REVIEW ARTICLE

Artificial Intelligence in Ultrasound Imaging: A Review of Progress from Machine Learning to Large Language Model

Tong Jin [a,1], ***Xiaohu Yu*** [b,1], ***Zheng Ai*** [b], ***Hongcheng Guo*** [b,*]

[a] *School of Life Sciences, Beijing University of Chinese Medicine, Beijing, China;* [b] *School of Data Science, Fudan University, Shanghai, China*

Received September 30, 2025; revision requested October 12, 2025; accepted October 26, 2025

Abstract: Biomedical ultrasound imaging, as one of the most common, safe, and cost-effective modalities in clinical diagnosis, witnesses remarkable progress with the integration of artificial intelligence (AI). Early studies based on traditional machine learning (ML) rely on handcrafted features and classical classifiers to achieve automatic recognition and quantitative analysis of ultrasound images. However, such methods are limited in feature representation capacity and generalizability. With the advent of deep learning (DL), convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based architectures are widely applied to tasks such as segmentation, detection, and lesion classification, significantly improving diagnostic accuracy and robustness. More recently, large language models (LLMs) and multimodal foundation models open new avenues for intelligent ultrasound analysis. These models not only integrate imaging and textual information to support automated report generation and cross-modal reasoning but also offer enhanced interpretability and greater potential for clinical adoption. In this review, we provide a systematic review of the evolution of AI in ultrasound image analysis, spanning from traditional ML to deep learning and LLMs, outlining a complete trajectory of methodological advances.

Key words: Ultrasound imaging; Artificial intelligence; Deep learning; Large language models

Advanced Ultrasound in Diagnosis and Therapy 2025; 04: 483-496 DOI: 10.26599/AUDT.2025.250104

Biomedical ultrasound imaging has become an indispensable tool in modern clinical practice due to its non-invasive nature, low cost, portability, and real-time imaging capability. It plays a critical role in screening, diagnosis, and treatment monitoring across diverse clinical domains, including cardiology, obstetrics, gynecology, hepatology, and oncology [1-2]. However, the interpretation of ultrasound images remains highly operator-dependent and is often challenged by issues such as low signal-to-noise ratio, speckle noise, and variability across machines and patient populations [3]. These limitations have motivated the integration of AI methods to improve the efficiency, accuracy, and consistency of ultrasound image analysis [4].

Figure 1 shows the development trajectory of intelligent ultrasound imaging. Early research is dominated by traditional machine learning approaches, where handcrafted features, such as texture, shape, and statistical descriptors, are combined with classifiers like support vector machines or random forests to detect lesions or characterize tissue properties. While these methods provide initial insights, their reliance on manually engineered features limits generalizability across clinical settings.

The emergence of deep learning revolutionizes ultrasound image analysis. Convolutional neural networks (CNNs) [5-7] enable end-to-end feature learning direct-

[1] *Tong Jin and Xiaohu Yu contributed equally to this work.*

[*] *Corresponding author: School of Data Science, Fudan University, Shanghai, China (Hongcheng Guo).*
e-mail: guohc@fudan.edu.cn (HC G)

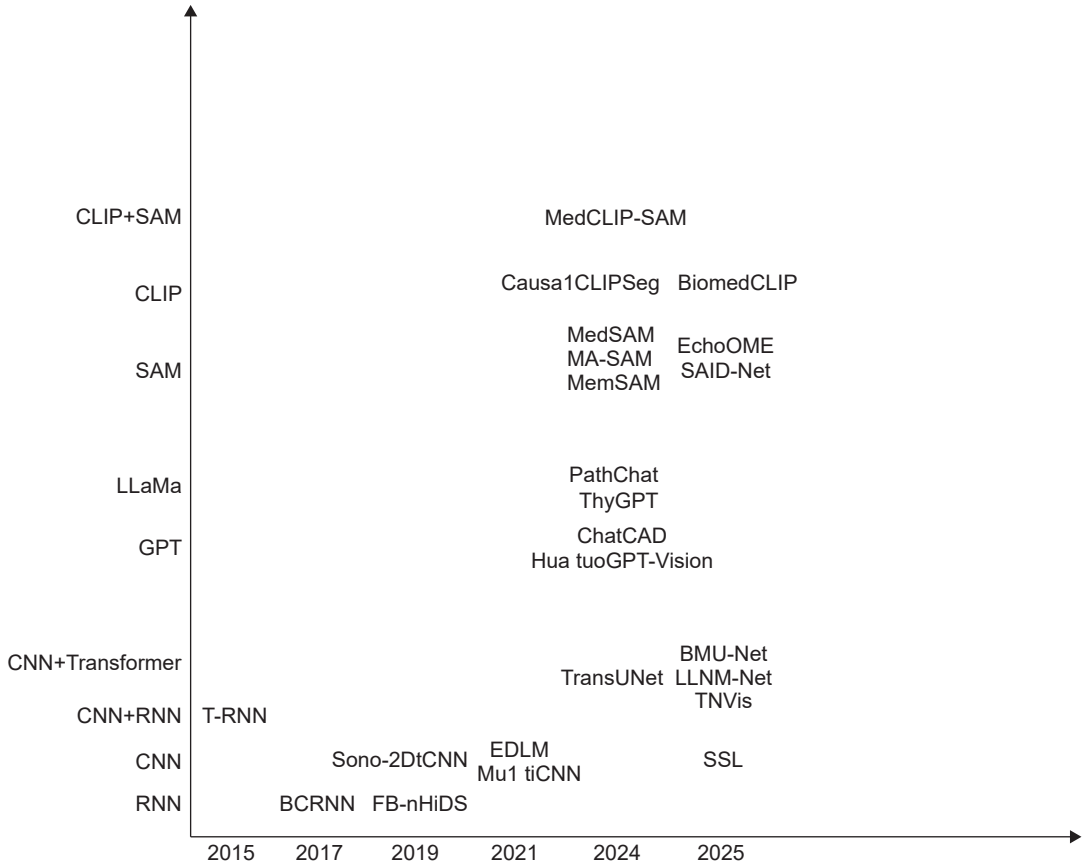

Figure 1 With the advent of deep neural networks, an increasing number of AI ultrasound studies have focused on various variants of RNNs, CNNs, and Transformers. Following the introduction of the large language models, most research has centered on utilizing diverse large foundation models for AI ultrasound applications.

ly from raw images, yielding superior performance in segmentation, classification, and detection tasks. More advanced architectures, including recurrent neural networks (RNNs) [8-9], attention-based mechanisms [10], and transformer models [11], further enhance the capacity to model temporal dynamics in ultrasound videos and capture long-range dependencies in image interpretation [12]. Despite these advances, challenges remain in terms of limited annotated datasets, lack of interpretability, and adaptation to diverse clinical environments [13].

In parallel, the rise of large language models [14-18] has opened new opportunities for intelligent ultrasound analysis. These models extend beyond visual recognition by integrating imaging with clinical text, structured data, and prior knowledge, enabling tasks such as automated report generation [19], visual question answering [20], and multimodal reasoning [21]. Importantly, LLMs provide a pathway toward more explainable AI systems that can align better with clinical workflows and decision-making processes.

As shown in figure 2, we present a comprehensive review of the application of AI in ultrasound imaging, spanning from traditional machine learning approaches to state-of-the-art deep learning and LLM-based methods. We categorize existing work across key tasks, including segmentation, classification, detection, and report generation while highlighting methodological innovations, clinical applications, and current challenges. Finally, we discuss future directions in intelligent ultrasound, including large-scale dataset construction, development of multimodal foundation models, improved interpretability, and clinical translation.

Traditional Machine Learning

Traditional machine learning techniques have been foundational in the development of computer-aided diagnosis (CAD) systems for biomedical ultrasound imaging. These methods typically involve handcrafted feature extraction followed by classification algorithms. Despite the rise of deep learning approaches, traditional ML methods remain relevant due to their interpretability and efficiency in certain clinical contexts.

