvement.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] Havaei M, Davy A, Warde-Farley D, et al. Brain tumor segmentation with deep neural networks. Medical image analysis, 2017, 35: 18 - 31.

[2] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv: 2010.11929, 2020.

[3] Chen W, Ayoub M, Liao M, et al. A fusion of VGG-16 and ViT models for improving bone tumor classification in computed tomography. Journal of Bone Oncology, 2023, 43: 100508.

[4] Ranneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation//International Conference on Medical image computing and computer-assisted intervention. Cham: Springer international publishing, 2015: 234 - 241.

[5] Qiao G X, Song J H. Cardiac image segmentation based on improved u-net//2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML). IEEE, 2022: 133 - 137.

[6] Aharon M, Elad M, Bruckstein A. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE Transactions on signal processing, 2006, 54 (11): 4311 - 4322.

[7] Srinivas A, Lin T Y, Parmar N, et al. Bottleneck transformers for visual recognition//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 16519 - 16529.

[8] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117 - 2125.

[9] Hatamizadeh A, Heinrich G, Yin H, et al. Fastervit: Fast vision transformers with hierarchical attention. arXiv preprint arXiv: 2306.06189, 2023.

[10] Li H, Su D, Cai Q, et al. Bsafusion: A bidirectional stepwise feature alignment network for unaligned medical image fusion//Proceedings of the AAAI Conference on Artificial Intelligence. 2025, 39 (5): 4725 - 4733.

[11] Gafar Y H E D, Bekhet S E, Hanafy H M, et al. Cloud-Enabled Diagnostic System: Leveraging PACS and AI for Enhanced Remote Carotid Atherosclerosis Managemen//2024 International Conference on Machine Intelligence and Smart Innovation (ICMISI). IEEE, 2024: 70 - 73.