ISSN: 2688-8645

Deep Learning in Medical Imaging: Application Progress and Future Challenges

Jingwen Wang *, Pingping Yang, Ziyi An, Run Luo, Jiayi Feng, and Linghan Zhang

School of Health and Medical Technology, Chengdu Neusoft University, Chengdu, Sichuan, China

* Corresponding author: Jingwen Wang (Email: WangJingwen@nsu.edu.cn)

Abstract

Medical Imaging occupies a central position in modern clinical diagnosis and treatment. However, faced with vast and continuously growing volumes of complex imaging data, traditional image processing algorithms reliant on manual features struggle to meet the demands for efficient and precise clinical applications. Against this backdrop, Deep Learning technology, leveraging its unique advantages, has demonstrated immense application potential in Medical Imaging, significantly enhancing the accuracy and efficiency of medical diagnosis. This paper systematically reviews the primary technological developments and practical application scenarios of Deep Learning within the Medical Imaging domain. The focus will be on key application areas such as image reconstruction, segmentation, and registration, alongside multimodal synthesis, intelligent lesion detection, and computer-aided diagnosis. Finally, the paper will delve into the current limitations of these technologies and explore future development directions, aiming to provide valuable reference for research and application in this field.

Keywords

Medical Imaging; Deep Learning; Image Reconstruction; Computer-aided diagnosis; Image Segmentation; Federated Learning; Multimodal Large Models.

1. Introduction

Medical Imaging constitutes an indispensable foundational element within clinical medicine, healthcare provision, and contemporary medical practice. It encompasses virtually every facet of disease diagnosis and treatment – from early disease screening, to furnishing critical information for accurate patient diagnosis, to formulating highly personalised therapeutic regimens. Furthermore, utilising various imaging techniques such as X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), Ultrasound (US), and Positron Emission Computed Tomography (PET) , Medical Imaging also serves to objectively evaluate and monitor treatment efficacy during the prognostic phase of therapeutic regimens.

Early Medical Imaging relied on traditional image processing techniques such as threshold segmentation, region growing, and edge detection. A typical approach involved designing manual features tailored to specific problems, which were highly dependent on the physician's accumulated expertise. While effective for simple scenarios, these methods often lacked robustness for complex structures and noisy images, and exhibited limited flexibility when processing images from different modalities. Over 90% of the medical data generated globally each day originates from Medical Imaging. Medical Imaging data continues to grow at an annual rate of approximately 30%, while the number of radiologists increases by only about 4% annually. This represents a significant disparity. The imbalance between supply and demand for Artificial Intelligence (AI)-assisted diagnostic tools not only exacerbates the burden on

ISSN: 2688-8645

clinicians but also heightens the risk of missed or misdiagnoses. Consequently, traditional algorithms fail to meet the demands of modern clinical medicine.

In recent years, rapid advances in computational power coupled with the accumulation of vast medical datasets have jointly propelled significant progress in Deep Learning applications within computer vision. Notable models include Convolutional Neural Network (CNN)[1][2][3], Transformer model[4], and Generative Adversarial Network (GAN)[5]. These models have been extended to Medical Imaging, achieving widespread adoption and seamless integration with Medical Imaging systems. Deep Learning technology possesses exceptional automated feature extraction capabilities. It can learn deep and complex abstract features within images through a data-driven approach. Furthermore, Deep Learning models can process multimodal images, integrating different types of information for comprehensive analysis, thereby providing clinicians with holistic diagnostic evidence and personalised treatment plans. Consequently, the significance of Deep Learning in Medical Imaging technology lies in its ability to effectively process complex medical data, enhancing diagnostic accuracy and speed.

This paper summarises the core algorithmic architecture of Deep Learning and introduces six application scenarios: medical image reconstruction, multimodal Medical Imaging, intelligent lesion detection, computer-aided diagnosis, medical image segmentation, and medical image registration. Additionally, it explores recent advancements in multimodal fusion and foundational large models. Finally, it conducts an in-depth analysis of systemic challenges encountered during clinical-scale implementation, including small-sample annotation, medical data silos, and limitations of unimodal models. Addressing these issues, the paper identifies semi-supervised learning, Federated Learning, and foundational multimodal medical large models as core future development directions.

2. Deep Learning Technologies

The remarkable achievements of Deep Learning in Medical Imaging owe much to the continuous innovation and evolution of underlying network architectures. To address challenges inherent to Medical Imaging—such as its distinctive 2D/3D spatial attributes, the diversity of multimodal data, and the relative scarcity of training samples—the academic community has developed multiple core architectures specifically optimised for medical tasks. Current mainstream research paradigms primarily encompass Convolutional Neural Network (CNN), Transformer models centred on Attention Mechanism, and generative model (GAN).

2.1. Convolutional Neural Network (CNN)

As the most representative and widely applied feedforward neural network architecture in Medical Imaging, CNN plays an irreplaceable role in this field due to their unique core mechanisms. It primary advantage lies in key mechanisms such as weight sharing and spatial downsampling. By focusing on local information for feature extraction, they can accurately capture fundamental local anatomical features such as edges and textures within medical images. Moreover, as the depth of the CNN network progressively increases, these extracted local basic features undergo continuous integration and abstraction, ultimately transforming into more valuable high-level semantic features. This provides robust support for the precise analysis of medical images.

The core workflow of CNN follows a progressive logic of "input – feature extraction – classification mapping". Each stage is tightly interconnected and builds upon the preceding one. The overall architecture is presented in Figure 1.

The CNN commences with the input layer, which receives the raw image as the initial data source for the entire network. The input image first undergoes processing in the first layer, where convolution operations perform preliminary feature extraction, ultimately outputting a

ISSN: 2688-8645

feature map with a depth of three to the second layer. To further compress the feature dimensions while preserving key information, pooling operations are applied to the three feature maps from the second layer, yielding a third-layer feature map that retains a depth of three. This "convolution-pooling" feature extraction process is repeated through multiple iterations, ultimately yielding five feature maps at layer five. These five feature maps (essentially five matrices) are then unrolled row-wise and concatenated into a continuous vector. This vector is directly fed into the network's fully connected layer.The fully connected layer essentially functions as a Back Propagation Neural Network (BP). Each feature map in the diagram may be conceptualised as a neuron arranged in matrix form, whose core functionality is fundamentally identical to that of a neuron within a BP neural network. Both perform feature transmission and signal processing.

