BioMedInformatics

MDPI

Systematic Review

Breast Tumor Segmentation in Automated Breast Ultrasound: A Scoping Review of Deep Learning Approaches (2020–2026)

Khadija Elkachachi *, My Hachem El Yousfi Alaoui and Atman Jbari

Laboratory of Electronic Systems, Sensors and Nanobiotechnology, National Graduate School of Arts and Crafts of Rabat (ENSAM-R), Mohammed V University in Rabat, Rabat 10100, Morocco; h.elyousfi@um5r.ac.ma (M.H.E.Y.A.); a.jbari@um5r.ac.ma (A.J.)
* Correspondence: khadija_elkachachi@um5.ac.ma

Abstract

Automated breast ultrasound (ABUS) acquires whole-breast volumes that can be reviewed independently of acquisition, but tumor segmentation in these volumes remains difficult. This scoping review maps deep learning methods for tumor segmentation in ABUS and examines how they were evaluated. Following PRISMA-ScR, PubMed, IEEE Xplore and ScienceDirect were searched for studies published between January 2020 and October 2026, without an open access restriction. Of 65 records identified, 32 sources were included: 27 primary segmentation studies, four segmentation-integrated computer-aided diagnosis studies, and the report of the TDSC-ABUS 2023 challenge, which describes 14 team methods evaluated on a common hidden test set. The primary studies were grouped into six families: CNN-based, attention-based, hybrid CNN–Transformer, lightweight/architecture-search/distillation, semi- and weakly supervised, and foundation-model-based approaches. Twelve studies used the same single-center dataset, one used more than one scanner vendor, and four tested on an external center. Reported Dice values depended strongly on the evaluation setting: values of 0.84 or higher were obtained on 2D slices, on localized regions (including 3D lesion-containing crops) or on small cohorts, whereas whole-volume evaluations on public data gave 0.58–0.65 and challenge entries 0.31–0.61. Self-configuring CNN pipelines gave the best challenge results, and most foundation-model approaches required adaptation and prompts. No family has been shown to outperform the others. Progress depends on shared multi-center benchmarks, standardized reporting of evaluation settings, data-efficient and privacy-preserving training, and evaluation of runtime and clinical workflow.

Keywords: automated breast ultrasound (ABUS); breast tumor segmentation; deep learning; 3D medical image analysis; scoping review; TDSC-ABUS; foundation model; computer-aided diagnosis

Check for updates

Academic Editor: James C. L. Chow

Received: 18 August 2026
Revised: 15 October 2026
Accepted: 15 October 2026
Published: 21 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

1.1. Breast Cancer and the Importance of Early Detection

Breast cancer remains a major public health issue and is the most common cancer among women worldwide. In 2022 alone, an estimated 2.3 million new cases and 670,000 deaths were recorded [1,2]. Breast cancer is now the second most commonly diagnosed cancer overall, and its burden is unevenly distributed: although most new cases occur in countries with a high Human Development Index (HDI), a disproportionate share of deaths occurs in low- and middle-HDI countries, where many women are diagnosed

late and lack access to appropriate treatment [1,2]. If current trends continue, 3.2 million new cases and 1.1 million deaths are projected annually by 2050 unless prevention, early detection and timely care improve [2].

Early detection is critical because the prognosis worsens sharply once the disease extends beyond the breast and regional lymph nodes. The World Health Organization reports that breast cancer detected while lesions are small and localized can be treated effectively, so that early diagnosis followed by prompt treatment reduces the risk of death [3]. Early detection includes both early diagnosis in women with signs or symptoms and population screening of asymptomatic women, most often by mammography from the age of 40 [3]. The evidence review for the US Preventive Services Task Force confirmed that mammography screening reduces breast cancer mortality [4], and collaborative modeling studies estimated that biennial screening of average-risk women could reduce breast cancer deaths by 25.4–41.7% compared with no screening [5].

Beyond survival, early detection reduces the need for extensive surgery, lowers the risk of metastatic presentation and allows curative multimodal treatment. The WHO Global Breast Cancer Initiative recommends that at least 60% of invasive breast cancers be diagnosed at stage I or II, that diagnosis be made within 60 days of the first visit, and that treatment start promptly [6]. These goals are particularly important in countries where women present at later stages and where survival differs because of unequal access to imaging, pathology and treatment [2,6].

Recent epidemiological data also raise concerns about the sustainability of progress. In the United States, breast cancer incidence increased by about 1% per year from 2012 to 2021, mainly for localized and hormone receptor-positive disease [7]. Although mortality has declined in many very-high-HDI countries, only a few meet the WHO target of a 2.5% annual reduction, and mortality is rising in some countries [2]. Early detection should therefore be seen not only as a diagnostic step but as a key factor for survival, treatment burden and equitable access to care.

1.2. Role of Ultrasound and 3D ABUS in Breast Imaging

Ultrasound is an important adjunct to mammography, for example in women with dense breast tissue, in the evaluation of palpable abnormalities and when ionizing radiation should be avoided, as in pregnant or lactating women. Its advantages include real-time imaging, wide availability, absence of radiation and better visibility of lesions within dense fibroglandular tissue. However, handheld breast ultrasound (HHUS) is operator-dependent and poorly reproducible, which lowers specificity and leads to false positives and variable performance across operators and institutions [8].

The role of ultrasound in screening is closely linked to breast density. Women with dense breasts have a higher risk of breast cancer and lower mammographic sensitivity, because dense parenchyma can mask tumors, and a large proportion of women of screening age have heterogeneously or extremely dense breasts. Recent evidence reviews indicate that supplemental imaging may be considered in selected women with dense breasts, but its benefits and harms depend on the modality and the clinical context [4,9].

Three-dimensional automated breast ultrasound (ABUS), also called automated breast volume scanning (ABVS), was developed to address the limitations of HHUS with a standardized volumetric acquisition. Whereas HHUS images are acquired and interpreted at the same time by the same operator, ABUS scans the whole breast with a standardized protocol and produces volumes that can be reviewed in the axial, sagittal and coronal planes after acquisition. Separating acquisition from interpretation reduces operator dependence, improves reproducibility and allows retrospective review and workflow standardization, which has increased interest in ABUS for both screening and diagnosis [9,10].

ABUS has mostly been evaluated together with mammography in women with dense breasts. In a real-world study, ABUS combined with mammography was as sensitive as HHUS combined with mammography, with slightly higher specificity, higher overall accuracy and fewer biopsy recommendations [11]. One retrospective study of clinical screening with 3D ABUS reported a cancer detection rate of 2.77 per 1000 examinations, comparable to earlier prospective studies, with mostly favorable recall and biopsy rates [10]. Together, these data suggest that ABUS is not merely a technical substitute for HHUS and may serve as a scalable adjunct for supplemental screening of women with dense breasts [10,11].

The standardized three-dimensional representation provided by ABUS can also support lesion localization, the detection of architectural distortion and review by several readers. Because the volumes are structured and reproducible, ABUS is increasingly used in computer-aided diagnosis (CAD) and artificial intelligence (AI) research on lesion detection, classification, and segmentation [12]. Whole-volume assessment and the coronal view may help identify lesions that are less visible on conventional two-dimensional images, particularly in dense breasts [9,12].

ABUS does not, however, remove all limitations of breast ultrasound: false positives, interpretation pitfalls, positioning artifacts (including incomplete peripheral coverage) and limited detection of microcalcifications remain. Although supplemental ultrasound detects additional cancers in women with dense breasts, long-term outcome data are still insufficient for professional societies to recommend it for all such women [9]. ABUS is therefore best seen as a complement to targeted diagnostic ultrasound: a standardized, reproducible pathway for whole-breast volumetric assessment that is increasingly used as a supplemental imaging test and as a platform for AI research [9,13].

1.3. Challenges of Breast Tumor Segmentation in ABUS

Tumor segmentation in ABUS poses considerable technical challenges. These arise from the intrinsic degradation of ultrasound images (speckle noise, low signal-to-noise ratio, acoustic shadowing and other artifacts), from the heterogeneity of tumors across a whole volume, and from the shortage of large annotated clinical datasets. Recent reviews of ABUS analysis consistently report segmentation as one of the most difficult components of computer-aided breast ultrasound analysis [12,14].

The first challenge is image quality. Speckle noise, low signal-to-noise ratio, acoustic shadowing and other artifacts obscure lesion margins. This matters more in ABUS than in single-slice analysis, because the algorithm must identify the tumor within a large volume of complex surrounding parenchyma. Low contrast and blurred boundaries between tumor and normal tissue are repeatedly reported as causes of inaccurate delineation in ABUS and in breast ultrasound more generally [14–16].

The second challenge is the variability of lesion appearance. Breast lesions in ABUS vary widely in size, shape, echotexture, orientation and margin: some are small and well defined, whereas others are irregular, spiculated or partly infiltrative. Heterogeneous internal texture in malignant lesions and weak contrast against adjacent dense tissue make it difficult for a single model to generalize across lesion types, and these factors have been shown to degrade segmentation performance [14,17,18].

The third challenge is dimensionality. An ABUS examination produces high-resolution whole-breast volumes of up to several hundred slices. Volumetric data provide richer anatomical context than 2D images, but they greatly increase memory use, training time and model complexity. Fully 3D networks are therefore computationally demanding, harder to optimize and prone to overfitting when training data are limited [12,17].

The fourth challenge is the scarcity of annotated data. Deep learning segmentation relies on precise voxel-level annotations, and manual delineation of tumors in 3D ABUS volumes is labor-intensive. The TDSC-ABUS challenge report noted the lack of publicly available, well-annotated ABUS datasets, and the challenge data remain the main shared resource [14,17]. This shortage limits external validation, makes comparisons between methods unfair and restricts generalization across institutions and scanners.

A fifth challenge is annotation uncertainty, including inter-observer variability. Tumor borders are often unclear on ultrasound, so different annotators may contour the same tumor differently. The resulting noisy labels mislead training, particularly at boundaries, and reduce segmentation performance in breast ultrasound; noise-tolerant and weakly supervised strategies are being investigated to address this [19,20]. Model performance therefore depends on both the architecture and the quality and consistency of the reference standard.

Class imbalance and small tumors are another major issue. A tumor usually occupies a very small fraction of an ABUS volume, which biases networks toward predicting background, particularly for small, low-contrast tumors or tumors partly hidden by shadowing. Challenge-based and architecture-oriented studies identify small, low-contrast tumors, noise and ill-defined boundaries as major causes of missed detections and poor masks [14,21].

Finally, a model trained on data from one scanner may not perform as well on another, because performance depends on the scanner vendor, acquisition technique, compression, body habitus and institutional workflow. Scanner and institutional variability, heterogeneous backgrounds and limited labeled data are repeatedly described as obstacles to robust clinical use [12,22]. The goal of ABUS segmentation is therefore not only a high Dice score on one dataset but also reproducible, reliable performance that transfers to diverse clinical settings.