One of the earliest attempts to apply machine learning in ultrasound-based diagnosis is conducted by Maclin and Dempsey in 1992 [1]. They develop a back-propagation neural network to classify hepatic masses into five categories (hepatoma, metastatic carcinoma, abscess, cavernous hemangioma, and cirrhosis) using 35

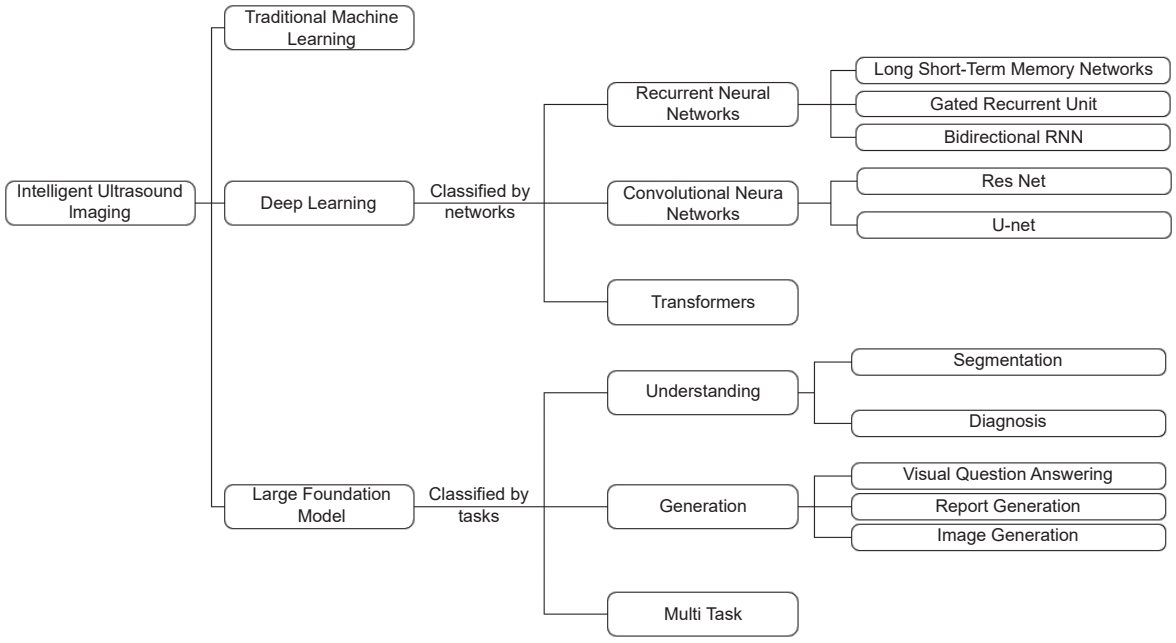

Figure 2 The application of AI in ultrasound imaging spans from traditional machine learning to state-of-the-art deep learning and LLM-based methods. While deep learning methods can be divided by networks such as CNN, RNN and Transformer, LLM-based methods can be classified by tasks.

input features derived from ultrasonographic data and laboratory tests. Although the model falls short of the performance of experienced radiologists (90%), it outperforms radiology residents (50%). This result provides early evidence for the feasibility and potential clinical utility of machine learning in ultrasound diagnosis. While Maclin and Dempsey pioneer the use of artificial neural networks in ultrasound diagnosis, Garra et al. [2] focus on quantitative sonographic texture analysis to distinguish benign from malignant breast lesions. By extracting statistical texture descriptors (e.g., gray-level co-occurrence matrices and fourier-based features) and applying discriminant analysis, their study demonstrates that computer-extracted imaging biomarkers significantly improve diagnostic accuracy. Although based on traditional statistical classifiers rather than modern neural networks, this work represents an important early step toward computer-aided diagnosis based on machine learning in ultrasound imaging.

The Different Features of ML in Ultrasound Imaging

Traditional machine learning approaches in ultrasound imaging have predominantly relied on carefully engineered features to achieve accurate diagnosis. Based on the type of extracted information, these features can be broadly categorized into three groups:

Texture features: Early studies demonstrate that statistical texture descriptors, including gray-level co-occurrence matrices (GLCM), histogram-based features, wavelet coefficients, and Fourier transforms, effectively distinguish benign from malignant breast lesions. Subsequent works [3,22] further refine speckle-emphasized and multi-scale texture features, often in combination with classical classifiers like support vector machines (SVM), to improve lesion discrimination.

Morphological features: The shape, boundary irregularity, orientation, and posterior acoustic characteristics of lesions provide complementary information to texture. Representative studies [23,24] show that quantitative morphological descriptors, either alone or fused with texture features, significantly enhance the accuracy of CAD systems.

Statistical features: Simple first-order statistics, such as mean, variance, skewness, and kurtosis, are widely used in early works [4,25] as robust and computationally inexpensive descriptors of lesion characteristics. These features often serve as a foundation for more complex texture-based representations.

The Different Models of ML in Ultrasound Imaging

Bioedical ultrasound images are often characterized by noise, artifacts, and variable quality. Specific task relies on the choice of classifier, such as Support Vector Machines (SVM), Random Forests (RF), and Logistic Regression, each offer unique advantages that make them suitable for biomedical ultrasound image analysis.

SVM is well known for its ability to handle high-dimensional feature spaces and robustness to overfitting with limited data. The fundamental principle of an SVM is to identify the "optimal hyperplane" that best sepa-

rates data points of different classes with the maximum possible margin. In biomedical imaging, this is crucial as we often extract a vast number of features from images—such as texture, shape, histogram intensities, and wavelet coefficients—creating a high-dimensional feature space. SVM excels in this context through the use of the "kernel trick," which allows it to implicitly map data into even higher-dimensional spaces where a separating hyperplane can be found, without the computational burden of explicitly performing the transformation. This enables SVM to construct highly complex, non-linear decision boundaries effectively, which is often the case in biomedical imaging [26].

RF provides strong generalization, interpretability through feature importance, and resilience to noisy data. Random Forest is an ensemble learning method that operates by constructing a multitude of decision trees during training and outputting the mode of the classes (classification) or mean prediction (regression) of the individual trees. This "wisdom of the crowd" approach is the source of its exceptional generalization ability, enabling it to model complex, non-linear relationships without easily overfitting, making it effective for heterogeneous ultrasound signals [27].

Logistic Regression, though relatively simple, offers probabilistic outputs and clear interpretability. Despite its relative simplicity, LR holds enduring value in clinical decision-support systems due to two paramount qualities. A clinician can then integrate this probability with other patient-specific factors, such as age, family history, and serum biomarkers to make a more nuanced and risk-aware final decision, which are valuable in clinical decision-making [28].

These classifiers are computationally efficient, require smaller datasets compared to deep learning methods, and align well with the need for transparency and reliability in medical applications.

In summary, these works illustrate that traditional machine learning in ultrasound imaging is fundamentally structured around the design and selection of informative features. By categorizing studies according to texture, morphology, statistical, this review highlights the evolution and relative strengths of each feature type, setting the stage for the transition to deep learning-based methods that can learn hierarchical features directly from raw images.

Deep Learning Methods

ML approaches, which rely on handcrafted features meticulously designed by domain experts, while promising, is fundamentally limited by the quality and generality of these manual features, often resulting in models that are brittle and fail to generalize across different datasets and scanning protocols. The breakthrough emerges with the advent of DL. While a significant portion of existing reviews on deep learning in medical ultrasound are organized by clinical application tasks (e.g., segmentation, classification, detection), this review adopts a model-centric taxonomy, categorizing the literature primarily by the underlying neural network architectures. We argue that this approach offers a more insightful perspective into the technological evolution of the field. As shown in table 1, it allows us to juxtapose how a single architecture paradigm empowers diverse applications. Furthermore, it facilitates a deeper discussion on the inherent strengths, limitations, and common optimization strategies specific to each architectural family, thereby providing a more valuable resource for methodology-oriented researchers aiming to push the technical boundaries forward.