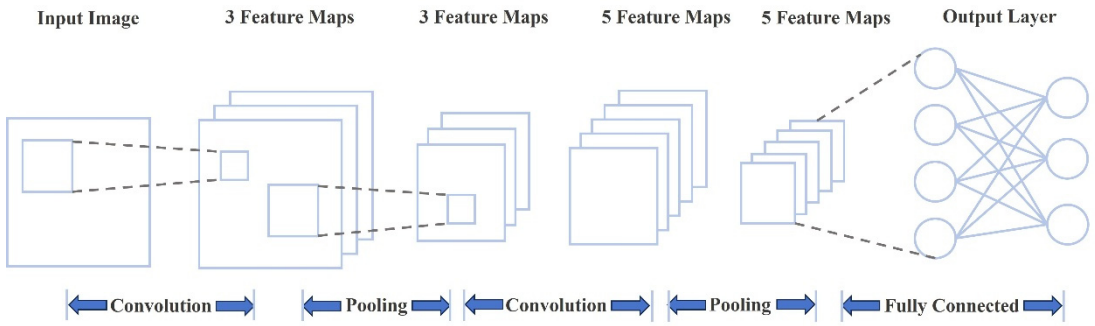

Figure 1. CNN Structural Diagram

In a standard convolutional layer, the spatial dimension computation of feature maps adheres to a rigorous mathematical paradigm, as expressed in Equations (1) and (2):

$$W_{output} = \frac{W_{input} - W_{filter}}{S} + 1 \tag{1}$$

$$H_{output} = \frac{H_{input} - H_{filter} + 2P}{S} + 1 \tag{2}$$

here,W and H denote the width and height of the feature map respectively,S represents the stride of the convolutional kernel, and P denotes the number of edge padding pixels. By integrating with pooling layers, CNN can reduce feature dimensions while endowing the model with a degree of translation invariance. Within pooling layers, the computational paradigm follows Equations (3) and (4):

$$W_{output} = \frac{W_{input} - W_{filter}}{S} + 1 \tag{3}$$

$$H_{output} = \frac{H_{input} - H_{filter}}{S} + 1 \tag{4}$$

In Medical Imaging, where CT, MRI, and similar data are inherently three-dimensional volumetric information, traditional two-dimensional convolutions struggle to meet practical demands. Consequently, academia has progressively extended these techniques into three-dimensional space. The V-Net model[6] stands as a particularly representative example. This model innovatively employs three-dimensional convolutional kernel designs, successfully achieving end-to-end feature extraction from three-dimensional voxel data and effectively adapting to the volumetric characteristics of medical images. Concurrently, the widespread issue of gradient vanishing during deep network training constrains performance improvements in Medical Imaging networks. To address this, residual connections[1] have been extensively applied within such networks. By effectively mitigating gradient vanishing,

ISSN: 2688-8645

they significantly enhance the model's capacity to learn and recognise complex anatomical structures within medical images.

2.2. Attention Mechanism and Transformers

Despite substantial progress, CNN still face numerous challenges requiring breakthroughs in the Medical Imaging domain. Issues such as insufficient data samples, susceptibility to overfitting, and persistently high computational costs have, to some extent, limited the full realisation of their application potential. To overcome these constraints, researchers are continuously deepening their exploration, developing novel model architectures, optimising algorithmic workflows, and refining data processing solutions, all aimed at further enhancing the accuracy and efficiency of CNN in medical image analysis.

Attention Mechanism[7] were introduced to enable feature re-weighting, overcoming the pronounced limitations of traditional methods in capturing spatial dependencies between distant organs within medical images. As a key implementation of Attention Mechanism, Squeeze-and-Excitation Network (SE Net)[8] employs global average pooling and fully connected layers to compute channel attention weights, thereby achieving the core objectives of enhancing effective feature representation and suppressing ineffective feature interference.

In recent years, the Transformer architecture[4], originating from natural language processing, has fundamentally disrupted conventional convolutional paradigms. At its core, the Transformer relies on a multi-head self-attention mechanism, enabling global interactions between image pixels or feature blocks. Its self-attention calculation formula is expressed as Equation (5):

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{5}$$

where the query vector Q , key vector K , and numerical vector V are given, and d_k denotes the vector dimension. The self-attention mechanism computes a weighted sum, with weights determined by the similarity between the query and key vectors.

As medical images lack absolute sequence information, Transformer typically require Positional Encoding (PE) to preserve the relative positions of anatomical structures, as shown in Equations (6) and (7):

$$PE(pos, 2i) = \sin(\frac{pos}{10000^{2i/d_{model}}}) \tag{6}$$

$$PE(pos, 2i+1) = \cos(\frac{pos}{10000^{2i/d_{model}}}) \tag{7}$$

where pos denotes the current position, i represents the feature vector dimension, and d_{model} indicates the model dimension.

Within medical vision research, TransUNet[9] represents a landmark breakthrough. As one of the first models to successfully integrate Transformers with CNN at depth, it laid crucial foundations for subsequent related studies. It employs CNN as the core feature extraction mechanism, capturing high-resolution low-level local details. Subsequently, feature maps undergo serialisation before input into the Transformer, enabling efficient extraction of global contextual information. Finally, precise upsampling is achieved through a cascaded decoder, enabling feature restoration and localisation. Building upon TransUNet[9], research has iterated further: Swin-Unet[10], based on a sliding window mechanism, and UNETR[11], specifically tailored for three-dimensional Medical Imaging. Both models optimise computational complexity, successfully overcoming technical bottlenecks to realise global feature modelling for large-scale three-dimensional medical images.

Through continuous technological iteration and deepening research, the application value of Transformer in medical image analysis will be further unlocked, with increasingly broad

ISSN: 2688-8645

prospects for development.Looking ahead, various Transformer-based models enhanced through structural and algorithmic optimisation are poised to play a pivotal role in medical image analysis tasks, providing more precise, efficient, and reliable technical support for clinical diagnosis and treatment planning. Concurrently, ongoing advancements in cutting-edge technologies such as multimodal fusion and weakly supervised learning will continue to empower Transformer models, driving sustained performance improvements and expanding their application boundaries within medical image analysis.

2.3. Generative Adversarial Network (GAN) and Diffusion Models

Building upon the Transformer's ability to accurately model global dependencies in medical images through Attention Mechanism, Generative Models further expand the application boundaries of medical image processing through their powerful feature generation and transformation capabilities. Among these, Generative Adversarial Network (GAN) and Diffusion Models represent two particularly representative approaches.