Together, these challenges explain why ABUS tumor segmentation remains an active research area. Attention mechanisms, hybrid CNN–Transformer architectures, semi-supervised learning, efficient 3D architectures and, more recently, foundation models have been proposed, but noisy images, ill-defined margins, small and heterogeneous tumors, limited annotations and poor cross-domain generalization are still far from solved [12,14].

1.4. Rationale, Objectives and Contributions

Most reviews of AI in breast imaging cover mammography, magnetic resonance imaging and ultrasound together and do not examine the specific constraints of 3D ABUS data in depth [23]. The review by Wang et al. [12] focuses on deep learning for ABUS but covers detection, classification and segmentation together. A review dedicated to ABUS tumor segmentation is therefore needed, and it must address a recurring problem in this literature: studies rely on private datasets with different splits, input settings and metric definitions, which makes it hard to separate genuine methodological gains from differences in study design.

This scoping review has three objectives, matching the research questions in Section 2.2: (1) to identify the deep learning methods proposed for tumor segmentation in ABUS and group them into methodological families; (2) to chart how they were evaluated and to what extent their results are comparable; and (3) to identify open technical, methodological and clinical challenges and future research directions. Its contributions are:

- a classification of 27 primary studies into six methodological families (CNN-based, attention-based, hybrid CNN–Transformer, lightweight/architecture-search/distillation, semi- and weakly supervised, and foundation-model-based approaches);

- a per-study charting of datasets, data splits, input dimensionality, level of automation and metric definitions, which shows how strongly these factors affect reported results;
- a separate analysis of the TDSC-ABUS 2023 challenge, the only benchmark in which methods were evaluated on the same hidden test set;
- an analysis of the design choices used to address ABUS-specific problems (loss functions, patch-based versus whole-volume processing, post-processing), and a discussion of ABUS segmentation in the context of general-purpose and foundation segmentation models.

2. Materials and Methods

2.1. Review Design

This study adopts a scoping review design, following the methodological framework of Arksey and O' Malley [24] and reported in accordance with the PRISMA-ScR guidelines [25]; the completed PRISMA-ScR checklist is provided in Table S1. Deep learning-based tumor segmentation in ABUS is addressed by a limited and heterogeneous body of primary studies, with datasets, metrics, and evaluation protocols that are not yet standardized across them. A systematic review or meta-analysis would have required a level of comparability this literature does not yet offer, so a scoping review was chosen instead. The review addressed the three objectives stated in Section 1.4, which correspond to the research questions in Section 2.2. The review focused on deep learning methods for tumor segmentation in ABUS volumes, whether these were processed in 3D, as stacks of adjacent slices (2.5D) or as individual slices (2D). Studies from the broader breast ultrasound segmentation literature were included only when they reported segmentation results on ABUS data; other relevant works are discussed as context in Section 4.2.

2.2. Research Questions

Three questions guided the review:

- Q1. Which deep learning methods have been proposed for tumor segmentation in ABUS, and how can they be grouped into methodological families?
- Q2. How were these methods evaluated (datasets, data splits, input dimensionality, level of automation, metrics and their aggregation), and to what extent are the reported results comparable across studies?
- Q3. Which technical, methodological and clinical challenges remain open, and which research directions do the included studies point to?

2.3. Search Strategy and Databases

Because this review sits at the intersection of medical imaging and artificial intelligence, the search was conducted across databases covering both sides of the field: PubMed and ScienceDirect for the medical and clinical literature, and IEEE Xplore for the engineering and AI side. Google Scholar was used only to locate abstracts and full texts of records already identified. The search combined four groups of terms: (i) the imaging modality, including the vendor-specific name "automated breast volume scanner" (ABVS) and the terms "3D," "whole," and "volumetric" breast ultrasound; (ii) the target (tumor, lesion, mass, nodule, cancer); (iii) the task (segmentation); and (iv) the method. The method group contains not only architecture terms but also learning paradigms (semi-supervised and weakly supervised learning) and foundation-model terms ("foundation model," "segment anything"). Boolean operators were used to combine the groups.

Each query was adapted to the syntax of the database (Table 1). Because ScienceDirect limits the number of Boolean operators per search field and does not support truncation,

its search was split into two shorter queries: one on deep learning architectures and one on foundation models and semi-supervised learning.

Table 1. Search queries by database (last search: 4 October 2026).

Database	Field Searched	Query	Records
PubMed	All fields	("automated breast ultrasound" OR "automated breast volume scanner" OR ABUS OR ABVS OR "3D breast ultrasound" OR "whole breast ultrasound" OR "volumetric breast ultrasound") AND (tumor OR tumour OR lesion OR mass OR nodule OR cancer) AND segment* AND ("deep learning" OR "neural network" OR CNN OR "U-Net" OR transformer OR attention OR "foundation model" OR "segment anything" OR "semi-supervised" OR "weakly supervised")	23
IEEE Xplore	All metadata	Same query as PubMed	22
ScienceDirect, query 1	Title, abstract, keywords	("automated breast ultrasound" OR ABUS OR "automated breast volume scanner") AND segmentation AND ("deep learning" OR "neural network" OR transformer)	12
ScienceDirect, query 2	Title, abstract, keywords	("automated breast ultrasound" OR ABUS) AND segmentation AND ("foundation model" OR "segment anything" OR "semi-supervised")	3
Total before removal of duplicates			60

, truncation operator: segment retrieves all terms beginning with "segment" (e.g., segmentation, segmenting).

The literature search was limited to publications released between January 2020 and 4 October 2026, the date of the last search, so as to include only the most recent technological advances in the field. No open access filter was applied. The search identified 60 records (23 from PubMed, 22 from IEEE Xplore and 15 from ScienceDirect). In addition, citation searching of the included studies was performed, and the TDSC-ABUS challenge report was added as a benchmark source (Section 2.5).

2.4. Eligibility Criteria

Inclusion criteria: To be included in this review, studies had to meet the following inclusion criteria: have been published between January 2020 and 4 October 2026; be written in English; be peer-reviewed journal articles or full-length conference papers; evaluate the use of automated breast ultrasound (ABUS or ABVS) for the segmentation of breast tumors or lesions, whether the volumes were processed in 3D, as stacks of adjacent slices (2.5D) or as individual slices (2D); and implement computational segmentation methodologies based on deep learning. They also had to report at least one quantitative segmentation metric on ABUS data; other methodological information (e.g., description of the model architecture, dataset used and evaluation protocol) was charted as "not reported" when missing and was not a reason for exclusion. Studies that also used handheld ultrasound images, as well as general-purpose and foundation segmentation models, were eligible under the same condition; only their ABUS results were charted. Open access was not required. As an exception to the peer-review criterion, the TDSC-ABUS challenge report, available as a preprint, was included as a benchmark source.

Exclusion criteria: Publications were excluded if they were published outside the study period or in a language other than English; if they did not use ABUS data (e.g., studies based only on handheld 2D breast ultrasound, mammography, magnetic resonance imaging [MRI] or computed tomography [CT]); if they addressed classification, detection or diagnosis without a segmentation component; if the segmentation method was not based on deep learning; or if no quantitative segmentation result was reported on ABUS data. Review articles, editorials, letters, opinion pieces, conference abstracts (including extended abstracts), posters, theses and non-peer-reviewed preprints were also excluded, except for the TDSC-ABUS challenge report described above. Duplicate records were removed before screening.

2.5. Study Selection

Records were exported from each database and duplicates were removed. Titles and abstracts were screened against the eligibility criteria, and the remaining reports were assessed for eligibility. Additional records were identified by citation searching of the included studies, and the TDSC-ABUS challenge report was added as a benchmark source.

Studies that used an ABUS segmentation network as one stage of a computer-aided diagnosis (CAD) pipeline were classified as segmentation-integrated CAD studies; they were charted separately and were not counted among the primary segmentation studies. Publications describing the method of a challenge team were linked to the challenge report and counted as a single source. The selection process and the number of sources at each stage are shown in Figure 1.

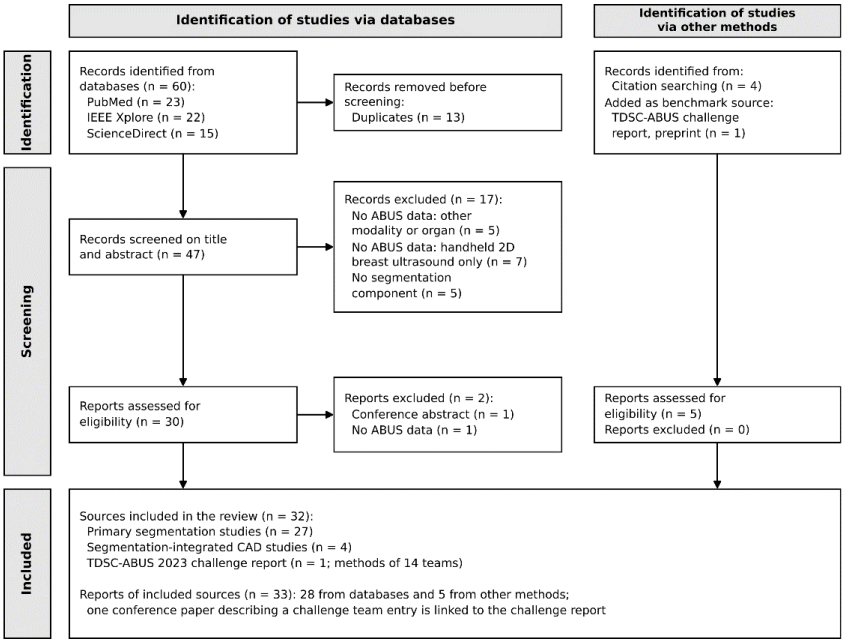


Figure 1. PRISMA-ScR flow diagram of the identification, screening and inclusion of sources. Sources are counted as studies; reports are the documents describing them.

2.6. Data Charting

A charting form was defined in a spreadsheet before extraction and applied to every included study. It recorded the following items:

- Study: bibliographic data and methodological family (Section 2.7).
- Dataset: public or private; number of centers; scanner vendor; number of patients, volumes and lesions; distribution of tumor size when reported.
- Data split: proportions for training, validation and testing; use of cross-validation; whether the split was made at patient level (patient-disjoint); use of an external test set.
- Input: 2D (individual slices), 2.5D (a stack of adjacent slices given as input channels) or 3D (whole volumes or 3D patches); patch-based or whole-volume training and inference.
- Level of automation: fully automatic on the whole volume; dependent on a region or volume of interest derived from the annotation; dependent on a detection stage; or interactive (click or bounding-box prompt).
- Training design: loss function, handling of class imbalance and of small tumors, and post-processing.