Recurrent Neural Networks

While convolutional neural networks (CNNs) have

Table 1 Comparison of DL based methods

Method	Framework	Task	Region	Dataset	Year
BCRNN	BiLSTM + ASM	Segmentation	Prostate	530 slices from 17 trans-rectal ultrasound	2017
MultiCNN	CNN	Classification	Breast	10,815 multimodal breast-ultrasound images of 721 biopsy-confirmed lesions from 634 patients	2021
EDLM	5 CNNs (Se-ResNet)	Classification	Gallbladder	3,705 sonographic gallbladder images from 1141 patients	2021
Sono-2DtCNN	CNN (SonoNet-64)	Classification	Fetus	Video clips from 25 full-length scans (average 45.7 ± 11.6 minutes)	2019
T-RNN	J-CNN + LSTM	Classification	Fetus	631 videos with 50,624 US images	2015
FB-nHiDS	3D FCN + BiLSTM	Segmentation	Fetus, gestational sac, placenta	104 prenatal ultrasound volumes from 104 volunteers	2019
SSL	HED + U-net	Segmentation	Cardiac chamber	4,569,266 images from 8,843 transthoracic echocardiograms	2025
BMU-Net	CNN+ Transformer	Classification	Breast	19,360 images of 5,216 breasts from 5,025 patients	2025
TransUNet	U-net+ Transformer	Segmentation	Multi-organ	3,779 axial contrast-enhanced abdominal clinical images	2024
LLNM-Net	YOLO-v8 + U-net++ + Transformer	Classification	Thyroid	Multimodal data from 29,615 patients and 9836 surgical cases	2025
TNVis	YOLO-v8 + Swin-Unet	Segmentation	Thyroid	9404 2D static ultrasound images from 5173 cases	2025

revolutionized the analysis of static ultrasound images by extracting spatial features, a significant portion of ultrasound diagnostics relies on interpreting dynamic sequences—such as cardiac motion, fetal activity, or blood flow. For these tasks, Recurrent Neural Networks (RNNs) [29-30], and particularly their advanced variants Long Short-Term Memory (LSTM) [31] and Gated Recurrent Unit (GRU) [32-33] networks, offer a powerful paradigm by inherently modeling temporal dependencies across video frames. This section reviews the application of RNN-based architectures in medical ultrasound, focusing on their unique ability to capture the spatiotemporal dynamics that are crucial for functional assessment and real-time monitoring.

BCRNN [8] addresses the critical challenge of boundary incompleteness in automatic prostate segmentation from ultrasound images, a key issue hindering accurate prostatic disease diagnosis and therapeutic planning (e.g., biopsy guidance, surgical anatomical modeling). Traditional methods are found inadequate: bottom-up approaches lack global shape prior to complement ambiguous/occluded boundaries, while top-down methods rely on hand-crafted descriptors and suffer from local information loss during shape fitting. To resolve these limitations, they propose a novel framework integrating feature extraction and shape prior learning seamlessly, centered on three core modules: first, a Boundary Completion RNN (BCRNN) based on bidirectional Long Short-Term Memory (BiLSTM), which serializes static 2D ultrasound images into dynamic sequences (via Cartesian-to-polar coordinate transformation) to sequentially infer missing boundaries using accumulated shape knowledge, eliminating the need for hand-designed features; second, a multi-view fusion strategy that merges predictions from three different serialization starting points to mitigate bias caused by varying sequence contexts; third, a multiscale Auto-Context scheme that cascades BCRNNs (coarse-to-fine scales), where each level's prediction is concatenated with the original image to refine boundary details, supplemented by an auxiliary ASM (trained on 300 annotated shape maps) for final gap-filling. Validated on 530 transrectal ultrasound (TRUS) slices, the framework outperforms pre-trained competitors.

Generally, RNNs are more suitable for time-series forecasting, natural language processing (e.g., machine translation, text generation), and audio analysis. Researchers tend to combine CNN and RNN to leverage both advantages. As shown in figure 3, by using a CNN to extract spatial features from images and then passing those features into an RNN, the model can not only understand what is in each frame but also how things evolve over time. This hybrid approach improves performance on tasks involving image sequences (such as videos or dynamic medical imaging) by capturing both spatial detail and temporal/contextual dependencies. T-RNN [9] is a typical example of combining RNN and CNN. The J-CNN in the proposed framework leverages multi-task joint learning with knowledge transfer across similar ultrasound detection tasks to address the insufficiency of training data, effectively extracting discriminative spatial features from ultrasound frames and accurately capturing key anatomical structures in regions of interest for standard plane identification. The LSTM exploits temporal information in consecutive ultrasound video frames, addressing the vanishing gradient issue of traditional RNNs via its gate mechanism to model inter-frame contextual correlations, which helps filter single-frame noise and improve the consistency and robustness of standard plane detection.

Convolutional Neural Networks

Biomedical ultrasound is among the most widely used imaging modalities in clinical practice, favored for its real-time capability, safety, low cost, and portability. However, ultrasound images suffer from unique challenges such as speckle, shadowing, and low signal-to-noise ratio, which complicate automatic analysis. In this

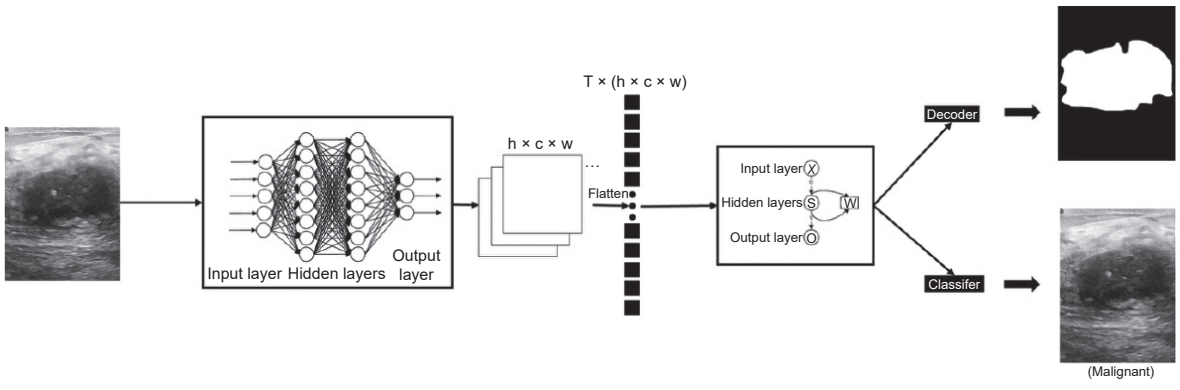

Figure 3 The CNN+RNN methods leverage a CNN to extract spatial features from images and then passing those features into an RNN, the model can not only understand what is in each frame but also how things evolve over time.

context, Convolutional Neural Networks (CNNs) [5,34] have revolutionized the field by enabling end-to-end learning of spatial features relevant for tasks such as lesion detection and classification, organ segmentation, and standard plane detection. In the past few years, numerous studies have applied CNNs in ultrasound imaging of the breast, heart, fetal anatomy, thyroid, and other organs, with improving performance owing to better network designs, data augmentation, and domain adaptation [35].

ResNet [36] addresses the long-standing challenge of accurate and accessible diagnosis of biliary atresia (BA) using an ensembled deep learning model (EDLM) based on sonographic gallbladder images. BA diagnosis is critically time-sensitive, yet conventional methods are limited: biochemical markers are resource-intensive, and ultrasound is highly dependent on expert experience, leading to delayed diagnosis. To tackle this, the team constructs a multi-center dataset involving 11 hospitals, including 3,705 sonographic images (from 330 BA and 811 non-BA patients) for model training and 841 independent images (from 102 BA and 196 non-BA patients) for external validation. The EDLM integrates five CNNs, based on the Se-ResNet architecture, trained via five-fold cross-validation (each CNN trained on 4 subsets and validated on 1 unique subset, with final predictions averaged to enhance robustness), and demonstrates superior diagnostic performance compared to human experts. Notably, combining EDLM predictions with expert diagnoses significantly improves expert sensitivity while only moderately reducing specificity, minimizing missed BA cases. The model also exhibits strong robustness across varying scanning conditions and extends to practical clinical scenarios: it maintains expert-level performance when using smartphone photos of sonographic images and fully automated video-based diagnosis, supported by a dedicated smartphone app for remote use. This study not only fills the gap of AI-based BA diagnosis using sonographic images but also provides a scalable solution to alleviate expertise shortages in rural and underdeveloped regions, offering substantial clinical value for timely BA detection and improved patient outcomes.