GAN[5] occupy an irreplaceable core position in key tasks such as medical image synthesis, denoising, and data augmentation. Figure 2 illustrates the GAN's architecture. This model comprises a Generator and a Discriminator, which co-evolve through a zero-sum game. Their core objective is to minimise the distributional divergence between generated data and authentic medical image data, thereby producing highly realistic image samples.

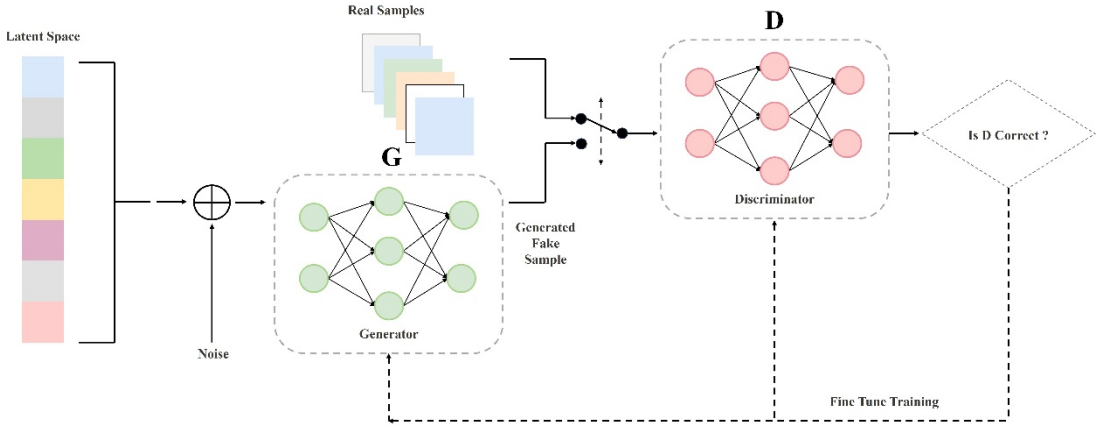

Figure 2. GAN Architecture Diagram

Faced with the technical challenge of multimodal medical image transformation lacking paired data, CycleGAN[12] innovatively introduced a cyclic consistency loss. This successfully achieved cross-modal mapping under unsupervised conditions, effectively resolving a core issue in this field.

In recent years, Diffusion Models[13] have gained prominence. By introducing Gaussian noise to original medical images and continuously learning denoising patterns during the reversal process, they demonstrate more stable training outcomes and higher generative fidelity compared to GAN. These models have now quietly become the leading research paradigm in medical image synthesis and anomaly detection.

3. Applications of Deep Learning Techniques in Medical Imaging

3.1. Image Reconstruction and Multimodal Synthesis

With the rapid iteration of Deep Learning techniques and the continuous advancement of research in Medical Imaging, it has become the core technological foundation for medical image reconstruction and multimodal synthesis tasks. Leveraging its robust capabilities in feature learning, non-linear fitting, and global modelling, it has effectively overcome the limitations of

ISSN: 2688-8645

traditional medical image processing methods in terms of accuracy, efficiency, and generalisability, playing an increasingly vital role in clinical diagnosis and treatment.

Medical image reconstruction, representing the core intersection of radiological physics and computational mathematics, fundamentally involves transforming raw digital information acquired during human imaging. This process employs specific mathematical algorithms to generate images that clearly display anatomical details and pathological features, facilitating interpretation. This step is crucial for visual output generation in imaging modalities such as CT, MRI, and PET, and fundamentally determines image quality and radiation dose. Deep Learning techniques have found extensive application within medical image reconstruction, spanning multiple clinically relevant scenarios including low-dose CT/MRI reconstruction, tomographic image restoration, image denoising, and artefact removal. These technologies effectively enhance image resolution and clarity while preserving critical anatomical structures and pathological features. Concurrently, they enable reduced radiation exposure or shorter scan durations, thereby providing high-quality imaging support for precise clinical diagnosis.

Complementing medical image reconstruction is multimodal Medical Imaging technology. This involves integrating image data from two or more imaging modalities with differing principles to provide comprehensive diagnostic information unattainable through a single modality. Crucially, multimodality is not merely the superimposition of multiple images; its core lies in employing image fusion techniques to achieve deep integration of information from different sources at the pixel, feature, or decision-making level.Within the field of multimodal medical image synthesis, Deep Learning has successfully achieved cross-modal conversion between different imaging modalities (such as CT and MRI, PET and CT), missing modality completion, and the generation of realistic image samples through various generative models. This application not only effectively addresses the clinical challenges of obtaining multimodal data and the scarcity of paired data. It also provides efficient solutions for multimodal fusion diagnosis, medical education and training, and data augmentation, further propelling Medical Imaging technology towards intelligent and precise development.

3.1.1. Intelligent Acceleration and De-interleaving in Magnetic Resonance Imaging

Traditional MRI reconstruction techniques have long relied on complex iterative optimisation algorithms. Representative examples include Sequential Enhancement of Nuclear Spin Echo (SENSE), Generalised Reciprocal Partial Parallel Acquisition with Automatic Partial Reconstruction (GRAPPA), and Compressed Sensing (CS). These conventional methods exhibit significant technical limitations. Not only do they entail high computational complexity and lengthy reconstruction times, but the rational selection of parameters also poses considerable challenges, failing to meet clinical demands for efficient and convenient imaging. In contrast, Deep Learning techniques effectively overcome these bottlenecks by directly learning end-to-end mappings between undersampled data and full-sampled images. This significantly reduces the overall time required for MRI scanning and image reconstruction, offering a superior solution for efficient clinical imaging. Currently, Deep Learning-based MRI reconstruction methods can be broadly categorised into two technical paradigms: data-driven and model-driven approaches.

Within the data-driven end-to-end reconstruction framework, the Deep Cascade Convolutional Neural Network[14] alternates CNN modules with data consistency layers to perform spatial de-aliasing and frequency-domain fidelity constraints respectively. This efficiently extracts spatial correlation information from dynamic MRI sequences. Addressing the inherent complex nature of raw MRI k-space data, the DeepcomplexMRI algorithm[15] innovatively constructs Convolutional Neural Networks within the complex domain. It autonomously learns cross-channel image correlations, achieving high-precision, high-fidelity image reconstruction without requiring additional coil sensitivity calculations.