- Evaluation: metrics and their exact definition; whether the Dice coefficient was computed per 2D slice or per volume, and averaged over cases or computed on pooled voxels; Hausdorff distance variant (maximum or 95th percentile) and unit; reported values with their dispersion.
- Clinical and practical aspects: inference time and hardware, multi-vendor or external testing, comparison with radiologists, use of uncertainty or calibration, and use of collaborative or privacy-preserving training.

Items that a study did not report were coded as "not reported" rather than inferred. For the TDSC-ABUS challenge, the same items were charted for each of the 14 teams whose segmentation results were scored on the hidden test set, using the challenge report.

2.7. Synthesis

Primary studies were grouped into six methodological families according to their main stated contribution: (1) CNN-based architectures, including studies centered on loss design or multi-task learning; (2) attention-based methods; (3) hybrid CNN–Transformer architectures; (4) lightweight, neural architecture search (NAS) and knowledge distillation approaches; (5) semi-supervised and weakly supervised learning approaches; and (6) foundation-model-based approaches. When a study combined several elements, it was assigned to the family that matched its main claim; for example, a semi-supervised method built on an attention-based network was assigned to family (5). The challenge team methods were mapped to the same families for description, but their results were reported separately. Segmentation-integrated CAD studies were not assigned to a family and are discussed separately. No review protocol was registered, and, in line with scoping review methodology [25], no critical appraisal of the included studies was performed.

3. Results

3.1. Overview of the Included Studies

The database searches returned 60 records and five further records were identified through other methods (Figure 1). After de-duplication, screening and full-text assessment, 32 sources were included: 27 primary segmentation studies, four segmentation-integrated CAD studies and the report of the TDSC-ABUS 2023 challenge, which describes the methods of 14 teams (one team entry was also published as a full conference paper and is linked to the challenge rather than counted separately).

The primary studies were published between 2020 and 2026; twenty are journal articles and seven are full conference papers. By main contribution, seven belong to the CNN family, six to the attention family, seven to the hybrid CNN–Transformer family, two to the lightweight, architecture-search and distillation family, four to the semi- and weakly supervised family and one to the foundation-model family. Nineteen studies processed 3D volumes or 3D patches and eight worked on 2D slices or 2.5D stacks. Almost all datasets whose origin was reported were acquired in China; the exceptions are a European multi-center dataset [26] and a Malaysian external test set [27]. The characteristics and reported results of each primary study are given in Table 2.

Table 2. Characteristics and reported segmentation results of the primary studies, grouped by methodological family and listed chronologically within each family. Each row should be read on its own: datasets, splits, input dimensionality, level of automation and metric definitions differ between studies, so values are not comparable across rows and no ranking is implied.

Study	Input	Dataset (Center; Scanner)	Cases	Split (Patient-Disjoint?)	Level of Automation	Dice (Evaluation Set)	Boundary Metric	Notes
Family 1: CNN-based architectures (including loss design and multi-task learning)								
Lei et al. [28]	3D, patches	ZJCH †; GE	25 patients	5-fold CV; yes	Fully automatic	0.84 ± 0.03 (CV) ‡	CMD 1.95 mm	Reported IoU (0.96) exceeds Dice although both are defined on the segmented volume; earlier study of the same group as Lei et al. [29]
Cao et al. [30]	3D, ROI 80 × 128 × 128 voxels	PUPH; Siemens	83 volumes/75 patients	53/30 volumes; yes	ROI centered on the mass (in most cases)	0.82 ± 0.08 (test)	HD 11.72 pixels (not HD95)	Boundary loss with signed non-Euclidean distance map
Lei et al. [29]	3D, patches 128 × 128 × 64	ZJCH †; GE	70 patients	40 (5-fold CV) + 30 hold-out; yes	Fully automatic	0.821 ± 0.145 (hold-out); 0.850 ± 0.104 (CV)	HD95 1.67 mm	Tumor volume 0.04–24.4 cm^3
Cao et al. [31]	3D, ROI 128 × 64 × 128 voxels (after resampling)	PUPH; Siemens	170 volumes/107 patients	5-fold CV; yes	ROI centered on the mass	0.690 (CV)	HD95 4.92 mm	Uncertainty focus loss; 2.0 M parameters
Zhou et al. [32]	3D, ROI 64 × 32 × 64 mm	PUPH; Siemens	170/107	4-fold CV; yes	ROI centered on the tumor	0.778 ± 0.145 (CV)	HD95 3.30 mm	Joint segmentation and classification
Cheng et al. [33]	2D slices (resized to 256 × 256)	PUPH; Siemens	170/107 (3742 slices)	120/50 volumes (75/32 patients); yes	Lesion-containing slices only	0.845 (test)	NR	Dice computed on 2D slices; mixed spatial–channel self-attention
He et al. [34]	3D, lesion-containing crop resized to 128^3	TDSC (200) + Jiangyin Hospital of TCM (400)	600 cases	447/153, stratified; case-level	Crop contains the lesion	0.839 (test)	HD95 7.28 mm	Boundary-aware multi-task (segmentation + classification); test set composition described inconsistently
Family 2: Attention-based models								
Pan et al. [35]	Sequences of 3 adjacent 2D slices + BLSTM	PUPH; Siemens	170 volumes/ 124 patients	104/10/56 volumes (75/8/41 patients); yes	Selected slices (3636 labeled slices; selection criterion NR)	0.818 (test) ‡	HD 11.14 (unit NR) ‡	Values from the 2D evaluation
Zhou et al. [36]	3D, whole volume (sliding window)	PUPH; Siemens	170/107	126/34/10 volumes; yes	Fully automatic	0.646 ± 0.291 (test)	HD95 11.01 mm ‡	Same group and data as Zhou et al. [32], without a tumor-centered ROI
Zhang et al. [37]	3D, input 128 × 64 × 128 voxels	PUPH; Siemens	170/107	109/15/46 volumes (64/11/32 patients); NR	ROI centered on the mass (probable; cropping NR)	0.745 (test)	HD95 5.39 mm	Spatial, channel, CBAM and SE attention compared

Table 2. *Cont.*

Study	Input	Dataset (Center; Scanner)	Cases	Split (Patient-Disjoint?)	Level of Automation	Dice (Evaluation Set)	Boundary Metric	Notes
Yang et al. [38]	2.5D (three adjacent slices)	Lanzhou † + external center; Siemens ABVS	101 + 87 cases	5-fold CV at patient level (8:2) (internal); external test; yes	Tumor-containing slices, cropped with the tumor near the center	0.850 (internal) ‡; 0.808 (external)	HD95 6.22 ‡/9.82 mm	External validation on a second center; selective-kernel encoder
Malek-mohammadi et al. [26]	2D + saliency map, after Bi-ConvLSTM detection	4 European centers; U-Systems + Siemens	60 volumes/43 patients	5-fold CV; yes	Fully automatic (detection then segmentation)	0.858 (CV)	HD95 5.23 (unit NR)	Only multi-center, multi-vendor dataset; 51% of masses smaller than 0.2 cm^3
Haqkiem et al. [27]	2D, twin-stream (adjacent slice)	TDSC + UMMC (external)	110 TDSC volumes (80 training/ 30 validation) + 184 UMMC volumes	TDSC validation; UMMC external test	Fully automatic	0.381 (TDSC); 0.238 (UMMC)	NR	0.639/0.697 on cases where all compared models reached Dice $\geq$ 0.1 ‡; Dice computed per volume
Family 3: Hybrid CNN–Transformer models								
Liu et al. [39]	3D; training crops around the label, sliding-window inference	Shenzhen †; Siemens	103 volumes	5-fold CV; NR	Fully automatic at inference	0.764 (CV)	HD95 15.47 voxels	Same group as Liu et al. [40]
Zhuang et al. [41]	2D	ZJCH; GE	256 patients	200/56 patients; yes	Fully automatic	0.604 (test)	HD95 7.74 mm	Consistency regularization with unlabeled and lesion-free images
Pan et al. [42]	3D; training crops 128 × 64 × 128 around the label (inference NR)	PUPH; Siemens	170 volumes	5-fold CV; NR	NR (inference not described)	0.813 ‡	HD95 2.21 mm ‡	Angular-margin contrastive loss on boundaries
Jeong et al. [43]	3D, patches 192 × 160 × 80 (nnU-Net)	GE Invenia; TDSC split (challenge not named)	200 volumes	100/30/70; NR	Fully automatic	0.583 ± 0.069	HD95 23.59 mm	Adaptive attention + channel-attentive gate + multi-task; nnU-Net baseline 0.547 on the same data; same group as Jeong et al. [44]
Li et al. [45]	3D, sliding window (nnU-Net)	PUPH; Siemens (+ TDSC)	170/107	120/50 volumes; NR	Fully automatic	0.755 (test)	HD95 22.26 mm	Over 50% of tumors < 1 cm^3; results reported by tumor size
Liu et al. [40]	3D, patches 128^3	Shenzhen; ABVS (+ TDSC as external test)	196 volumes (225 tumors)	166/30 volumes; 5-fold CV; NR	Fully automatic, with false-positive post-processing	0.737 (CV); 0.653 (TDSC, external)	HD95 23.23 (unit NR); 25.21 (TDSC; unit NR)	Self-supervised (masked image modeling) pre-training; code available
Jeong et al. [44]	3D (nnU-Net framework)	TDSC; GE	200 volumes	5-fold CV; NR	Fully automatic	0.598 (CV)	HD95 17.85 mm	Parallel vision Mamba module; Dice averaged per patient; results by tumor volume (37.5% of tumors < 0.5 cm^3); code available