Sono-2DtCNN [37] proposes a multi-stream framework to leverage the spatio-temporal dynamics of fetal anomaly ultrasound videos across different temporal scales, with the Sono-2DtCNN serving as the base model for each stream. It incorporates three individual streams trained on video clips with varying frame rates to capture distinct temporal changes: one stream uses near-consecutive frames at FR/2 to model fine short-term temporal dynamics, another at FR/4 (the original frame rate for single-model training) to capture standard short-term motion, and the third at FR/8 to capture coarse long-term temporal dependencies. When applied to 10 unlabeled full-length scans, the automated classification shows high correlation with manually computed workflow statistics, demonstrating the practicality of the proposed spatio-temporal partitioning and description approach for analyzing full-length fetal anomaly ultrasound scans.

Fully Convolutional Network (FCN) [6] represents a pioneering and foundational architecture for end-to-end semantic segmentation. Its key innovation lies in transforming classical classification-oriented CNNs into fully convolutional models by replacing the final fully connected layers with convolutional layers. This crucial modification enables the network to accept input images of any size and generate correspondingly sized spatial output maps, rather than a single classification label. FCNs recover fine-grained spatial information through learned transposed convolutions (or deconvolutions), which perform upsampling on the coarse feature maps. Additionally, it incorporates skip connections from earlier, higher-resolution layers to combine deep, semantically strong features with shallow, detailed features, thereby refining the segmentation output. As a milestone architecture, FCN establishes the encoder-decoder paradigm as a standard for dense prediction tasks in computer vision, directly influencing subsequent, more complex architectures.

The integration of FCN with RNN creates a powerful architecture for spatiotemporal analysis. The FCN excels at extracting dense, pixel-level spatial features from each individual frame, preserving crucial locational and structural information. The RNN, typically an LSTM or GRU, then processes this sequence of feature maps to model complex temporal dependencies and dynamic evolution across time. The primary benefit of this union is the ability to perform fine-grained, pixel-wise prediction while simultaneously understanding contextual changes over time [38]. This makes it exceptionally well-suited for tasks like video semantic segmentation, dynamic medical image analysis (e.g., quantifying cardiac function from ultrasound videos), and any application requiring both precise spatial understanding and robust temporal modeling within a sequence of images. The entire system can be trained end-to-end, optimizing the spatial feature extractor for the ultimate temporal task.

While FCN introduces the core idea of end-to-end, pixel-wise semantic segmentation by replacing fully connected layers with convolutional layers and using upsampling, U-Net [7] enhances this architecture via a symmetric encoder-decoder (“U-shaped”) design. The contracting encoder path captures contextual informa-

tion, while the expanding decoder path enables precise localization. The key innovation of U-Net lies in its skip connections, which bridge corresponding layers in the encoder and decoder. These connections fuse high-resolution feature maps from the encoder with the upsampled outputs in the decoder, thereby preserving fine-grained spatial details that are crucial for accurate pixel-wise segmentation. Originally developed for cell tracking in microscopic images, U-Net's exceptional performance and elegant design lead to its widespread adoption and inspire numerous variants across diverse medical imaging modalities and beyond.

SSL [39] proposes a label-free self-supervised learning pipeline for the segmentation of cardiac chambers and calculation of cardiac function parameters from cardiac ultrasound images, aiming to address the limitations of traditional cardiac ultrasound analysis—specifically the labor intensity and poor reproducibility of manual annotations, as well as the heavy reliance of supervised learning on high-quality manual labels. The pipeline integrates traditional computer vision techniques, clinical domain knowledge, and deep learning models (UNet for semantic segmentation and Holistically Nested Edge Detection (HED) network for enhancing blurry ultrasound edges). It first generates weak labels without manual input, then refines these labels through successive training steps involving early stopping and self-learning, ultimately achieving precise segmentation. This work realizes clinical-valid, scalable cardiac ultrasound analysis without manual annotations, offering a practical solution for automated cardiac assessment and a paradigm for label-free medical image segmentation.

Transformers

Over the past decade, deep learning, especially CNNs has achieved remarkable success in a variety of ultrasound tasks: classification and segmentation. CNNs excel at capturing local patterns through convolutional filters, but they are intrinsically limited in modeling long-range dependencies, global context, and relationships across distant image regions—features that can be critical in ultrasound where context matters.

The Transformer architecture [10], originally developed for natural language processing and later adapted to vision, provides powerful self-attention mechanisms for capturing long-range interactions and global context. In medical imaging broadly, Transformers have begun to show strong performance in segmentation, classification, detection, registration, reconstruction, and other tasks. However, their application to ultrasound imaging is still emerging, and specific challenges arise in the ultrasound context.

The CNN-Transformer hybrid architecture synergistically combines the strengths of both models to create a powerful framework for visual understanding, especially in complex domains like medical imaging [40]. The CNN component acts as a proficient local feature extractor, efficiently capturing fine-grained patterns, textures, and edges from pixel data due to its innate inductive biases like translation invariance. However, its ability to model long-range dependencies between distant parts of an image is limited. The Transformer component excels at global context modeling. Its self-attention mechanism allows every part of the feature map to interact directly with all others, enabling the model to understand complex relationships and the overall structural context of an entire image.

By integrating them, the hybrid model leverages the CNN as a front-end to create meaningful feature maps from raw pixels and uses the Transformer as a back-end to interpret those features within a global context. This results in a model that achieves superior performance by simultaneously appreciating minute details and holistic anatomy. Furthermore, the architecture offers exceptional flexibility for fusing multimodal data using cross-attention mechanisms, making it exceptionally well-suited for clinically translatable AI systems that require robust, explainable, and comprehensive analysis [12].

Yolo-v8 and U-net are the most popular foundation framework for biomedical imaging [41]. TransUNet [40] is one of the first to combine Transformer and U-net. It leverages the Transformer to encode tokenized image patches from a CNN feature map as the input sequence for extracting global contexts. On the other hand, the decoder upsamples the encoded features which are then combined with the high-resolution CNN feature maps to enable precise localization. LLNM-Net [11] addresses the critical clinical challenge of insufficiently precise preoperative prediction of lateral lymph node metastasis (LLNM) in thyroid cancer by developing an explainable multimodal Transformer based model. This bidirectional attention-based model integrates multimodal data, including preoperative ultrasound images, radiological reports, pathological findings, and demographic information. LLNM-Net leverages optimized YOLO-v8 and U-Net++ for nodule/thyroid segmentation, and a novel central point distance transformation to extract precise locational and morphological features; its core Thyroid Multimodal Deep Learning (TMDL) transformer enables cross-modal information fusion via four bidirectional attention blocks and twelve self-attention blocks. Notably, LLNM-Net outperformed both human experts and previous models.

Large Foundation Models

Deep learning approaches, particularly convolutional

and transformer-based architectures, have markedly improved ultrasound image analysis by enabling automated feature extraction, segmentation, and classification. However, these models remain largely task-specific, requiring extensive annotated data and offering limited generalization across diverse clinical scenarios. Moreover, they often fail to integrate multimodal clinical information, such as radiology reports or patient histories, which are essential for real-world decision-making. These limitations have motivated a paradigm shift toward large-scale foundation models [14-15]. By leveraging massive pretraining corpora, cross-modal alignment, and instruction-tuning, large models extend beyond single tasks to provide more generalizable, versatile, and clinically relevant solutions, thereby representing the next frontier in AI-driven ultrasound imaging.

The development of LLMs has revolutionized natural language processing and multimodal reasoning. Models such as GPT-4 [16], PaLM-E [17], and LLaVA [18] have demonstrated remarkable abilities in understanding complex language prompts, integrating visual inputs, and generating medically relevant responses. Transformer based models, especially LLMs have demonstrated the capability to perform tasks beyond the reach of traditional deep learning approaches, including biomedical image segmentation at cellular granularity [13,42]. Recent works tend to explore the integration of vision-language models (VLMs) in the biomedical domain, particularly for tasks like medical image captioning, question answering, and report generation. However, most of these approaches remain limited to coarse-grained image-level reasoning, lacking the capacity for fine-grained, pixel-level understanding essential for clinical pathology.