ISSN: 2688-8645

In Deep Learning reconstruction based on prior models, researchers have integrated traditional optimisation algorithms such as the Alternating Direction Method of Multiples (ADMM) and Split Bergman Iteration with deep networks, achieving algorithmic networking. Taking ADMM-Net[16] as an example: this model decomposes the iterative process of imaging optimisation into a network architecture with clear physical meaning. While preserving the core theory of compressed sensing, it adapts model parameters through a data-driven mode, significantly enhancing reconstruction acceleration factors and cross-device generalisation capabilities.

3.1.2. Noise Suppression and Artifact Elimination in Low-Dose CT

Driven by the clinical imperative to reduce the potential carcinogenic risks to patients from ionizing X-ray radiation, Low-Dose CT (LDCT) has been widely adopted in clinical screening for diseases such as pulmonary nodules. However, the low-dose scanning modality directly induces severe quantum noise and streak artifacts, leading to a significant degradation in imaging quality. Under these conditions, the traditional Filtered Back Projection (FBP) algorithm[17] generally fails to reconstruct CT images that meet clinical diagnostic standards.

Deep Learning-based LDCT denoising primarily operates in the image domain or sine map domain. For instance, RED-CNN[18] employs a symmetric residual encoding-decoding structure to restore image size and detail while preserving information.The further developed WavResNet algorithm[19] integrates wavelet transform with deep CNNs. It first decomposes LDCT images into subbands of different frequencies and orientations via contour wavelet transform, then processes high- and low-frequency components separately through the network for denoising. This strategy smooths background noise while significantly preserving minute high-frequency edges of lesions. Moreover, Wasserstein GAN (WGAN)[20] incorporating perceptual loss functions has seen widespread application in LDCT reconstruction, yielding normal-dose CT images that exhibit enhanced visual realism and anatomical fidelity.

3.1.3. Cross-Modal Synthesis and Super-Resolution Reconstruction in Medical Imaging

The core advantage of multimodal Medical Imaging lies in its capacity to provide complementary physiological and pathological information, offering more comprehensive support for clinical diagnosis and treatment planning. In clinical practice, scenarios such as positron emission tomography (PET) or radiotherapy planning often rely on the electron density information provided by CT images. However, to avoid additional radiation exposure, synthesising CT directly from MRI has become an urgent clinical need.

The CycleGAN-based model[21] effectively addresses such challenges. Without requiring strictly paired data, it leverages the mutual constraint mechanism of dual generators and dual discriminators, combined with gradient consistency loss, to successfully construct a high-precision nonlinear mapping from MRI to CT. This achieves information conversion without additional radiation exposure.

Beyond intermodal conversion, multimodal Medical Imaging techniques demonstrate significant value in image super-resolution enhancement. Arterial spin labelling (ASL) MRI, constrained by physical limitations, often suffers from substantial noise during acquisition, severely compromising image quality.To address this limitation, researchers proposed a two-stage CNN architecture employing multiple loss functions[22]. This achieves quality enhancement through phased collaborative optimisation: the first stage prioritises removing large-scale noise from the image, while the second stage focuses on refining high-frequency details to further enhance image clarity. Ultimately, weighted fusion techniques generate high-resolution cerebral blood flow images. The application of such techniques effectively elevates imaging capabilities in primary healthcare institutions without requiring costly hardware upgrades, demonstrating exceptional clinical utility.

ISSN: 2688-8645

3.2. Intelligent Lesion Detection and Computer-Assisted Diagnosis with

Intelligent lesion detection, empowered by computer vision and Deep Learning technologies, automatically identifies, localizes, and quantitatively analyzes pathological abnormalities within complex medical images. It represents the most fundamental and actively researched direction of AI in radiology. Meanwhile, computer-aided diagnosis (CAD) serves as an auxiliary technology designed to provide radiologists with a "second opinion." It strictly adheres to the core principle of "assistance rather than replacement," ultimately ensuring that medical decision-making authority always remains firmly in the hands of physicians.

Within Deep Learning applications for Medical Imaging, these two domains represent the most clinically actionable and commercially transformative core areas. Unlike conventional object detection tasks on natural images, medical lesions possess distinct complex characteristics: they are extremely small in size, exhibit highly heterogeneous morphological presentations, and often have indistinct boundaries with surrounding healthy tissue. It is precisely these inherent features that impose far stricter demands on the sensitivity and specificity metrics of Deep Learning models than conventional tasks.

3.2.1. Superhuman Expert-Level Detection and Diagnosis for Specific Diseases

Ophthalmic disease auxiliary diagnosis represents a particularly representative application. In 2016, the medical flagship journal "JAMA" published a study by Gulshan's team[23]. Employing the Inception-v3 architecture, the team utilised over 128,000 retinal fundus images as training data, with images undergoing multiple cross-annotations by 54 American ophthalmologists.On an independent test set, this intelligent diagnostic system achieved an area under the receiver operating characteristic curve (AUC) of 0.99 for detecting referral-level diabetic retinopathy (DR). Subsequent clinical validation further demonstrated a sensitivity of 92.7% and specificity of 95.5%. This fully illustrates the immense public health value of this technology in the large-scale prevention and control of blinding diseases.

In skin cancer screening, Esteva et al.'s study[24] published in "Nature" similarly caused a sensation. The research team employed transfer learning techniques, fine-tuning a deep neural network using 129,000 dermatological lesion images encompassing 2,032 conditions. When distinguishing the most lethal melanoma from benign moles, the AI system achieved a diagnostic accuracy of 69.4%. A panel of 21 specialist dermatologists achieved a diagnostic accuracy of 66.0%. The performance of both was comparable. This demonstrates the feasibility of low-cost early skin cancer screening via smartphone terminals.

3.2.2. Three-dimensional lesion detection and general target analysis

Devices such as CT and MRI generate volumetric data. To detect lesions, the relevant detection networks must evolve towards three-dimensional capabilities.In lung nodule screening, researchers including Wentao Zhu conducted relevant studies. They extended the Faster R-CNN network into the three-dimensional 3D R-CNN[25]. This network incorporates a 3D Feature Pyramid Network (FPN) and a dual-encoder-decoder architecture. It effectively captures the spatial characteristics of lung nodules while significantly reducing false positives, which are predominantly caused by vascular cross-sections or pleural margins.