Table 2. *Cont.*

Study	Input	Dataset (Center; Scanner)	Cases	Split (Patient-Disjoint?)	Level of Automation	Dice (Evaluation Set)	Boundary Metric	Notes
Family 4: Lightweight models, neural architecture search and knowledge distillation								
Cao et al. [46]	3D, ROI 128 × 64 × 128 voxels (after resampling)	PUPH; Siemens	170/107	120/50 volumes (75/32 patients); yes	ROI centered on the mass	0.778 ± 0.103 (test)	HD95 2.21 mm; maximum HD 4.63 mm	Neural architecture search
Pan et al. [47]	3D, ROI 64 × 128 × 128 voxels (after resampling)	PUPH; Siemens	170/107	NR	ROI centered on the tumor	0.768 ± 0.14 (3D HR-Net student) ‡	HD95 2.46 mm	Decoupled contrastive knowledge distillation
Family 5: Semi-supervised and weakly supervised learning								
Cao et al. [48]	2D slices	PUPH; Siemens	170/107	109/15/46 volumes; yes	Selected slices (criterion NR)	0.743 (test)	HD 3.81 mm (not HD95)	Uncertainty-aware temporal ensembling; 13,382 slices used
Li et al. [49]	3D, ROI 128 × 64 × 128 cropped at the tumor center	PUPH; Siemens (+ TDSC, 80/20)	170/107	120/50 volumes (75/32 patients); yes	ROI centered on the tumor	0.668 (50% labels); 0.580 (20%) ‡	HD95 9.33 mm (50% labels)	TDSC: 0.621 (50% labels)
Pan et al. [50]	2D slices	PUPH; Siemens	170/107 (3736 slices)	62/8/37 patients; yes	Lesion-containing slices (probable)	0.746 (test)	NR	Dual-teacher mean-teacher variant
Huang et al. [51]	3D	In-house	605 cases ‡/1735 volumes	1124/125/486 volumes; yes	3D bounding box required at inference	0.755 ± 0.127 (test)	HD 20.08 voxels	Box-level supervision; prompted SAM-Med3D reached 0.38–0.48 on the same data
Family 6: Foundation-model-based approaches								
Zhou et al. [52]	3D	Lishui; GE (+ TDSC as external test)	292 patients analyzed (329 collected) (+ 200 TDSC)	204/29/59 (5-fold CV on training + validation); external TDSC	Semi-automatic: 128^3 crop from the lesion area and 11 simulated corrective clicks derived from the ground truth	0.75 (test); 0.79 (TDSC) ‡	HD 60.67 (maximum; unit NR)	SAM-Med3D with LoRA; malignant tumors only; lesion volumes 1.64–100.03 cm^3

Abbreviations: PUPH, Peking University People's Hospital (Siemens ACUSON S2000 ABVS); ZJCH, Zhejiang Cancer Hospital (GE Invenia ABUS); TDSC, TDSC-ABUS 2023 challenge data (Harbin Medical University Cancer Hospital, GE Invenia ABUS); UMMC, Universiti Malaya Medical Centre; CV, cross-validation; ROI, region of interest; CMD, center-of-mass distance; HD, Hausdorff distance; HD95, 95th-percentile Hausdorff distance; NR, not reported; "probable", inferred from the article but not stated explicitly; †, center inferred from the authors' affiliation; ‡, the article reports other values for this metric (abstract, text or another table), and the value of the main results table is given; "Fully automatic", no manual input described in the article. "Patient-disjoint" is marked "yes" only when the article states that the split was made at the patient level. Dice values are reported as given by each study; for studies with 2D or 2.5D input, Dice may have been computed on individual slices rather than on whole volumes and is therefore not equivalent to a volumetric Dice. Boundary metrics are given with the unit reported by each study. Results of the TDSC-ABUS 2023 challenge teams, obtained on a common hidden test set, are reported separately in Section 3.11 .

Four further studies used an ABUS segmentation network as one stage of a diagnostic pipeline (Table 3). Their main outcome is benign/malignant classification; two did not report a segmentation metric [53,54], and all four applied the segmentation network to a region or images selected manually [53–56]. They were therefore kept apart from the primary studies and are used only in the Discussion, to show how segmentation feeds clinical decision support.

Table 3. Segmentation-integrated CAD studies. These studies are not included in the comparison of segmentation methods. HMUCH, Harbin Medical University Cancer Hospital; SYSUCC, Sun Yat-sen University Cancer Center; VOI, volume of interest.

Study	Pipeline	Segmentation Step	Segmentation Evaluated?	Data
Xiang et al. [55]	3D U-Net segmentation, then 3D Res-CapsNet benign/malignant classification	On a VOI defined manually by a radiologist	Yes (Dice 0.88)	SYSUCC; 396 patients, 444 tumors
Wang et al. [53]	3D segmentation network (Dice loss) providing tumor shape to ResNet34/50/101 v2 benign/malignant classifiers	Automatic, on lesion crops sized from the manual annotation	No	HMUCH; GE Invenia; 769 lesions/743 patients; random lesion-level split 600/169
Li et al. [56]	Two-stage segmentation and multi-view classification of BI-RADS 4 lesions	Automatic, on 2D images of the largest lesion section selected manually in three views (axial, sagittal, coronal)	Yes (Dice 0.817 ± 0.142, external center)	Two centers; Siemens ABVS; 251 lesions/216 patients (178 training, 73 external test)
Zhang et al. [54]	3D U-Net++ segmentation, then 3D attention-inception classification	On a VOI cropped manually; hole filling and largest-region selection	No	SYSUCC; same cohort as Xiang et al. [55]

3.2. Recurrent Technical Challenges

The included studies describe a consistent set of difficulties: speckle noise and blurred boundaries, wide variation in tumor size with many small lesions, the memory cost of whole-breast volumes, the scarcity of voxel-level annotations, false positives caused by artifacts and anatomical structures such as the nipple, and limited generalization across centers and scanners. Table 4 lists, for each difficulty, the approaches used and examples of studies; the families in Sections 3.3–3.8 and the design choices in Section 3.10 describe them in more detail.

Table 4. Recurrent technical challenges in ABUS tumor segmentation and how the included studies addressed them.

Challenge	Approaches in the Included Studies	Examples
Speckle noise, low contrast, blurred boundaries	Boundary-oriented losses and branches; attention to salient regions	Cao et al. [30]; Pan et al. [42]; He et al. [34]; Zhou et al. [36]
Variability in tumor size; small tumors	Multi-scale fusion, dilated convolutions, modules for small tumors; results reported by tumor size	Cao et al. [31]; Cheng et al. [33]; Yang et al. [38]; Li et al. [45]; Jeong et al. [44]
Size of the volumes (memory, computation)	3D patches with sliding-window inference; 2D/2.5D inputs; architecture search and distillation	Lei et al. [29]; Liu et al. [40]; Yang et al. [38]; Cao et al. [46]; Pan et al. [47]
Limited annotated data	Semi- and weakly supervised learning; self-supervised pre-training; adaptation of foundation models	Cao et al. [48]; Li et al. [49]; Huang et al. [51]; Liu et al. [40]; Zhou et al. [52]
False positives (artifacts, nipple, shadowing)	Inter-slice context; multi-task consistency; dedicated post-processing	Pan et al. [35]; Jeong et al. [44]; Liu et al. [40]; Haqkiem et al. [27]
Generalization across centers and scanners	Multi-center training data; testing on an external center	Yang et al. [38]; Malekmohammadi et al. [26]; Liu et al. [40]; Zhou et al. [52]; Haqkiem et al. [27]

The following subsections describe the primary studies of each methodological family, followed by the TDSC-ABUS 2023 challenge entries assigned to the same family (Table 5). Characteristics and reported results of every primary study are given in Table 2. Because most families contain only a few studies, which often share a single private dataset,

the statements below describe what individual studies found and are not generalizations about the family as a whole.

3.3. CNN-Based Architectures

Seven primary studies and nine challenge entries fall into this family. Lei et al. applied a 3D Mask R-CNN to 25 patients [28] and later extended it to a Mask scoring R-CNN evaluated on 70 patients [29], with a mean Dice of 0.82 on a 30-patient hold-out set. In both studies a region proposal stage localizes the tumor before voxel-wise segmentation of 3D patches. In the 2020 report, the intersection over union (0.96) is higher than the Dice (0.84). This cannot occur when both are computed on the same volumes, yet the article defines both on the segmented tumor volume (while also describing the IoU as a detection metric), so the two values cannot be reconciled.

Two studies from the Peking University People's Hospital group addressed class imbalance through the loss function. Cao et al. [30] added a boundary loss based on a signed non-Euclidean distance map to the Dice loss, which raised the Dice from 0.77 to 0.82 on 83 volumes; the network was applied to a region of interest centered on the mass. Cao et al. [31] combined a densely connected U-Net with hybrid dilated convolutions (2.0 M parameters) and an uncertainty focus loss that gives more weight to voxels with uncertain predictions, and obtained a Dice of 0.69 in five-fold cross-validation on 170 volumes, within a region of interest centered on the mass.

Multi-task designs were used in two studies. Zhou et al. [32] trained a V-Net jointly for segmentation and benign/malignant classification, with iterative refinement of the feature maps (Dice 0.78), and He et al. [34] added a boundary-prediction branch to a 3D multi-task network trained on 600 cases, including the TDSC-ABUS data (Dice 0.84). In both studies the network received a crop that contained the lesion, so these values describe segmentation within an already localized region. Cheng et al. [33] worked on 3742 2D slices from 170 volumes, which were the slices containing the annotated lesion (slices without lesion were excluded), with a decoder guided by the deepest encoder features, a multi-scale fusion module and a mixed spatial–channel self-attention module (Dice 0.85, computed on slices).

Challenge entries: CNNs were also the most common choice in the challenge, with nine of the fourteen entries. Most relied on self-configuring pipelines (nnU-Net [57], STU-Net, MONAI Auto3DSeg) with limited changes, such as residual encoders or ensembling. The two highest Dice values on the hidden test set belong to this group (T2, 0.615; T8, 0.602), which suggests that a well-configured standard CNN remained competitive with the more specialized designs submitted.

3.4. Attention-Based Models

Six primary studies and one challenge entry use attention as their main contribution. Pan et al. [35] inserted spatial–channel attention into a 2D fully convolutional network and linked consecutive slices with a bidirectional LSTM to remove false positives (Dice 0.82 on selected slices); the authors note that the nipple region was sometimes segmented together with the tumor. Zhou et al. [36] passed the probability map of a 3D Mask R-CNN head to every level of a V-Net through attention blocks. Evaluated on whole volumes with a sliding window, it reached a Dice of 0.65, lower than the 0.78 that the same group obtained on the same data when the input was a tumor-centered region [32]. The comparison shows how much the input setting, rather than the architecture alone, affects the reported value. Zhang et al. [37] compared spatial, channel, CBAM and squeeze-and-excitation attention within a dense network; spatial attention gave the highest Dice (0.75).

Three further studies used 2D or 2.5D inputs. Yang et al. [38] stacked each slice with its two neighbors as a three-channel image and added an improved channel attention module to the skip connections and a selective-kernel encoder that mixes dilated and standard convolutions for tumors of different sizes. The Dice fell from 0.85 on the internal center to 0.81 on an external center; only slices that contained a tumor were segmented. Malekmohammadi et al. [26] first detected masses with a Bi-ConvLSTM network and then segmented each slice with an attentive U-Net guided by a saliency map (Dice 0.86), on 60 volumes from four European centers acquired with two scanner vendors, the only multi-vendor dataset in the corpus. Haqkiem et al. [27] fused adjacent slices in a twin-stream U^2-Net with efficient channel attention. Trained on the TDSC-ABUS data, it reached a Dice of 0.38 on the TDSC validation set and 0.24 on an external Malaysian dataset; on the subset of cases in which all compared models reached a Dice of at least 0.1, the values were 0.64 and 0.70. The values of 0.66 and 0.65 given in the abstract for cases with a Dice of at least 0.1 correspond, in the result tables, to the subgroup of cases with a shadow ratio of at least 1.0.