Understanding

Biomedical images understanding, including segmentation and diagnosis, is one of the most important applications of large foundation models. Most of the biomedical image segmentation works are based on SAM [43]. As shown in figure 4, SAM adopts a promotable segmentation framework that links an image encoder, a prompt encoder, and a lightweight mask decoder. Its transformer-based image encoder captures rich visual features, while the decoder efficiently generates segmentation masks guided by prompts such as points, boxes, or masks. MedSAM [44] is one of the first to apply Vision Foundation Models (VFM) to biomedical image segmentation. MedSAM introduces a self-attention-based large model. Ultrasound images often exhibit low contrast and blurred boundaries, making it difficult to delineate structures accurately. Additionally, the complex and variable nature of anatomical features in ultrasound images adds to the segmentation difficulty. Small-sized lesions or subtle structures are particularly challenging to detect due to their size and the presence of noise. These challenges underscore the need for specialized adaptations and training of SAM to improve its applicability and performance in medical ultrasound image segmentation [45-47]. To address these challenges, researchers have developed various fine-tuning strategies and adaptation frameworks. Table 2 compares the different SAM based methods.

EchoONE [48] introduces a unified framework for multi-view cardiac ultrasound segmentation that addresses the fundamental challenge of processing structurally different echocardiographic views (such as two-chamber and four-chamber) within a single model architecture. Its innovative Prior-Composable Mask Learning module (PC-Mask) automatically generates semantic-aware dense prompts to reduce manual prompting dependency, while the Local Feature Fusion and Adaptation module (LFFA) enhances SAM's adaptability to ultrasound-specific characteristics. The framework demonstrated exceptional performance on large-scale multi-center datasets comprising 22,044 private images and public CAMUS data, achieving strong cross-center generalization with a Dice coefficient of 73.94% on the external HMC_QU test set, significantly outperforming traditional view-specific models.

SAID-Net [49] enhances SAM by integrating Implicit Neural Representations (INR) to address the chal-

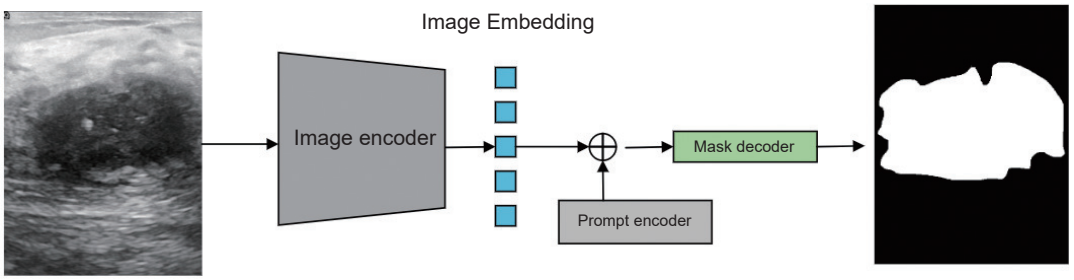

Figure 4 SAM adopts a promotable segmentation framework that links an image encoder, a prompt encoder, and a lightweight mask decoder. Its transformer-based image encoder captures rich visual features, while the decoder efficiently generates segmentation masks guided by prompts such as points, boxes, or masks.

Table 2 Comparison of different SAM variants

Method	Framework	Region	Dataset	Year
MedSAM	SAM + Self-attention fine-tuning	Multi-modal medical images	1,570,263 image–mask pairs, 10 modalities, 86 internal & 60 external tasks	2024
MA-SAM	SAM + 3D adapter	CT, MRI, ultrasound and other 3D medical images	Multiple 3D medical datasets cross-modal tasks	2024
MemSAM	SAM + Spatio-temporal memory module + Noise-robust prompts	Cardiac ultrasound	Echocardiography video dataset	2024
EchoONE	SAM + PC-Mask + LFFA module	Multi-view cardiac ultrasound (2/4-chamber)	Multi-center 22,044 private images + public CAMUS data	2025
SAID-Net	SAM + INR + Hiera encoder + Attention decoder	Cardiac ultrasound	CAMUS, EchoNet-Dynamics	2025
MedCLIP-SAM	BiomedCLIP + SAM + Residual UNet	Multi-modal medical images	MedPix dataset (image-text pairs), with pseudo-labels	2024

lenges of segmenting complex cardiac anatomy and subtle boundaries in echocardiography sequences. It employs a Hiera-based encoder for multi-scale feature extraction and a Mask Unit Attention Decoder for capturing fine details, complemented by orthogonalization to boost feature diversity. When tested on CAMUS and EchoNet-Dynamics datasets, SAID-Net achieved state-of-the-art performance with a Dice Similarity Coefficient of 93.2% and Hausdorff Distance of 5.02 mm on CAMUS, demonstrating its superior accuracy and robustness for cardiac ultrasound segmentation.

Beyond SAM, numerous studies have leveraged CLIP [50] for biomedical ultrasound image segmentation. BiomedCLIP [51] addresses these by using PubMedBERT as the text encoder and a Vision Transformer (ViT) as the image encoder, incorporating domain-specific adaptations tailored to biomedical vision-language processing. BiomedCLIP has demonstrated superior performance compared to previous biomedical vision-language models across a wide range of standard datasets. CausalCLIPSeg [52] leverages CLIP to align textual descriptions with image pixels. The authors address key challenges in medical image analysis, such as ambiguous lesion boundaries and spurious correlations caused by confounding factors. The proposed method includes CLIP-driven text and vision encoders for multi-modal feature extraction, a tailored cross-modal decoder that achieves text-pixel alignment via cross-correlating projected textual features with upsampled visual features, and a causal intervention module. CausalCLIPSeg achieves state-of-the-art performance, demonstrating effective transfer of CLIP's semantic knowledge to pixel-level segmentation and improved robustness against misleading contextual biases.

MedCLIP-SAM [53] proposes a framework integrating BiomedCLIP and SAM to solve low data efficiency, poor generalizability, and limited interactability in medical image segmentation, supporting zero-shot and weakly supervised settings. It involves three core steps: fine-tuning BiomedCLIP on the MedPix dataset via the proposed DHN-NCE loss (addressing small-batch inefficiency and similar sample discrimination), achieving zero-shot segmentation by using gScoreCAM (first applied in medical imaging) to generate saliency maps (post-processed with CRF for SAM bounding box prompts), and optimizing via weakly supervised Residual UNet training with zero-shot pseudo-masks.

Beyond fine-tuning existing pre-trained models on biomedical ultrasound data, a growing number of researchers are focusing on building ultrasound specific foundation models. Foundation models represent a paradigm shift beyond traditional AI modeling, they are large-scale, self-supervised models trained on vast, diverse datasets, capable of adapting to a wide range of downstream tasks. One notable example is the Universal Ultrasound Foundation Model (USFM) [54]. USFM is a self-supervised foundation model tailored specifically for ultrasound image analysis, it was pre-trained on an expansive multi-organ, multi-center, and multi-device ultrasound dataset containing over two million images, using an organ-balanced sampling strategy to ensure equitable learning across anatomical regions. It demonstrates strong generalization across diverse downstream tasks, including segmentation, classification, and image enhancement. MedSegX [55] is a vision foundation model developed to address the challenges of generalist medical image segmentation in open-world scenarios. MedSegX introduces a model-agnostic approach called Contextual Mixture of Adapter Experts (ConMoAE), which enables the model to adapt to various medical segmentation tasks effectively. MedSegDB employs a tree-structured hierarchy to organize the data, facilitating comprehensive training across a wide range of medical imaging modalities and organ systems. This work underscores the potential of MedSegX as a versatile,

task-agnostic solution for medical image segmentation, capable of adapting to diverse and previously unseen tasks in open-world environments.