Presently, numerous large-scale, multi-organ whole-body image datasets have been made open-source, such as the DeepLesion dataset released by the National Institutes of Health Clinical Center[37]. With the aid of these datasets, lesion detection models have undergone new developments. Previous models could only detect lesions in a single organ or for a single disease. Current models, such as MVP-Net[26], can now perform generalised lesion detection across multiple organs throughout the body. This model incorporates cross-slice spatial attention and multi-view fusion techniques. Using a single neural network, it achieves high-precision, universal lesion localisation across multiple regions including bones, lungs, liver, kidneys, and

ISSN: 2688-8645

abdominal soft tissues. This significantly enhances radiologists' efficiency in reviewing entire images.

3.2.3. Feature-Level Fusion Diagnosis Using Multimodal Data

In clinical practice, achieving precise diagnostic outcomes typically requires the complementary and mutually corroborative use of multi-dimensional information. The early diagnosis of Alzheimer's disease exemplifies this scenario, where single-modality examination data alone struggles to support reliable conclusions.

Addressing this challenge, Shen Dinggang et al.[27] adopted a multimodal data fusion approach. They utilised both morphological features of brain atrophy reflected in structural MRI and metabolic functional characteristics of brain regions revealed by PET scans. Building upon this foundation, the team employed models such as deep typical correlation analysis networks and multimodal Boltzmann machines to achieve joint representation of different types of heterogeneous data at the level of high-level semantic features. Ultimately, in the prediction task of transition from mild cognitive impairment to Alzheimer's disease, this approach achieved superior classification accuracy compared to single-modality data.

3.3. Precise Image Segmentation and Deformation Registration

Medical image segmentation is a core technology for precisely delineating the boundaries of lesions, organs, and normal tissues within complex medical images, as shown in Figure 3; whereas deformable image registration is a critical step for the spatial alignment and morphological correction of multi-modal and multi-temporal images, as illustrated in Figure 4. Together, these two techniques constitute the fundamental prerequisites for three-dimensional (3D) reconstruction, quantitative volumetric measurement, and image-guided targeted surgery.

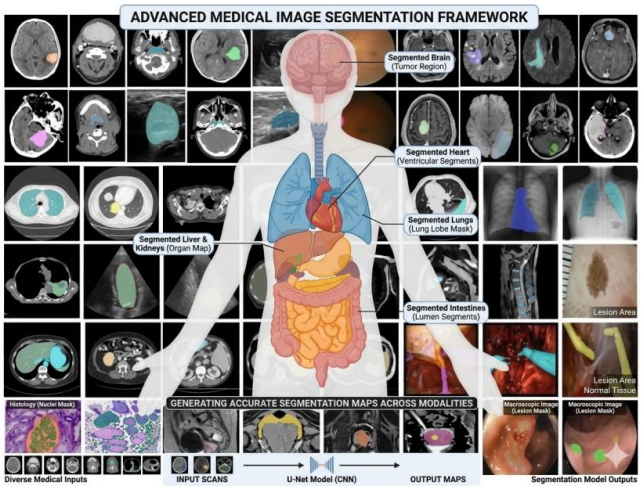

Figure 3. Primary Targets in Medical Image Segmentation

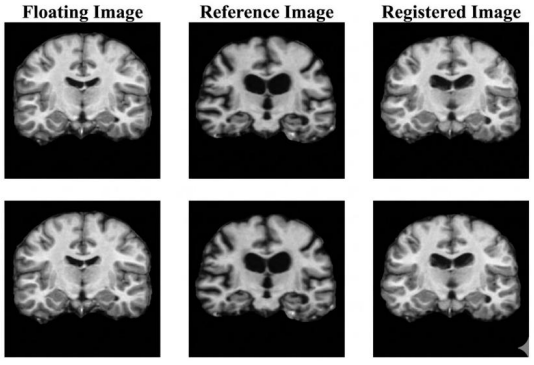

Figure 4. Brain Medical Image Registration

ISSN: 2688-8645

Deep Learning overcomes the limitations of traditional methods reliant on complex mathematical optimisation, transforming cumbersome optimisation problems into end-to-end non-linear regression tasks. This approach not only enhances the precision of tissue boundary extraction in image segmentation but also enables more efficient and robust spatial alignment in image registration. By significantly boosting computational speed while maintaining processing accuracy, it provides core technological support for precision medicine in post-processing of medical images.

3.3.1. Innovations in Symmetric Segmentation Networks and Adaptive Frameworks

In 2015, Ronneberger et al. introduced U-Net[28] . Since then, the symmetric encoder-decoder architecture with skip connections has become the mainstream approach in medical segmentation. U-Net demonstrated exceptional performance. Its downsampling path extracts high-level semantic and contextual information, while its upsampling path integrates shallow feature maps to fully reconstruct lesion locations and edge details. To address severe foreground-background class imbalance in medical images (e.g., lesions occupying merely 1% of total volume), cross-entropy loss is frequently replaced by Dice Loss or Focal Loss. Dice Loss directly optimises the overlap between predicted regions and ground truth labels. Focal Loss incorporates an exponential penalty term into the cross-entropy loss. This automatically reduces the weight assigned to easily distinguishable background samples, compelling the model to prioritise samples with blurred boundaries and challenging distinctions. Details are illustrated in Figure 5.