Challenge entries: One team (T18) combined 3D attention (DAF3D) and multi-dimensional attention networks with U-Net and nnU-Net in 2.5D and 3D ensembles, with a Dice of 0.489.

3.5. Hybrid CNN–Transformer Models

Seven primary studies and two challenge entries combine convolutional layers with self-attention. Liu et al. [39] added Transformer layers and atrous spatial pyramid pooling to a 3D attention U-Net (Dice 0.76 on 103 ABVS volumes). The same group later proposed DST-C [40], in which a CNN branch and a Swin-type branch with dilated sampling self-attention exchange information through spatial–channel attention; the model was pre-trained by masked image modeling and followed by a post-processing step that removes false positives (Dice 0.74). Trained on the in-house data only, it reached 0.65 on the 100 public TDSC-ABUS training volumes used as an external test set. Zhuang et al. [41] built a residual Swin Transformer U-Net with an image-reconstruction decoder that also learns from unlabeled and lesion-free images (Dice 0.60 on 2D images from 256 patients).

Pan et al. [42] placed a CNN decoder and a decoder with Gaussian-weighted self-attention after a shallow CNN encoder, and trained them with an angular-margin contrastive loss focused on boundaries (Dice 0.81). Li et al. [45] used a dual-branch encoder–decoder in which a Transformer branch and a convolutional branch interact through cross-attention, with a graph-convolution module for small tumors (Dice 0.75 on a test set where more than half of the tumors were smaller than 1 cm^3). Jeong et al. [43] designed a 3D hybrid Transformer with adaptive attention blocks, a channel-attentive gate and a classification branch for multi-task learning, which raised the Dice of an nnU-Net baseline from 0.55 to 0.58 and reduced the HD95 from 37 to 24 mm on the same data. A later journal article by the same group [44] built the model on the nnU-Net framework, added a parallel vision Mamba module and a consistency loss between the segmentation and classification outputs, and reached a Dice of 0.60 in five-fold cross-validation on the 200 TDSC-ABUS volumes.

The three hybrid models evaluated on TDSC-ABUS data reached Dice values of 0.58 to 0.65, in the range of the best challenge entries although not on the same test cases, whereas values between 0.60 and 0.81 were reported on private datasets. With the present evidence it is not possible to say whether hybrid models segment ABUS tumors better than CNNs, because no study compared the two on a common external test set.

Challenge entries: Two teams used hybrid models: an ensemble of Swin-UNETR models (T6, Dice 0.540) and a vision Transformer encoder with a 3D CNN decoder (T12, 0.467). Both scored below the best CNN entries on the same hidden test set.

3.6. Lightweight, Neural Architecture Search and Knowledge Distillation Approaches

Two primary studies target model efficiency. Cao et al. [46] used neural architecture search to select the cell structure and multi-scale fusion of a densely connected encoder–decoder, with search and training decoupled to limit memory use (Dice 0.78 on 50 test volumes, within a region of interest centered on the mass). Pan et al. [47] distilled a 3D U-Net teacher into smaller students through contrastive learning applied separately to tumor and background features, together with a ranking loss for hard pixels; a 3D HR-Net student reached a Dice of 0.77 (0.78 in the abstract), and a small 3D U-Net student gained 7 percentage points, also within a region centered on the tumor. Both studies report efficiency as parameter counts; neither reports inference time on clinical hardware.

Challenge entries: No challenge entry was designed primarily for efficiency, and the challenge did not evaluate runtime or memory use.

3.7. Semi-Supervised and Weakly Supervised Learning

Four primary studies reduce the need for voxel-level annotations. Cao et al. [48] used temporal ensembling in which a pixel-wise uncertainty map weights both the pseudo-labels and the unsupervised loss, so that unreliable predictions contribute less (Dice 0.74 on 2D slices). Pan et al. [50] extended the mean-teacher framework with an auxiliary teacher and kept only high-confidence pseudo-labels (Dice 0.75 on slices). Li et al. [49] used an adversarial framework with two discriminators that enforce consistency between labeled and unlabeled predictions and between the segmentation and a distance-map representation; with 50% of the labels, it reached a Dice of 0.67 within a region centered on the tumor. Huang et al. [51] trained with 3D bounding boxes only, using reinforcement-learning agents that progressively erase the nodule inside the box (Dice 0.75 on 486 test volumes); the box is still needed at inference.

Three of the four studies used data from the same center, and only one [49] also reported results on a second, public dataset.

Challenge entries: No team reported using unlabeled data, although only 100 annotated training volumes were available.

3.8. Foundation-Model-Based Approaches

One primary study and two challenge entries adapt segment-anything models. Zhou et al. [52] fine-tuned SAM-Med3D [58] with low-rank adaptation and simulated clicks on 292 patients, with a Dice of 0.75 on the internal test set and 0.79 on the TDSC-ABUS data used as external test. Without ABUS-specific adaptation, the same model performed poorly: in the experiments of Huang et al. [51], prompted SAM-Med3D reached a Dice of only 0.38 to 0.48 with one to ten clicks. Only one primary study met the inclusion criteria in this family; promptable, few-shot and universal segmentation models evaluated outside ABUS are discussed in Section 4.2.

Challenge entries: T7 fine-tuned MedSAM [59] inside boxes detected by YOLOv8 (Dice 0.498), and T9 fine-tuned SAMed on 2D slices (Dice 0.585), which gave the lowest Hausdorff distance of all entries (80.2). Taken together, these results suggest that foundation models need adaptation to ABUS data to be useful, and that they currently depend on clicks or a detection stage.

3.9. Datasets and Evaluation Protocols

The 27 primary studies rely on a small number of datasets. Twelve used the same 170-volume private dataset from Peking University People's Hospital (107 patients, Siemens ACUSON S2000); two earlier studies from the same group used 83 volumes and a 124-patient version of this cohort. Three studies used data from Zhejiang Cancer Hos-

pital (GE Invenia; two of them as inferred from the authors' affiliation), and at least eight used the TDSC-ABUS 2023 data as their main, additional or external dataset. Only one study used images from more than one scanner vendor [26], and four evaluated a trained model on data from a center not used for training [27,38,40,52].

Evaluation protocols differed even within the shared dataset. Studies on the 170-volume cohort used five-fold or four-fold cross-validation, or fixed splits of 104/10/56, 120/50, 126/34/10 or 109/15/46 volumes. The same split sizes recur within the group (109/15/46 volumes: Cao et al. [48], and Zhang et al. [37]; 120/50 volumes: Cheng et al. [33], Cao et al. [46], Li et al. [45], and Li et al. [49]), but whether the same volumes were used is not stated. Seventeen of the 27 studies stated that the split was made at patient or case level; the other ten did not report it.

Three further factors affect how the reported Dice values should be read. First, eight studies processed 2D slices or stacks of adjacent slices, and in several of them the Dice was computed per slice rather than per volume. Second, eleven studies gave the network information about the tumor location: a region of interest centered on the mass, a crop containing the lesion, only the slices that contained a tumor, a bounding box or user clicks; four more probably did: two used only a few thousand slices from 170 volumes (Pan et al. [35]; Pan et al. [50]), one used 13,382 slices from 170 volumes without stating how they were selected (Cao et al. [48]), and one used the same 128 × 64 × 128 input as the tumor-centered studies of the same group (Zhang et al. [37]). Third, boundary metrics were reported in different forms, as the 95th-percentile or maximum Hausdorff distance, in millimeters, pixels or voxels, or not at all. Only a few studies stated how the Dice was aggregated (for example, per patient in Jeong et al. [44] and per volume in Haqkiem et al. [27]), and three reported results by tumor size [44,45,52]. These factors are recorded for each study in Table 2.

Their effect can be seen within studies from the same group on the same data. Zhou et al. obtained a Dice of 0.78 when the network received a tumor-centered region [32] and 0.65 when it segmented whole volumes [36]. In Haqkiem et al. [27], the Dice on the full test sets (0.38 and 0.24) rose to 0.64 and 0.70 on the subset of cases in which all compared models reached a Dice of at least 0.1. Values of 0.84 or higher in Table 2 were all obtained on 2D slices, on a localized region (including 3D lesion-containing crops), or on 70 patients or fewer, whereas whole-volume evaluations on the TDSC-ABUS data [40,43,44] and on the challenge test set (Section 3.11) gave values between 0.58 and 0.65.

3.10. Design Choices Addressing ABUS-Specific Challenges

This subsection groups the design choices that the included studies and challenge entries used to deal with class imbalance, volume size and false positives.

3.10.1. Loss Functions

Most studies combined an overlap term with a voxel-wise term, typically Dice with cross-entropy or binary cross-entropy (e.g., [29,39–41,45,52]), and this was also the most frequent choice in the challenge (T5, T10, T12, T13). Some studies used Dice loss alone [26,37,46]. Others targeted boundaries: a boundary loss based on a signed distance map [30], an angular-margin contrastive loss [42], a boundary-prediction branch [34], or a surface loss combined with Dice and focal terms (challenge team T6). Two studies weighted the loss by predictive uncertainty [31,48], and four added classification or consistency terms through multi-task learning [32,34,43,44]. No included study used Tversky or focal-Tversky losses, which are designed to trade false negatives against false positives and could suit small tumors.

3.10.2. Patch-Based Versus Whole-Volume Processing

Because an ABUS volume can contain several hundred slices of more than 800 × 500 pixels, no study fed full-resolution volumes to the network in one pass. Three strategies were

used. Most of the 19 studies with 3D input trained on 3D patches or crops and, when reported, reconstructed the whole volume at inference with a sliding window (e.g., Lei et al. [29], 128 × 128 × 64 patches; Li et al. [45], 208 × 80 × 208 patches selected by nnU-Net; Liu et al. [40], 128^3; challenge teams T2, T3, T8 and T12). Two challenge teams down-sampled the whole volume, either alone (T5) or as a coarse first stage followed by refinement of each detected component (T13). Eight primary studies and five challenge entries avoided 3D memory costs by working on 2D slices or 2.5D stacks, sometimes adding inter-slice context through an LSTM [35], adjacent-slice channels [38] or twin streams [27]. Patch- and slice-based processing keep the native resolution but lose wider context, which can lead to false positives on structures that resemble tumors, such as the nipple region in 2D slices [35].