Instead of models themselves, some researchers focus on bias mitigation tasks in models when segmenting ultrasound images. Apple [56] addresses the critical issue of algorithmic fairness in models for ultrasound image diagnosis, particularly in lesion segmentation. The authors note that both train-from-scratch and pretrained models often exhibit performance disparities across sensitive attributes like sex and age, leading to biased diagnoses for different patient subgroups. To mitigate this unfairness without modifying the base model's parameters, they propose a novel approach called APPLE (Adversarial Protected attribute aware Perturbations on Latent Embeddings). Extensive experiments on both public and in-house ultrasound datasets demonstrate that APPLE successfully improves segmentation performance and enhances fairness across all tested sensitive attributes and various model architectures, contributing to the development of more equitable AI-powered healthcare systems.

Besides segmentation, diagnosis is another significant application of Biomedical LLMs. Originally, researchers primarily employed LLMs for basic biomedical image classification, such as tumor detection or disease screening [57]. One of the first attempta to evaluate in-context learning models on biomedical images uses GPT-4V on cancer image processing with in-context learning on three cancer histopathology tasks of high importance [58]. Results prove that LLMs can perform biomedical image classification tasks.

Classification of Benign and Malignant is one of the most important tasks of biomedical ultrasound image diagnosis. Researchers [59] develop and validate transformer-based models using DeiT architecture initialized with ImageNet pretraining. The models demonstrated robust performance across centers, ultrasound systems, histological diagnoses, and patient age groups, significantly outperforming both expert and non-expert examiners on all evaluated metrics. Furthermore, in a retrospective triage simulation, AI-driven diagnostic support reduced referrals to experts by 63% while significantly surpassing the diagnostic performance of the current practice. These results demonstrate that transformer-based models exhibit strong generalization and above human expert-level diagnostic accuracy, with the potential to alleviate the shortage of expert ultrasound examiners and improve patient outcomes. In addition, some researchers focus on the use of LLMs to assess the quality of ultrasound imaging. Ultrasound-QBench [60] decompose the quality assessment task into three dimensionalities: qualitative classification, quantitative scoring, and comparative assessment. Results show that LLMs possess preliminary capabilities for low-level visual tasks in the classification of ultrasound image quality.

Generation

In biomedical imaging, the generation capability of Large Language Models (LLMs) is primarily used not only to create the images themselves, but also to produce textual outputs [61-62]. Acting as an intelligent interpreter, an LLM takes visual features (extracted by a vision encoder) and generates comprehensive radiology reports, answers clinical questions through Visual Question Answering (VQA) about the images [63] and provides data-driven health recommendations. This application bridges the gap between complex visual medical data and actionable clinical language, enhancing workflow efficiency and decision support.

Instead of generating structured report from ultrasound reports [64], LLMs are increasingly applied to automated ultrasound report generation by integrating image analysis with clinical language production. Typically, a vision encoder first extracts visual features from ultrasound images. These features are then mapped into the embedding space of an LLM, which functions as a powerful text decoder. The LLM, often fine-tuned on medical corpora, generates coherent, structured reports by describing findings, measurements, and impressions based on the visual input. This approach leverages the capability of LLM to produce fluent and clinically relevant text, significantly reducing radiologists' documentation burden. Advanced systems further incorporate context-aware prompting and multi-modal alignment techniques to enhance report accuracy, consistency, and adherence to clinical standards, demonstrating performance comparable to or even surpassing human experts in specific tasks. ChatCAD [65] uses medical knowledgeof LLMs and reasoning to enhance Computer-aided diagnosis (CAD) network outputs, such as diagnosis, lesion segmentation, and report generation, by summarizing information in natural language. The generated reports are of higher quality and can improve the performance of vision-based CAD models. This approach shows the potential of LLMs to revolutionize clinical decision-making and patient communication.

ThyGPT [66] proposes a multimodal generative pretrained transformer (GPT) model designed to assist in thyroid nodule diagnosis and management, addressing critical limitations of traditional AI-based CAD systems. Conventional diagnosis relies heavily on radiologists' experience and traditional CAD models suffer from "black box" (no diagnostic rationale) and "mute box" (no interaction) issues, causing over-diagnosis and over-

treatment. Built on the LLaMA3 model and Transformer architecture, ThyGPT proposes the concept of AIGC-CAD in the field of medical image diagnosis. By combining AIGC-CAD techniques, the ThyGPT model can function as an AI copilot model that intelligently interacts with radiologists, automates thyroid nodule risk assessment, provides decision support, and detects potential errors in diagnostic reports.

UltrasoundSG [19] is the first to combine semantic scene graphs (SGs) with LLMs. It proposes a framework centered on SGs for US images: first, it uses the transformer-based one-stage model RelTR to generate US SGs directly, which capture key anatomical structures and their relationships via <entity-predicate-entity> triplets. Then, the SG paired with side information (left/right neck) and probe movement data is input to LLMs to perform two tasks: generating user-friendly US image summaries and providing natural-language scanning guidance to locate missing anatomies.

Li et al. [67] propose a multimodal LLM framework ultrasound report generation by leveraging fuzzy theory to extract essential anatomical knowledge from statistical features, thereby providing more accurate and context-aware guidance throughout the report generation process. Experiments are conducted on both a publicly available dataset and a proprietary dataset. Results demonstrate that this approach consistently achieves state-of-the-art performance across multiple evaluation metrics, highlighting its robustness and adaptability. These findings underscore the potential of LLMs in advancing the accuracy and clinical applicability of ultrasound report generation.

PathChat [68] is a multimodal generative AI copilot specifically tailored for human pathology, aiming to address the gap between visual analysis capabilities and natural language interaction in computational pathology. Researchers curate a large-scale pathology-specific instruction dataset comprising 456,916 visual-language instructions covering diverse formats and sources. The model architecture integrates three core components: a pathology-optimized vision encoder fine-tuned from the UNI foundation model, a multimodal projector module aligning visual features with text embeddings, and a 13-billion-parameter Llama 2 model, which underwent two-stage training. This work provides a foundational framework for multimodal AI applications in pathology and releasing PathQABench and training code to facilitate further research.

When solving biomedical VQA tasks, existing multimodal large language models (MLLMs) like GPT-4V suffers from insufficient quantity and quality of biomedical vision-text data. HuatuoGPT-Vision [20] addresses the limitation by refining high-quality medical image-text pairs from PubMed through a rigorous pipeline—including text filtering, image filtering and deduplication. They then employed an "unblinded" MLLM to denoise and reformat the data, generating two types of VQA samples to construct the dataset. Experimental validations show that HuatuoGPT-Vision demonstrates superior performance among open-source models on biomedical VQA tasks.

LLaVA-Ultra [69] is a fine-tuned Multimodal Large Language Model designed for Chinese medical visual question answering (Med-VQA), specifically for ultrasound data. To address the limitations of general and medical-specific VLMs that often produce vague or visually-irrelevant answers, the authors propose a novel architecture featuring a fine-grained vision encoder for enhanced medical semantic understanding. They also tackle data redundancy common in medical reports by using a weighted scoring and knowledge distillation method to adaptively select the most relevant images for a given text. The model is efficiently tuned on a large-scale, high-quality Chinese ultrasound dataset with instructions crafted from doctors' text, resulting in a robust system that achieves new state-of-the-art performance across multiple Med-VQA benchmarks.

Multi Task

In addition to conventional tasks such as diagnosis, classification and segmentation, recent research efforts have increasingly explored novel and broader applications of LLMs and multimodal models in biomedical imaging. Researchers tend to focus on multi-task frameworks that simultaneously address tasks like report generation, disease localization, visual question answering, and segmentation [21,70-72]. U2-Bench [73] designed eight ultrasound-related multi-organ tasks to comprehensively test the ultrasound understanding capabilities of large models. BiomedGPT [74-75] is one of the first open-source, lightweight vision–language foundation model tailored for diverse biomedical tasks. Inspired by encoder–decoder transformer architecture like BART [76], BiomedGPT is implemented with a BERT-style encoder over corrupted text and a GPT-style left-to-right autoregressive decoder. It processes multi-modal inputs through self-supervised and multi-task pretraining spanning images, text, and image–text pairs. BiomedGPT achieves state-of-the-art performance in most experiments across modalities including MRI, CT, pathology, and microscopy, while remaining computing-efficient and accessible. It excels in tasks such as radiology VQA with a low error rate.