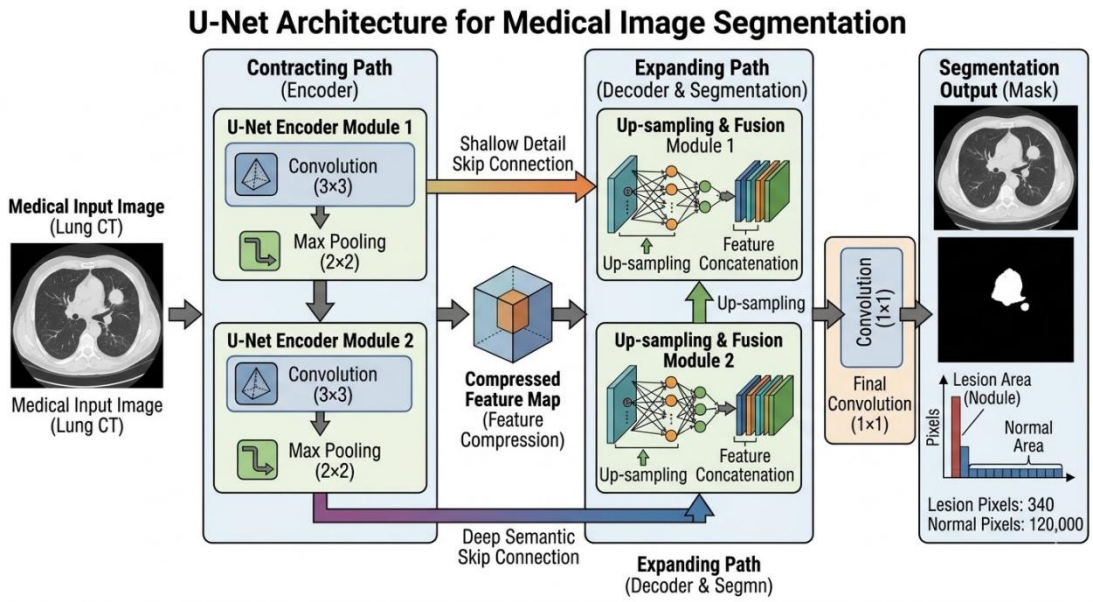

Figure 5. U-Net Architecture Diagram

In recent years, medical image segmentation technology has advanced towards greater automation and versatility. As an adaptive configuration framework, nnU-Net [29] prioritises analysing dataset attributes—such as image type, clarity, and aspect ratio—over pursuing complex, novel network architectures. It then performs data attribute extraction, automatic preprocessing configuration, network hyperparameter tuning, and subsequent optimisation. This framework requires no manual intervention.Across dozens of international medical segmentation competitions, it has outperformed highly customised complex models. It has now become the standard reference model in the field of medical image segmentation.

TotalSegmentator[30] demonstrates immense potential for multi-organ, large-scale general segmentation. In routine CT scans, it reliably segments up to 104 anatomical structures, encompassing various visceral organs, blood vessels, muscles, and bones. Experiments confirm

ISSN: 2688-8645

an average Dice score of 0.943, achieved with remarkably short inference times. This significantly advances automation in whole-body organ volumetry and radiotherapy planning.

Furthermore, with the emergence of the image segmentation AI model paradigm- Segment Anything, the medical-grade visual large model MedSAM[31] utilises over 1.5 million pairs of medical images and masks to form an extensive multimodal dataset. Researchers fine-tuned this model to enable accurate segmentation of unseen lesions through zero-shot or bounding box prompts, demonstrating MedSAM's robust generalisation capabilities.

3.3.2. Deep Deformable Registration of Medical Images

Medical image registration involves computing spatial transformation matrices or dense deformation fields. It requires precisely aligning images taken at different times (e.g., during disease progression follow-ups) or from different modalities (e.g., CT and MRI) within a common coordinate system. Traditional registration methods employ similarity measures such as mutual information and normalised cross-correlation. These require iterative numerical optimisation and are computationally intensive. Deep Learning techniques transform image registration into an end-to-end mapping process.

Table 1. Deep Learning segmentation and registration models

Comparative Analysis of Representative Deep Learning Segmentation and Registration Models	Core Architecture/ Mechanism	Supervision Approach	Core Advantages and Clinical Applications
U-Net / V-Net[28]	Encoder-Decoder Architecture, SkipConnections,3D Convolution	Fully supervised	High-precision extraction of lesions, vessels, and organ contours for small-sample datasets.
TransUNet[9]	CNN-Transformer hybrid architecture	Fully supervised	Combines local edge details with global spatial dependencies, suitable for complex tumour segmentation with blurred boundaries.
nnU-Net[29]	Adaptive heuristic rule-driven variant of U-Net	Fully supervised	Automatically adapts to medical datasets of varying modalities and resolutions without manual parameter tuning, demonstrating exceptional generalisation capability.
MedSAM[31]	Prompt-driven Visual Large Model	Large-scale pre-training	Possesses exceptional zero-shot transfer learning capabilities, enabling interactive segmentation of rare lesions via click or box selection.
VoxelMorph[32]	CNN integrated with Spatial Transformation Networks (STN)	Unsupervised / Weakly supervised	Generates differential isomorphous deformation fields in sub-second timeframes, suitable for intraoperative multimodal alignment and organ deformation tracking.

VoxelMorph[32] represents a pioneering unsupervised registration framework. It employs a spatial transformation network to concatenate the floating and fixed images. The data is then fed into a U-Net-like architecture, enabling the network to directly predict smooth displacement vector fields. The loss function comprises two components: one derived from image similarity calculations, and another involving a smoothing regularisation term for the deformation field, such as the L2 norm of gradients. This ensures image deformations adhere to physical principles, preventing spatial crossovers, folding, or tearing. Conclusions

ISSN: 2688-8645

demonstrate VoxelMorph achieves registration accuracy comparable to traditional state-of-the-art algorithms while reducing computation time from tens of minutes to sub-second levels. This represents a revolutionary advancement for intraoperative real-time image navigation and dynamic tumour target tracking. Table 1 presents representative Deep Learning segmentation and registration models within Medical Imaging.

4. Systemic Challenges and Future Development

Whilst Deep Learning consistently demonstrates strong performance on academic open datasets, its large-scale implementation in real clinical practice faces numerous intractable challenges. Resolving these issues constitutes a core task for future medical AI research.

4.1. The Small-Sample Dilemma and the Rise of Semi-Supervised Learning

Medical Imaging exhibits extreme specialisation in anatomy and pathology. Acquiring high-quality pixel-level annotations from senior specialists often demands considerable time investment. Furthermore, data on rare diseases remains scarce. Consequently, constructing fully supervised datasets comprising tens of millions of examples proves exceptionally challenging in the medical domain.

To address the challenges of sparse data and weak annotations, researchers have turned to semi-supervised learning approaches. In semi-supervised learning, researchers utilise a small amount of labelled data alongside a large volume of unlabelled data. They employ methods such as pseudo-label self-iteration or Mean Teacher networks[33] to train models. This enhances the stability of the model's feature extraction capabilities. In weakly supervised learning, models require only coarse-grained image-level labels, such as simple presence/absence of lesions or basic bounding boxes. Models then utilise class activation mapping or clustering-constrained attention multi-instance learning (CLAM)[34] to precisely localise lesions and achieve pixel-level segmentation.