3.10.3. Post-Processing

Post-processing was rarely described in the primary studies: only Liu et al. [40] reported a dedicated step to reduce false positives; Pan et al. [35] removed them within the network through inter-slice context. Challenge reports were more detailed: teams kept the largest connected component (T9, T12, T16), refined masks with a conditional random field (T6), removed candidates with low classification confidence (T6), averaged ensembles of cross-validation models (T3, T6) or used flip test-time augmentation (T16). Because these steps can change Dice and Hausdorff values substantially, their absence from most method descriptions is a further obstacle to reproducing published results.

3.11. Benchmark Evidence from the TDSC-ABUS 2023 Challenge

The TDSC-ABUS 2023 challenge, organized at MICCAI 2023, is currently the only shared benchmark for 3D ABUS tumor segmentation [14]. It used 200 volumes acquired at a single center (Harbin Medical University Cancer Hospital), split by stratified sampling into training (n = 100), validation (n = 30) and test (n = 70) sets, with a malignant-to-benign ratio of about 0.58:0.42; the data are distributed through the challenge platform (tdsc-abus2023.grand-challenge.org). Test labels were not released during the challenge: each team submitted a Docker container that the organizers ran in a standardized environment, and Dice and Hausdorff distance (HD) were computed by the organizers for every test case. Because all teams were evaluated on the same hidden cases with the same metric implementation, the values in Table 5 can be compared with each other, unlike those reported by the primary studies (Table 2).

Table 5 summarizes the 14 teams whose segmentation results were scored on the hidden test set; four of them (T1, T4, T8 and T10) were ranked only after the organizers replaced infinite Hausdorff distances with a penalty value. Three observations are relevant to this review. First, mean Dice on the hidden test set ranged from 0.31 to 0.61, below most values reported on private datasets (Table 2), which suggests that the latter should not be read as estimates of performance on unseen patients. Second, overlap and boundary accuracy were weakly related: the team with the highest Dice (T2, 0.615) did not obtain the lowest HD, which was achieved by a Segment Anything-based model (T9, HD = 80.2). Reporting Dice alone therefore hides part of the picture. Third, most top-ranked entries relied on self-configuring CNN pipelines (nnU-Net, STU-Net, MONAI Auto3DSeg) rather than on new architectures, and neither foundation-model entry was superior in overlap.

The organizers noted three limitations that bear directly on clinical use: residual label noise in the training annotations, data from a single center and a limited set of acquisition protocols, and the absence of any runtime or hardware evaluation.

Table 5. Segmentation methods and results of the 14 scored teams in the TDSC-ABUS 2023 challenge (hidden test set, $n = 70$ volumes). Only the 14 teams whose segmentation results were scored are listed (T1, T4, T8 and T10 only after the penalty adjustment for infinite Hausdorff distances), so the team identifiers (T1–T18) are not consecutive. Teams are listed by identifier, not by rank. Because all teams were evaluated by the organizers on the same hidden cases with the same metric implementation, these values can be compared with each other, unlike those in Table 2.

Team	Architecture	Input	Training/Inference Strategy	Loss	Post-Processing	Dice	HD
T1	SEGNet (U-Net with multi-scale features)	3D	NR	NR	NR	0.586	117.2
T2	Multi-task 3D nnU-Net with residual encoder	3D	Lesion-positive and random negative patches; sliding window	NR	NR	0.615	90.5
T3	nnU-Net with residual encoder	3D	Patches 80 × 160 × 192; 5-fold CV	NR	Ensemble of 5 folds	0.562	162.9
T4	STU-Net (large)	3D	NR	NR	NR	0.589	159.0
T5	nnU-Net 3D low-resolution	3D (down-sampled)	5-fold CV	Dice + BCE	NR	0.538	96.5
T6	Swin-UNETR ensemble + SegResNetVAE for small tumors	3D	NR	Dice + focal + surface	CRF; confidence-based false-positive removal	0.540	121.2
T7	YOLOv8 detection, then fine-tuned MedSAM in detected boxes	2D slices	Box-prompted	NR	Slice stacking into a 3D mask	0.498	153.1
T8	SegResNet (MONAI Auto3DSeg)	3D	Sliding window	Dice	Resampling to native resolution	0.602	82.9
T9	SAMed (fine-tuned SAM)	2D slices (512 × 512)	5-fold CV	NR	Probability-map fusion; largest 3D connected component	0.585	80.2
T10	2D U-Net with classification branch	2.5D (contiguous slices)	Multi-scale 3D crops	Dice + CE	NR	0.534	105.1
T12	ViT encoder with cross-attention + 3D CNN decoder	3D	Patches 128^3; sliding window	Dice + CE	Largest connected component	0.467	266.1
T13	ResUNet with ASPP, coarse-to-fine	3D	Whole volume at 256^3, then per-component ROI	Dice + 0.5 × BCE	Per-component refinement	0.306	203.4
T16	2D U-Net (full report: Barkhof et al. [60])	2D slices	NR	NR	Flip TTA; largest 3D connected component	0.441	101.4
T18	MDA-Net, U-Net, nnU-Net, DAF3D	2.5D (3 slices) and 3D	NR	NR	NR	0.489	81.7

Dice, mean per-case Dice similarity coefficient; HD, Hausdorff distance computed by the organizers, with infinite values replaced by 105% of the worst valid HD for the case (penalty-adjusted leaderboard; units not specified in the challenge report). BCE, binary cross-entropy; CE, cross-entropy; CRF, conditional random field; CV, cross-validation; TTA, test-time augmentation; NR, not reported in the challenge report. Values from the penalty-adjusted results table (Table 4) of the challenge report [14].

3.12. Comparative Summary

Taken together, the included studies show a field that has moved from CNN baselines to attention, hybrid, data-efficient and foundation-model approaches, but they do not show that any family outperforms the others. Within single studies, specific design choices improved results on the authors' own data, such as boundary or uncertainty-aware losses, inter-slice context and pre-training. Across studies, however, differences in data, input setting and metric definition are larger than the differences between methods: on the TDSC-ABUS data, the best fully automatic CNN, hybrid and foundation-model approaches reached Dice values between 0.58 and 0.65 (on different test cases), whereas a click-prompted model reached 0.79 [52]. Table 6 summarizes what the evidence of each family supports and where it is insufficient. Most families rest on fewer than eight studies, often from the same group and dataset, so these observations should be read as describing the current literature rather than intrinsic properties of the architectures.

Table 6. Summary of the evidence by methodological family. The table describes what the included studies reported; it does not rank families.

Family	Primary Studies (Challenge Entries)	Input and Data Settings	What the Included Studies Show	What the Evidence Does Not Allow
CNN-based	7 (9)	Mostly 3D; four studies on a localized region and one on lesion-containing slices only; three data sources	Loss design (boundary, uncertainty) improved Dice within single studies; standard self-configuring CNNs gave the best challenge results	Conclusions on long-range context: no CNN was compared with a hybrid model on a common external test set
Attention-based	6 (1)	Four 2D/2.5D studies; the only multi-vendor and two external-center evaluations	Within-study ablations favored attention modules; performance dropped on external centers (0.85 to 0.81; 0.38 to 0.24)	A general robustness advantage: external results come from two studies only
Hybrid CNN–Transformer	7 (2)	Mostly 3D; three studies on TDSC data	Dice of 0.58–0.65 on TDSC data and 0.60–0.81 on private datasets; one study used self-supervised pre-training and dropped from 0.74 to 0.65 on external TDSC data	Superiority over CNNs: hybrid challenge entries scored below the best CNN entries
Lightweight/NAS/ distillation	2 (0)	3D; same single-center dataset; both on a tumor-centered region	Smaller models reached Dice close to larger ones (0.77–0.78)	Deployment benefit: no study reported inference time or hardware
Semi-/weakly supervised	4 (0)	Two 2D, two 3D; three on the same dataset; one on a tumor-centered region	Uncertainty-weighted pseudo-labels, teacher–student and adversarial consistency reduced the need for labels; box supervision reached Dice 0.75	Generalization: only one study reported results on a second dataset; the box-supervised model still needs a box at inference
Foundation-model-based	1 (2)	3D or 2D slices; clicks or detected boxes required	Adapted SAM-Med3D reached Dice 0.75–0.79; unadapted SAM-Med3D reached 0.38–0.48	Fully automatic use: two of the three approaches depend on prompts or a detection stage; only SAMed (T9) does not

4. Discussion

4.1. Principal Findings

This review identified 27 primary studies and one challenge benchmark on deep learning segmentation of tumors in ABUS. Figure 2 places the primary studies by year and methodological family. Over 2020–2026, the methods moved from CNNs with region proposals or modified losses toward attention modules, hybrid CNN–Transformer encoders, architecture search and distillation, semi- and weakly supervised training and, from 2025, adapted foundation models. This progression is visible in the methods proposed, but not in the results: only one study compared CNN and hybrid models on a common external

test set [40], and on the only shared benchmark the best entries were self-configuring CNN pipelines rather than newer architectures (Section 3.11).

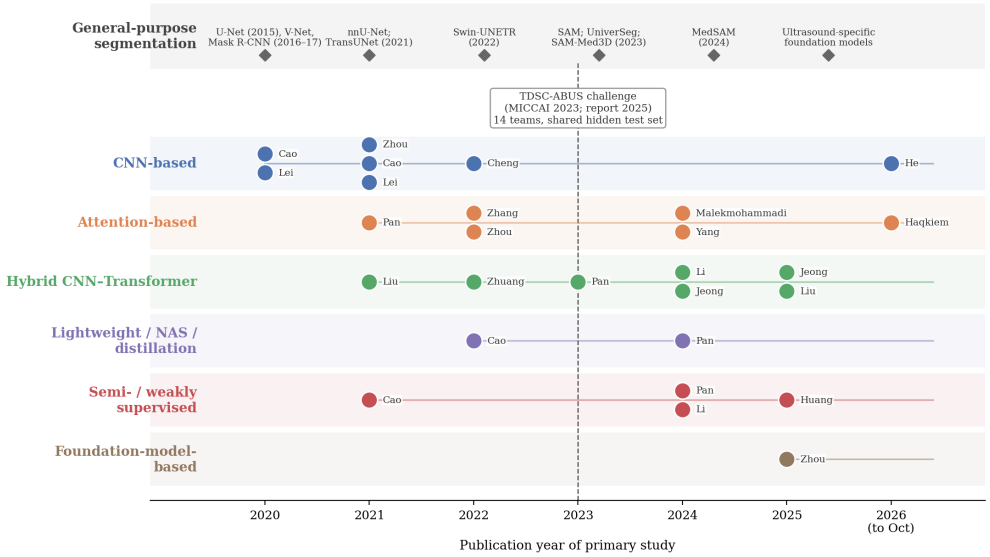


Figure 2. Methodological roadmap of deep learning tumor segmentation in ABUS. Each marker is a primary study included in this review, placed by publication year and methodological family (first author shown); marker colors indicate the methodological family. The upper band shows general-purpose segmentation methods on which the ABUS studies built: U-Net [61], V-Net [62], Mask R-CNN [63], nnU-Net [57], TransUNet [64], Swin-UNETR [65], Segment Anything [66], UniverSeg [67], SAM-Med3D [58], MedSAM [59] and ultrasound-specific foundation models [68]. The dashed line marks the TDSC-ABUS 2023 challenge. The roadmap shows when approaches appeared; it does not imply that later families perform better.