Future work

For the future development of intelligent ultrasound,

we should explore several key directions, including the construction of large-scale datasets, the development of multimodal large language models, the enhancement of model interpretability, and the advancement of clinical translation. Large-scale datasets provide diverse and high-quality data to support robust model training and improve generalization across various clinical scenarios. Multimodal large language models integrate information from multiple sources, such as images, text, and patient data, enabling comprehensive and context-aware analysis. Enhancing interpretability addresses the "black box" nature of AI models, fosters trust, and facilitates adoption in clinical practice. Finally, advancing clinical translation bridges the gap between research and real-world applications, ensuring that intelligent ultrasound systems effectively improve diagnostic accuracy, workflow efficiency, and patient outcomes.

Conclusions

This review summarizes the evolution of AI in biomedical ultrasound imaging, from traditional machine learning with handcrafted features to deep learning architectures and, most recently, large language models and multimodal foundation models. While these advances have enabled robust feature learning, improved diagnostic accuracy, real-time applications, and greater task generalization, challenges remain, such as limited annotated data, insufficient model generalizability, and poor interpretability remain obstacles to real-world deployment. Future efforts should focus on developing ultrasound-specific foundation models, strengthening multimodal learning and interpretability, and fostering interdisciplinary collaboration to translate technical advances into clinical practice, ultimately advancing precision medicine and improving diagnostic outcomes worldwide.

Acknowledgements

Not applicable.

Funding

Not applicable.

Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate

Not applicable.

Conflict of interest

The authors have no conflict of interest to declare.

Consent for publication

Not applicable.

References

[1] Maclin, Philip S, Jack Dempsey. Using an artificial neural network to diagnose hepatic masses. J Med Syst 1992;16.5:215-225.

[2] Garra BS, Krasner BH, Horii SC, Ascher S, Mun SK, Zeman RK. Improving the distinction between benign and malignant breast lesions: the value of sonographic texture analysis. Ultrason Imaging 1993;15:267-285.

[3] Lee HW, Liu BD, Hung KC, Lei SF, Wang PC. Breast tumor classification of ultrasound images using wavelet-based channel energy and imageJ. IEEE J Sel Top Signal Process 2009;3:81-93.

[4] Oelze ML, Mamou J. Review of quantitative ultrasound: Envelope statistics and backscatter coefficient imaging and contributions to diagnostic ultrasound. IEEE T Ultrason Ferr 2016;63:336-351.

[5] Krizhevsky, Alex, Ilya Sutskever, Geoffrey EH. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25.

[6] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. Proc IEEE Conf Comput Vision Pattern Recogn 2015;39:640-651.

[7] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. Int Conf Med Image Comput Computer-Assisted Intervention. Cham: Springer international publishing 2015.

[8] Yang X, Yu L, Wu L, Wang Y, Ni D, Qin J, et al. Fine-grained recurrent neural networks for automatic prostate segmentation in ultrasound images. Proc AAAI Conf Artif Intelligence 2016;31.

[9] Chen H, Dou Q, Ni D, Cheng JZ, Heng PA. Automatic fetal ultrasound standard plane detection using knowledge transferred recurrent neural networks. Int Conf Med Image Comput Computer-Assisted Intervention. Cham: Springer International Publishing 2015.

[10] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Adv Neural Inf Process Syst. https://doi.org/10.48550/arXiv.1706.03762.

[11] Shen P, Yang Z, Sun J, Wang Y, Qiu C, Wang Y, et al. Explainable multimodal deep learning for predicting thyroid cancer lateral lymph node metastasis using ultrasound imaging. Nat Commun 2025;16:7052.

[12] Qian X, Pei J, Han C, Liang Z, Zhang G, Chen N, et al. A multimodal machine learning model for the stratification of breast cancer risk. Nat Biomed Eng 2025;9:356-370.

[13] Yao W, Bai J, Liao W, Chen Y, Liu M, Xie Y. From cnn to transformer: A review of medical image segmentation models. J Imaging Inform Med 2024;37:1529-1547.

[14] Singhal K, Azizi S, Tu T, Mahdavi, SS, Wei J, Chung, HW, et al. Large language models encode clinical knowledge. Nature 2023;620:172-180.

[15] Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med 2023;29:1930-1940.

[16] OpenAI, Achiam J, Adler S, Agarwal S, Lama A, Ilge A, et al. Gpt-4 technical report. https://doi.org/10.48550/arXiv.2303.08774.

[17] Driess D, Xia F, Sajjadi MSM, Lynch, C, Chowdhery A, Ichter B, et

al. Palm-e: an embodied multimodal language model. https://doi.org/10.48550/arXiv.2303.03378.
[18] Liu H, Li C, Wu Q, Lee YL. Visual instruction tuning. Adv Neural Inf Process Syst 2023;36:34892-34916.
[19] Li XS, Huang DY, Zhang YM, Navab N, Jiang ZL. Semantic Scene Graph for Ultrasound Image Explanation and Scanning Guidance. https://doi.org/10.48550/arXiv.2506.19683.
[20] Chen J, Gui C, Ouyang R, Gao A, Chen S, Chen GH, et al. Towards injecting medical visual knowledge into multimodal llms at scale. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. 2024.
[21] Huang X, Shen L, Liu J, Shang F, Li H, Huang H, et al. Towards a multimodal large language model with pixel-level insight for biomedicine. Proceedings of the AAAI Conference on Artificial Intelligence. 2024.
[22] Chang RF, Wu WJ, Moon WK, Chen, DR, et al. Improvement in breast tumor discrimination by support vector machines and speckle-emphasis texture analysis. Ultrasound Med Biol 2003;29:679-686.
[23] Wu WJ, Woo KM. Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features. Acad Radio 2008;l15:873-880.
[24] Huang YL, Chen DR, Jiang YR, Kuo SJ, Wu HK, Moon WK. Computer - aided diagnosis using morphological features for classifying breast lesions on ultrasound. Ultrasound Obstet Gynecol 2008;32:565-572.
[25] Gomez W, Pereira WCA, Infantosi AFC. Analysis of co-occurrence texture statistics as a function of gray-level quantization for classifying breast ultrasound. IEEE Trans Med Imaging 2012;31:1889-1899.
[26] Cai L, Wang X, Wang Y, Guo Y, Yu J, Wang Y. Robust phase-based texture descriptor for classification of breast ultrasound images. Biomed Eng Online 2015;14:26.
[27] Uniyal N, Eskandari H, Abolmaesumi P, Sojoudi S, Gordon P, Warren L, et al. Ultrasound RF time series for classification of breast lesions. IEEE Trans Med Imaging 2014;34:652-661.
[28] Xia F, Wei W, Wang J, Wang Y, Wang K, Zhang C, Zhu Q. Ultrasound radiomics-based logistic regression model for fibrotic NASH. BMC Gastroenterol 2025;25:66.
[29] Jordan MI. Serial order: A parallel distributed processing approach. Adv Psychol 1997;121:471-495.
[30] Elman JL. Finding structure in time. Cogn Sci 1990;14:179-211.
[31] Hochreiter S, Jürgen S. Long short-term memory. Neural computation 1997;9:1735-1780.
[32] Chung JY, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. https://doi.org/10.48550/arXiv.1412.3555.
[33] Cho K, Merrienboer BV, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. https://doi.org/10.48550/arXiv.1406.1078.
[34] Lecun Y, Bottou L. Gradient-based learning applied to document recognition. Proceedings of the IEEE 1998;86:2278-2324.
[35] Qian X, Pei J, Zheng H, Xie X, Yan L, Zhang H, et al. Prospective assessment of breast cancer risk from multimodal multiview ultrasound images via clinically applicable deep learning. Nat Biomed Eng 2021;5:522-532.
[36] Wu WJ, Woo KM. Ultrasound breast tumor image computer-aided diagnosis with texture and morphological features. Acad Radiol 2008;15:873-880.
[37] Sharma H, Droste R, Chatelain P, Drukker L, Papageorghiou AT, Noble JA. Spatio-temporal partitioning and description of full-length routine fetal anomaly ultrasound scans. Proc IEEE Int Symp Biomed Imaging 2019;16:987-990.
[38] Yang X, Yu L, Li S, Wen H, Luo D, Bian C, et al. Towards automated semantic segmentation in prenatal volumetric ultrasound. IEEE Trans Med Imaging 2019;38:180-193.
[39] Ferreira DL, Lau C, Salaymang Z, Arnaout R. Self-supervised learning for label-free segmentation in cardiac ultrasound. Nat Commun 2025;16:4070.
[40] Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q, et al. TransUNet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal 2024;97:103280.
[41] Zhou Y, Chen C, Yao J, Yu J, Feng B, Sui L, et al. A deep learning based ultrasound diagnostic tool driven by 3D visualization of thyroid nodules. NPJ Digit Med 2025;8:126.
[42] Takahashi S, Sakaguchi Y, Kouno N, Takasawa K, Ishizu K, Akagi Y, et al. Comparison of vision transformers and convolutional neural networks in medical image analysis: A systematic review. J Med Syst 2024;48:84.
[43] Kirillov A, Mintun E, Ravi N, Mao H, Rolland C, Gustafson L, et al. Segment anything. Proceedings of the IEEE/CVF international conference on computer vision. 2023.
[44] Ma J, He Y, Li F, Han L, You C, Wang B. Segment anything in medical images. Nat Commun 2024;15:654.
[45] Chen C, Miao J, Wu D, Zhong A, Yan Z, Kim S, et al. MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Med Image Anal 2024;98:103310.
[46] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods 2021;18:203-211.
[47] Deng X, Wu H, Zeng R, Qin J, et al. Memsam: Taming segment anything model for echocardiography video segmentation. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2024.
[48] Hu JT, Zhuo W, Cheng J, Liu YY, Xue WF, Ni D. EchoONE: segmenting multiple echocardiography planes in one model. Proceedings of the computer vision and pattern recognition conference. 2025.
[49] Wu Y, Zhao T, Hu S, Wu Q, Huang X, Chen Y, et al. SAID-Net: enhancing segment anything model with implicit decoding for echocardiography sequences segmentation. Med Biol Eng Comput 2025.
[50] Koleilat T, Asgariandehkordi H, Rivaz H, Xiao YM. Medclip-sam: Bridging text and image towards universal medical image segmentation. International conference on medical image computing and computer-assisted intervention. Cham: Springer Nature Switzerland 2024.
[51] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. International conference on machine learning. PmLR, 2021.
[52] Zhang S, Xu YB, Usuyama N, Xu HW, Bagga J, Tinn R, et al. Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. https://doi.org/10.48550/arXiv.2303.00915.
[53] Chen YX, Wei MH, Zheng ZX, Hu JL, Shi YL, Xiong SW, et al. Causalclipseg: Unlocking clip's potential in referring medical image segmentation with causal intervention. https://doi.org/10.48550/arXiv.2503.15949.
[54] Jiao J, Zhou J, Li X, Xia M, Huang Y, Huang L, et al. Usfm: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. Med Image Anal 2024;96:103202.