4.2. Data Silos and Privacy Protection Mechanisms in Federated Learning

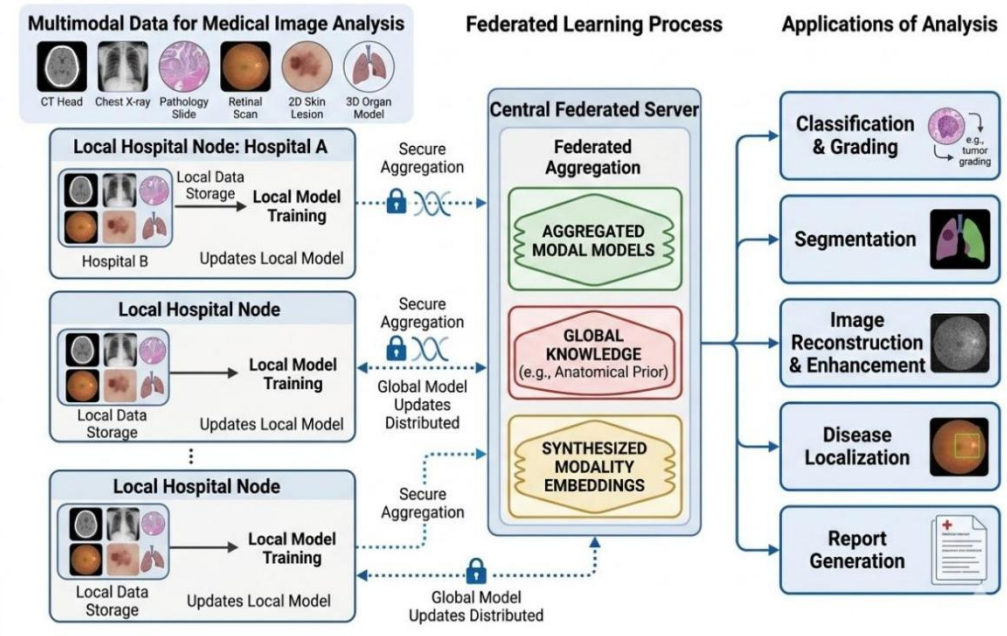

Figure 6. Example of medical image analysis based on Federated Learning

ISSN: 2688-8645

Medical data involves sensitive patient privacy information. Nations impose stringent regulations on cross-hospital sharing and cross-border transfer of such data. Data from a single healthcare institution often exhibits significant bias, manifesting in geographical, ethnic, and scanning equipment variations. Consequently, Deep Learning models trained on single-hospital datasets are prone to overfitting. When validated in other hospitals, their performance often deteriorates substantially.

Federated Learning[35][36] presents an ideal paradigm for resolving this issue. Within the Federated Learning framework, raw medical imaging data strictly remains on the local servers of individual hospitals. The client at each hospital initially trains the model utilizing its local data. Upon completion of training, only the encrypted local model gradients or weight parameters are uploaded to the central server. The central server aggregates these parameters to update the global model, which is subsequently distributed back to the hospital clients, as illustrated in Figure 6. This distributed learning approach, characterized by "stationary data and mobile models," offers two primary advantages. First, it complies with the most stringent regulatory requirements for medical privacy. Second, it indirectly facilitates collaborative training on massive datasets across multiple institutions. Consequently, this represents an inevitable path toward constructing highly stable and broadly adaptable large medical models in the future.

4.3. The Dawn of Multimodal Foundational Models in Medicine

Following the rise of large language models, Medical Imaging is rapidly entering the era of "multimodal foundational large models." Traditional Deep Learning networks predominantly feature standalone architectures designed for single tasks and single modalities. However, recent research, such as Google's 2024 Med-Gemini series models[38], demonstrates the immense potential of deeply integrating visual and textual data. These models inherit the core capabilities of Gemini while being fine-tuned for medical applications through training on 2D and 3D radiology, histopathology, ophthalmology, dermatology, and genomic data. As illustrated in Figure 7.

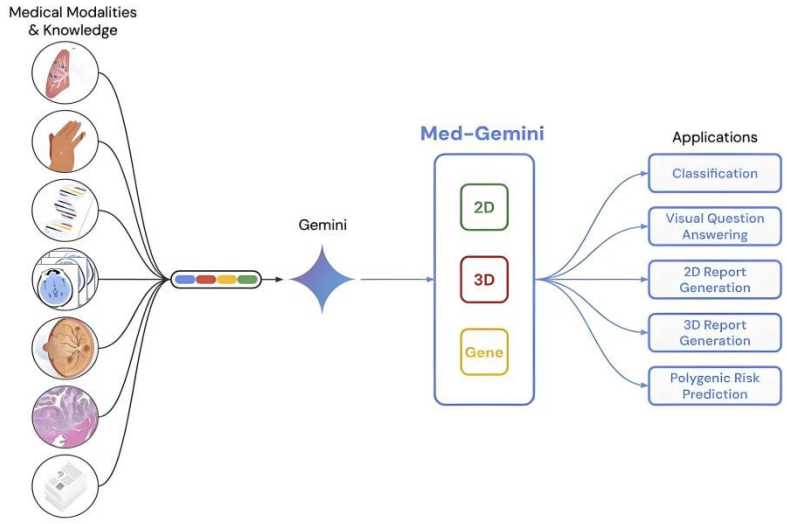

Figure 7. Flowchart of Med-Gemini application

Such multimodal models can simultaneously process diverse data streams, including patients' three-dimensional volumetric imaging data, electronic medical record texts, laboratory test results, and genomic data. They not only perform precise medical visual question-answering across modalities but also autonomously generate imaging diagnostic reports compliant with clinical standards. Furthermore, they are capable of complex pathological logical reasoning. In

ISSN: 2688-8645

the future, this technology will fundamentally transform how clinicians view and analyse medical images.

5. Conclusion

The convergence of Deep Learning and Medical Imaging has ushered in a historic transformation in modern medicine. It no longer relies solely on physicians' subjective experience and historical precedents, but has shifted towards objective data computation. From fundamental signal reconstruction and image noise reduction, through intermediate stages of multi-organ automatic segmentation and registration, to higher-level advanced diagnostic assistance and disease risk prediction, deep neural networks have permeated every facet of radiomics. Looking ahead, as Transformer mechanisms continue to evolve, generative models will effectively supplement data for rare diseases. Federated Learning will break down data silos between hospitals. Coupled with the comprehensive application of multimodal medical large models, Deep Learning technologies will play a decisive role in building a reliable healthcare ecosystem. This healthcare ecosystem will enable users to comprehend model functioning, perceive potential risks, and safeguard patient privacy. It will provide an unceasing stream of technological momentum to enhance global human health and quality of life.