The main finding concerns comparability. Differences in input setting, data and metric definition were larger than the differences between methods (Section 3.9). Within one group and dataset, the Dice fell from 0.78 with a tumor-centered region to 0.65 with whole-volume inference [32,36], and, within one study, restricting the evaluation to cases that all compared models segmented at least partially raised it from 0.38 to 0.64 [27]. Values reported on private data, up to 0.86, should therefore be read as performance under each study's own conditions, not as estimates for unseen patients. The challenge results, with mean Dice between 0.31 and 0.61 on 70 hidden volumes, are closer to the latter.

A second finding is that design choices responded to ABUS-specific constraints rather than to a general search for accuracy: patch-based and slice-based processing for memory, boundary- and uncertainty-weighted losses for blurred margins and class imbalance, inter-slice context and post-processing for false positives, and semi-supervised or box-level supervision for annotation cost (Section 3.10). These choices were usually tested only by ablation within a single study on a single dataset, so their benefit outside that setting is unknown.

4.2. ABUS Segmentation in the Context of General-Purpose and Foundation Models

Medical image segmentation is moving from task-specific architectures toward general-purpose models. Self-configuring frameworks such as nnU-Net [57] already provide strong baselines and underlay most of the best challenge entries. Promptable models derived from Segment Anything [66], including MedSAM [59] and the volumetric SAM-Med3D [58], segment a structure from a click or a box. In-context models such as UniverSeg [67] segment new tasks from a few labeled examples without retraining, ultrasound-specific foundation models such as SAM-MedUS [68] are trained across many ultrasound tasks, and vision–

language models use text prompts, as in breast ultrasound [69]. Many of these models report results on 2D handheld breast ultrasound but not on ABUS volumes, which is why they fall outside the inclusion criteria of this review and are discussed here as context.

The ABUS evidence on these models is still limited but consistent. Without adaptation, prompted SAM-Med3D reached a Dice of only 0.38–0.48 on ABUS [51]; after low-rank adaptation with simulated clicks, it reached 0.75–0.79, including on external data [52]. In the challenge, a fine-tuned SAM variant gave the lowest Hausdorff distance but not the highest overlap [14]. Two practical points follow. First, these models need adaptation to ABUS speckle and anatomy. Second, they depend on prompts, so they either shift effort to the reader or require a detection stage, as in the challenge entry that fed YOLOv8 boxes to MedSAM. Whether this is acceptable depends on the workflow: when the radiologist has already found the lesion, a click is cheap; in whole-breast screening, it is not.

Few-shot adaptation and universal models are a promising answer to the scarcity of ABUS annotations, because they could reduce the number of annotated volumes needed for a new center or scanner. This has not yet been tested: no included study evaluated few-shot or in-context segmentation on ABUS, and, to our knowledge, no foundation model has been trained on ABUS volumes at scale. Evaluating these models on the TDSC-ABUS data, with a patient-level split and the challenge metrics, would be a simple first step.

4.3. Clinical Applicability and Deployment

Four studies used segmentation as one stage of a diagnostic pipeline (Table 3): to provide tumor shape or a volume of interest to a benign/malignant classifier [53–55], or to support the classification of BI-RADS 4 lesions across two centers [56]. They show where segmentation enters clinical decision support, but all four started from a region or images selected manually and two did not evaluate the segmentation itself. The contribution of fully automatic segmentation to diagnostic accuracy therefore remains to be measured.

Evidence on deployment is thin. Almost all datasets came from Chinese centers, one study used two scanner vendors [26], and four tested on an external center, where performance fell in three of them (for example, Dice from 0.85 to 0.81 in Yang et al. [38], and from 0.74 to 0.65 in Liu et al. [40]). No study reported inference time on clinical hardware (one reported about 1 s per patient on a research GPU [29]); the efficiency-oriented studies reported only parameter counts, and the challenge did not evaluate runtime. Segmentation accuracy also misses what matters most in whole-breast reading: how many tumors are found and how many false positives are produced per volume. Lesion-level metrics, such as sensitivity and false positives per volume, capture this better than Dice and should be reported alongside it.

To fit radiology practice, a tool must run within the reading time of a volume, cope with different scanners and acquisition settings, and produce output that a reader can check quickly, for example a contour on the coronal view with an indication of its reliability. Semi-automated approaches suit a review station where the radiologist confirms or corrects a proposal, whereas fully automatic approaches are needed for pre-reading or triage. Studies should state which of these uses they target, because it determines the appropriate input setting and evaluation.

4.4. Methodological Limitations of the Current Literature

The current evidence base is narrow. Twelve of the 27 primary studies used the same 170-volume dataset from one hospital, three used data from a second hospital, and at least eight used the TDSC-ABUS data, which come from a single center and vendor. Even on the shared private dataset, split sizes differed between studies and the reuse of identical splits was not stated, and, across all 27 studies, ten did not state whether the split was

made at the patient level. Dice was computed per slice in several 2D studies, eleven studies gave the network information about tumor location, post-processing was rarely described, and boundary metrics were reported as different Hausdorff variants in millimeters, pixels or voxels (Sections 3.9 and 3.10).

These are reporting gaps as much as methodological ones, and existing guidance could close them. The CLAIM checklist [70] covers data partitioning and the reference standard, and Metrics Reloaded [71] recommends metrics matched to the task, such as per-case Dice with a boundary metric in physical units and lesion-level metrics for detection. A minimum reporting set for ABUS segmentation could include: patient-level splits; per-volume Dice averaged over cases, with dispersion; HD95 in millimeters; results stratified by tumor size; the input setting and any localization information given to the network; post-processing; runtime and hardware; and, where possible, results on the TDSC-ABUS data.

4.5. Future Research Directions

4.5.1. Shared Benchmarks and Multi-Center, Multi-Vendor Data

The TDSC-ABUS challenge showed the value of a hidden test set evaluated by the organizers, but it covers one center and one vendor. Future benchmarks should include several centers and both major ABUS systems, keep test labels hidden, include volumes without tumors so that false positives can be measured, and evaluate runtime. Such benchmarks are a scientific prerequisite for the field, not a secondary organizational task [12,14].

4.5.2. Privacy-Preserving Collaborative Learning

Pooling ABUS volumes across hospitals raises privacy, consent and governance constraints, which partly explains why datasets remain local. Federated learning trains a shared model while the data stay at each site [72]. Split learning divides the network between the sites and a server and can be combined with homomorphic encryption to protect the exchanged features, including for U-shaped segmentation networks [73]. None of the included studies used collaborative or privacy-preserving training. Because ABUS data come from a few vendors with standardized acquisition, they are a good candidate for federated benchmarks that also measure cross-site generalization.

4.5.3. Data-Efficient Learning, Uncertainty and Calibration

Annotation cost is the main bottleneck of ABUS segmentation. Four primary studies reduced the need for voxel-level labels through semi-supervised learning or box-level supervision [48–51], and two others used unlabeled images for consistency regularization [41] or self-supervised pre-training [40]. In contrast, no challenge team used unlabeled data, although only 100 annotated training volumes were available and the 30 validation volumes had been released without labels [14]. Uncertainty-aware pseudo-labeling and consistency learning are natural candidates for this setting [74,75]. Uncertainty was used as a training signal in two studies, which weighted the loss by predictive uncertainty [31,48]. Jeong et al. [44] reported that multi-task learning improved calibration against lesion-related artifacts; however, their consistency loss was designed to suppress false positives, and their reported metrics did not include a calibration measure. Because boundaries are ambiguous and readers disagree, uncertainty should also be an output of the model, not only a training signal: models should report voxel- or lesion-level uncertainty and be checked for calibration [76], so that a reader knows when not to trust a contour.

4.5.4. Small Tumors and Fair Comparison of Architectures

Small tumors are both clinically important and the main source of failure. No included study used Tversky or focal-Tversky losses [77,78], which trade false negatives against false positives, and only three reported results by tumor size [44,45,52]. Few studies used

2.5D inputs (Yang et al. [38], and two challenge entries) or otherwise added adjacent-slice context [27,35], although 2.5D inputs are a natural compromise between 2D efficiency and 3D context in ABUS. Finally, the question of whether hybrid or foundation models outperform well-configured CNNs can only be answered by head-to-head comparisons on a common external test set. Only Liu et al. [40] made such a comparison: all models were trained on their private data and tested on the TDSC-ABUS data, where their hybrid model reached a Dice of 0.65 against 0.61 for nnU-Net. This comparison, however, comes from a single study, with baselines re-implemented by the authors, and no foundation model was included.

4.5.5. Clinical Reliability and Workflow Integration

Future work should move beyond segmentation accuracy as an isolated measure. Clinical value will depend on consistent performance across devices and sites, acceptable runtime on clinical hardware, lesion-level detection performance, outputs whose reliability the reader can check (Section 4.5.3), and evidence that segmentation improves downstream decisions, ideally in reader studies and prospective evaluations within the ABUS reading workflow [12].

5. Conclusions

Deep learning for tumor segmentation in ABUS has diversified since 2020 into six methodological families, from CNNs to adapted foundation models. The evidence does not show that any family is superior. Most studies used private, single-center data, often the same dataset, with different splits, input settings and metric definitions, and the reported Dice values depend more on these settings than on the architecture. Almost all datasets came from Chinese centers and were acquired with Siemens or GE systems, so it is not known whether the reported results hold for other populations, scanners and acquisition protocols. On the only shared benchmark, self-configuring CNN pipelines performed best, and Dice values were well below most of those reported on private data.

The next step for the field is therefore not mainly a new architecture but better evidence: multi-center, multi-vendor benchmarks with hidden test sets; standardized reporting of evaluation settings; data-efficient and privacy-preserving training; uncertainty estimates; and evaluation of runtime and clinical use within the ABUS reading workflow.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/biomedinformatics6050095/s1, Table S1: PRISMA-ScR checklist.