[55] Zhang S, Zhang Q, Zhang S, Liu X, Yue J, Lu M, et al. A generalist foundation model and database for open-world medical image segmentation. Nat Biomed Eng 2025:1-16.

[56] Xu Z, Tang F, Quan Q, Yao Q, Kong Q, Ding J, et al. Fair ultrasound diagnosis via adversarial protected attribute aware perturbations on latent embeddings. NPJ Digit Med 2025;8:291.

[57] Fang X, Lin Y, Zhang D, Cheng KT, Chen H. Aligning medical images with general knowledge from large language models. https://doi.org/10.48550/arXiv.2409.00341.

[58] Ferber D, Wölflein G, Wiest IC, Ligero M, Sainath S, Ghaffari Laleh N, et al. In-context learning enables multimodal large language models to classify cancer pathology images. Nat Commun 2024;15:10104.

[59] Christiansen F, Konuk E, Ganeshan AR, Welch R, Palés Huix J, Czekierdowski A, et al. International multicenter validation of AI-driven ultrasound detection of ovarian cancer. Nat Med 2025;31:189-196.

[60] Miao HY, Jia J, Cao YK, Zhou YJ, Jiang YW, Liu Z, et al. Ultrasound-qbench: Can llms aid in quality assessment of ultrasound imaging?. https://doi.org/10.48550/arXiv.2501.02751.

[61] Van Veen D, Van Uden C, Blankemeier L, Delbrouck JB, Aali A, Bluethgen C, et al. Adapted large language models can outperform medical experts in clinical text summarization. Nat Med 2024;30:1134-1142.

[62] Chen YX, Yang HY, Pan HK, Siddiqui F, Verdone A, Zhang QY, et al. Burextract-llama: An llm for clinical concept extraction in breast ultrasound reports. Proceedings of the 1st International Workshop on Multimedia Computing for Health and Medicine. 2024.

[63] Chen MC, Fan SQ, Cao GL, Liu YH, Liu HB. USPilot: An embodied robotic assistant ultrasound system with large language model enhanced graph planner. https://doi.org/10.48550/arXiv.2502.12498.

[64] Liu C, Wei M, Qin Y, Zhang M, Jiang H, Xu J, et al. Harnessing large language models for structured reporting in breast ultrasound: a comparative study of Open AI (GPT-4. 0) and Microsoft Bing (GPT-4). Ultrasound Med Biol 2024;50:1697-1703.

[65] Wang S, Zhao Z, Ouyang X, Liu T, Wang Q, Shen D, et al. Interactive computer-aided diagnosis on medical image using large language models. Commun Eng 2024;3:133.

[66] Yao J, Wang Y, Lei Z, Wang K, Feng N, Dong F, et al. Multimodal GPT model for assisting thyroid nodule diagnosis and management. NPJ Digit Med 2025;8:245.

[67] Li ZM, Li MD, Wang W, Huang QH. Ultrasound report generation with fuzzy knowledge and multi-modal large language model. Expert Syst Appl 2025;292:128555.

[68] Lu MY, Chen B, Williamson DFK, Chen RJ, Zhao M, Chow AK, et al. A multimodal generative AI copilot for human pathology. Nature 2024;634:466-473.

[69] Guo XC, Chai WH, Li SY, Wang G. LLaVA-ultra: Large chinese language and vision assistant for ultrasound. Proceedings of the 32nd ACM international conference on multimedia. 2024.

[70] Zhu, KY, Qin ZY, Yi HH, Jiang ZK, Lao QC, Zhang, ST et al. Guiding medical vision-language models with explicit visual prompts: framework design and comprehensive exploration of prompt variations. https://doi.org/10.48550/arXiv.2501.02385.

[71] Chen YY, Xu DX, Huang Y, Zhan, SK, Wang HP, Chen DX, et al. MIMO: A medical vision language model with visual referring multimodal input and pixel grounding multimodal output. Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.

[72] Rasheed H, Maaz M, Mullappilly SS, Shaker A, Khan S, Cholakkal H, et al. Glamm: Pixel grounding large multimodal model. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.

[73] Le AJ, Liu H, Wang Y, Liu ZY, Zhu RK, Wenig TH, et al. U2-BENCH: Benchmarking large vision-language models on ultrasound understanding. 2025.arXiv preprint arXiv:2505.17779.

[74] Zhang K, Zhou R, Adhikarla E, Yan Z, Liu Y, Yu J, et al. A generalist vision–language foundation model for diverse biomedical tasks. Nat Med 2024;30:3129-3141.

[75] Peng C, Zhang K, Lyu MX, Liu HF, Sun LC, Wu YH. Scaling up biomedical vision-language models: fine-tuning, instruction tuning, and multi-modal learning. https://doi.org/10.48550/arXiv.2505.17436.

[76] Lewis M, Liu YH, Goyal N, Ghazvininejad M, Mohamed A, Levy O, et al. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. https://doi.org/10.48550/arXiv.1910.13461.