References

[1] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[2] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[3] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1-9.

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[5] Yi X, Walia E, Babyn P. Generative adversarial networks in Medical Imaging: A review[J]. Medical Image Analysis, 2019, 58: 101552.

[6] Milletari F, Navab N, Ahmadi S A. V-net: Fully Convolutional Neural Networks for volumetric medical image segmentation. . [C]//2016 fourth international conference on 3D vision (3DV). Ieee, 2016: 565-571.

[7] Niu Z, Zhong G, Yu H. A review on the attention mechanism of Deep Learning[J]. Neurocomputing, 2021, 452: 48-62.

[8] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 7132-7141.

[9] Chen J, Lu Y, Yu Q, et al. TransUNet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.

[10] Cao H, Wang Y, Chen J, et al. SWIN-UNet: UNet-like pure transformer for medical image segmentation[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022: 205-218.

[11] Hatamizadeh A, Tang Y, Nath V, et al. UNETR: Transformers for 3D medical image segmentation [C] //Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 574-584.

[12] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2223-2232.

[13] Kazerouni A, Aghdam E K, Heidari M, et al. Diffusion models in Medical Imaging: A comprehensive survey[J]. Medical image analysis, 2023, 88: 102846.

[14] Huang Q, Zhao C, Jiang M, et al. Cascade-Net: A new Deep Learning architecture for OFDM detection[J]. arXiv preprint arXiv:1812.00023, 2018.

[15] Wang S, Cheng H, Ying L, et al. DeepcomplexMRI: Exploiting deep residual networks for fast parallel MR imaging with complex convolution[J]. Magnetic Resonance Imaging, 2020, 68: 136-147.

[16] Sun J, Li H, Xu Z. Deep ADMM-Net for compressive sensing MRI[J]. Advances in neural information processing systems, 2016, 29.

[17] Willemink MJ, Noël PB. The evolution of image reconstruction for CT—from filtered back projection to artificial intelligence[J]. European Radiology, 2019, 29(5): 2185-2195.

[18] Chen H, Zhang Y, Kalra M K, et al. Low-dose CT with a residual encoder-decoder convolutional neural network[J]. IEEE transactions on Medical Imaging, 2017, 36(12): 2524-2535.

[19] Kang E, Ye J C. Wavelet domain residual network (WavResNet) for low-dose X-ray CT reconstruction[J]. arXiv preprint arXiv:1703.01383, 2017.

[20] Adler J, Lunz S. Banach Wasserstein GAN[J]. Advances in Neural Information Processing Systems, 2018, 31.

[21] Sandfort V, Yan K, Pickhardt P J, et al. Data augmentation using generative adversarial networks (CycleGAN) to improve generalisability in CT segmentation tasks[J]. Scientific reports, 2019, 9(1): 16884.

[22] Li Z, Liu Q, Li Y, et al. A two-stage multi-loss super-resolution network for arterial spin labelling magnetic resonance imaging[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2019: 12-20.

[23] Gulshan V, Peng L, Coram M, et al. Development and validation of a Deep Learning algorithm for detection of diabetic retinopathy in retinal fundus photographs[J]. JAMA, 2016, 316(22): 2402-2410.

[24] Esteva A, Kuprel B, Novoa R A, et al. Dermatologist-level classification of skin cancer with deep neural networks[J]. Nature, 2017, 542(7639): 115-118.

[25] Zhu W, Liu C, Fan W, et al. Deeplung: Deep 3D dual path nets for automated pulmonary nodule detection and classification[C]//2018 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE, 2018: 673-681.

[26] Li Z, Zhang S, Zhang J, et al. MVP-Net: multi-view FPN with position-aware attention for deep universal lesion detection[C]//International conference on medical image computing and computer-assisted intervention. Cham: Springer International Publishing, 2019: 13-21.

[27] Suk H I, Lee S W, Shen D, et al. Hierarchical feature representation and multimodal fusion with Deep Learning for AD/MCI diagnosis[J]. NeuroImage, 2014, 101: 569-582.

[28] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer International Publishing, 2015: 234-241.

[29] Isensee F, Jaeger P F, Kohl S A A, et al. nnU-Net: a self-configuring method for Deep Learning-based biomedical image segmentation[J]. Nature methods, 2021, 18(2): 203-211.

[30] Wasserthal J, Breit H C, Meyer M T, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images[J]. Radiology: Artificial Intelligence, 2023, 5(5): e230024.

[31] Ma J, He Y, Li F, et al. Segment anything in medical images[J]. Nature communications, 2024, 15(1): 654.

[32] Balakrishnan G, Zhao A, Sabuncu M R, et al. Voxelmorph: a learning framework for deformable medical image registration[J]. IEEE Transactions on Medical Imaging, 2019, 38(8): 1788-1800.

[33] Luo X, Wang G, Liao W, et al. Semi-supervised medical image segmentation via uncertainty rectified pyramid consistency[J]. Medical Image Analysis, 2022, 80: 102517.

[34] Lu M Y, Williamson D F K, Chen T Y, et al. Data-efficient and weakly supervised computational pathology on whole-slide images: [J]. Nature Biomedical Engineering, 2021, 5(6): 555-570.

[35] Kaissis GA, Makowski MR, Rückert D, et al. Secure, privacy-preserving and federated machine learning in Medical Imaging. Nature Machine Intelligence, 2020, 2(6): 305-311.

ISSN: 2688-8645

[36] Rieke N, Hancox J, Li W, et al. The future of digital health with Federated Learning. NPJ Digital Medicine, 3, 119[J]. DOI: https://doi.org/10.1038/s41746-020-00323-1, 2020.

[37] Yan K, Wang X, Lu L, et al. Deeplesion: Automated deep mining, categorisation and detection of significant radiology image findings using large-scale clinical lesion annotations[J]. arXiv preprint arXiv:1710.01766, 2017.

[38] Saab K, Tu T, Weng W H, et al. Capabilities of Gemini models in medicine[J]. arXiv preprint arXiv:2404.18416, 2024.