Author Contributions: Conceptualization, K.E., M.H.E.Y.A. and A.J.; methodology, K.E.; investigation, K.E.; resources, K.E.; data curation, K.E.; writing—original draft preparation, K.E.; writing—review and editing, M.H.E.Y.A. and A.J.; visualization, K.E.; supervision, M.H.E.Y.A. and A.J.; project administration, A.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created in this study. Data sharing is not applicable to this article as it is a literature review synthesizing previously published studies, all of which are cited in the reference list.

Acknowledgments: The authors wish to express their sincere gratitude to the researchers and clinicians whose publications have made significant contributions to the fields of automated breast ultrasound and deep learning-based medical imaging systems, and who formed the scientific basis

of this literature review. The authors would also like to thank all individuals who provided advice, technical assistance, and constructive criticism during the preparation of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

2D/2.5D/3D	Two-dimensional/stack of adjacent slices/three-dimensional
ABUS	Automated breast ultrasound
ABVS	Automated breast volume scanner
AI	Artificial intelligence
BCE/CE	Binary cross-entropy/cross-entropy
BI-RADS	Breast Imaging Reporting and Data System
BLSTM	Bidirectional long short-term memory
CAD	Computer-aided diagnosis
CNN	Convolutional neural network
CRF	Conditional random field
CV	Cross-validation
DSC	Dice similarity coefficient
HD/HD95	Hausdorff distance/95th-percentile Hausdorff distance
HDI	Human Development Index
HHUS	Handheld breast ultrasound
LoRA	Low-rank adaptation
MTL	Multi-task learning
NAS	Neural architecture search
PRISMA-ScR	Preferred Reporting Items for Systematic Reviews and Meta-Analyses extension for Scoping Reviews
ROI/VOI	Region of interest/volume of interest
SAM	Segment Anything Model
TDSC-ABUS	Tumor Detection, Segmentation and Classification Challenge on Automated 3D Breast Ultrasound
WHO	World Health Organization

References

1. Bray, F.; Laversanne, M.; Sung, H.; Ferlay, J.; Siegel, R.L.; Soerjomataram, I.; Jemal, A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **2024**, *74*, 229–263. [CrossRef] [PubMed]
2. Kim, J.; Harper, A.; McCormack, V.; Sung, H.; Houssami, N.; Morgan, E.; Mutebi, M.; Garvey, G.; Soerjomataram, I.; Fidler-Benaoudia, M.M. Global patterns and trends in breast cancer incidence and mortality across 185 countries. *Nat. Med.* **2025**, *31*, 1154–1162. [CrossRef] [PubMed]
3. World Health Organization. Breast Cancer. 2026. Available online: https://www.who.int/news-room/fact-sheets/detail/breast-cancer (accessed on 18 August 2026).
4. Henderson, J.T.; Webber, E.M.; Weyrich, M.S.; Miller, M.; Melnikow, J. Screening for breast cancer: Evidence report and systematic review for the US Preventive Services Task Force. *JAMA* **2024**, *331*, 1931–1946. [CrossRef] [PubMed]
5. Trentham-Dietz, A.; Chapman, C.H.; Jayasekera, J.; Lowry, K.P.; Heckman-Stoddard, B.; Hampton, J.M.; Caswell-Jin, J.; Lu, Y.; Gangnon, R.E.; Sun, L.; et al. *Breast Cancer Screening with Mammography: An Updated Decision Analysis*; Technical Report No. 231s; AHRQ: Rockville, MD, USA, 2024.
6. World Health Organization. *Global Breast Cancer Initiative Implementation Framework: Executive Summary*; WHO: Geneva, Switzerland, 2023.
7. Giaquinto, A.N.; Sung, H.; Newman, L.A.; Freedman, R.A.; Smith, R.A.; Star, J.; Jemal, A.; Siegel, R.L. Breast cancer statistics 2024. *CA Cancer J. Clin.* **2024**, *74*, 477–495. [CrossRef] [PubMed]

8. Iacob, R.; Iacob, E.R.; Stoicescu, E.R.; Ghenciu, D.M.; Cocolea, D.M.; Constantinescu, A.; Ghenciu, L.A.; Manolescu, D.L. Evaluating the role of breast ultrasound in early detection of breast cancer in low- and middle-income countries: A Comprehensive Narrative Review. *Bioengineering* **2024**, *11*, 262. [CrossRef] [PubMed]
9. Allajbeu, I.; Morris, K.; Nanaa, M.; Payne, N.; Charnley, K.; Moyle, P.L.; Taylor, K.; Sharma, N.; Lim, Y.; Gilbert, F.J. Introduction of automated breast ultrasound as an additional screening tool for dense breasts in the UK: A practical approach from the BRAID trial. *Clin. Radiol.* **2024**, *79*, e641–e650. [CrossRef] [PubMed]
10. Klein Wolterink, F.; Ab Mumin, N.; Appelman, L.; Derks-Rekers, M.; Imhof-Tas, M.; Lardenoije, S.; van der Leest, M.; Mann, R.M. Diagnostic performance of 3D automated breast ultrasound (3D-ABUS) in a clinical screening setting—A retrospective study. *Eur. Radiol.* **2024**, *34*, 5451–5460. [CrossRef] [PubMed]
11. Winkelman, A.J.; Tulenko, K.; Epstein, S.H.; Nguyen, J.V.; Ford, C.; Miller, M.M. Breast cancer screening with automated breast US and mammography vs handheld US and mammography in Women with Dense Breasts in a Real-World Clinical Setting. *J. Breast Imaging* **2024**, *6*, 493–501. [CrossRef] [PubMed]
12. Wang, R.; Wang, Z.; Xiao, Y.; Liu, X.; Tan, G.; Liu, J. Application of deep learning on automated breast ultrasound: Current developments, challenges, and opportunities. *Meta-Radiology* **2025**, *3*, 100138. [CrossRef]
13. Yu, S.; Liang, X.; Zhao, S.; Xie, Y.; Sun, Q. Three-Dimensional Automated Breast Ultrasound (ABUS) Tumor Classification Using a 2D-Input Network: Soft Voting or Hard Voting? *Appl. Sci.* **2024**, *14*, 11611. [CrossRef]
14. Luo, G.; Xu, M.; Chen, H.; Liang, X.; Tao, X.; Ni, D.; Jeong, H.; Kim, C.; Stock, R.; Baumgartner, M.; et al. Tumor detection, segmentation and classification challenge on automated 3D breast ultrasound: The TDSC-ABUS challenge. *arXiv* **2025**, arXiv:2501.15588. [CrossRef]
15. Manh, V.; Jia, X.; Xue, W.; Xu, W.; Mei, Z.; Dong, Y.; Zhou, J.; Huang, R.; Ni, D. An efficient framework for lesion segmentation in ultrasound images using global adversarial learning and region-invariant loss. *Comput. Biol. Med.* **2024**, *171*, 108137. [CrossRef] [PubMed]
16. Aslam, M.A.; Naveed, A.; Ahmed, N.; Ke, Z. A hybrid attention network for accurate breast tumor segmentation in ultrasound images. *Sci. Rep.* **2025**, *15*, 39633. [CrossRef] [PubMed]
17. Oladimeji, O.O.; Imran, A.A.-Z.; Wang, X.; Unnikrishnan, S. Deep learning advances in breast medical imaging with a focus on clinical readiness and radiologists' perspective. *Image Vis. Comput.* **2025**, *161*, 105601. [CrossRef]
18. Andreadis, T.; Gasteratos, A.; Seimenis, I.; Koulouriotis, D. Application of CAD systems in breast cancer diagnosis using machine learning techniques: An overview of systematic reviews. *Bioengineering* **2025**, *12*, 1160. [CrossRef] [PubMed]
19. Zou, H.; Gong, X.; Luo, J.; Li, T. A robust breast ultrasound segmentation method under noisy annotations. *Comput. Methods Programs Biomed.* **2021**, *209*, 106327. [CrossRef] [PubMed]
20. Chen, Y.; Shao, X.; Shi, K.; Rominger, A.; Caobelli, F. AI in breast cancer imaging: An update and future trends. *Semin. Nucl. Med.* **2025**, *55*, 358–370. [CrossRef] [PubMed]
21. Sun, D.; Dong, C.; Yan, Y.; Jiang, B.; Duan, Y.; Tu, Z.; Zhang, C. Challenge-aware U-Net for breast lesion segmentation in ultrasound images. *Pattern Recognit.* **2025**, *168*, 111851. [CrossRef]
22. Pérez-Núñez, J.-R.; Rodríguez, C.; Vásquez-Serpa, L.-J.; Navarro, C. The challenge of deep learning for the prevention and automatic diagnosis of breast cancer: A systematic review. *Diagnostics* **2024**, *14*, 2896. [CrossRef] [PubMed]
23. Al-Karawi, D.; Al-Zaidi, S.; Helael, K.A.; Obeidat, N.; Mouhsen, A.M.; Ajam, T.; Alshalabi, B.A.; Salman, M.; Ahmed, M.H. A review of artificial intelligence in breast imaging. *Tomography* **2024**, *10*, 705–726. [CrossRef] [PubMed]
24. Arksey, H.; O'Malley, L. Scoping studies: Towards a methodological framework. *Int. J. Soc. Res. Methodol.* **2005**, *8*, 19–32. [CrossRef]
25. Tricco, A.C.; Lillie, E.; Zarin, W.; O'Brien, K.K.; Colquhoun, H.; Levac, D.; Moher, D.; Peters, M.D.J.; Horsley, T.; Weeks, L.; et al. PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Ann. Intern. Med.* **2018**, *169*, 467–473. [CrossRef] [PubMed]
26. Malekmohammadi, A.; Soryani, M.; Kozegar, E. Mass segmentation in automated breast ultrasound using an enhanced attentive UNet. *Expert Syst. Appl.* **2024**, *245*, 123095. [CrossRef]
27. Haqkiem, N.; Abdullah, E.R.M.F.; Saw, S.N.; Tan, L.K.; Wong, J.H.D.; Rahmat, K.; Ian, P.T.W.; Ab Mumin, N. BreAST-U^{2}Net: A twin-stream U^{2}Net with attention-based tumor fusion for 2-D tumor segmentation in automated breast ultrasound. *Ultrasound Med. Biol.* **2026**, *52*, 315–327. [CrossRef] [PubMed]
28. Lei, Y.; Yao, J.; He, X.; Xu, D.; Wang, L.; Li, W.; Curran, W.J.; Liu, T.; Yang, X. Deep learning-based breast tumor detection and segmentation in 3D ultrasound image. In *Proceedings of the SPIE Medical Imaging 2020: Ultrasonic Imaging and Tomography, Houston, TX, USA, 15–20 February 2020*; SPIE: Bellingham, WA, USA, 2020; Volume 11319, p. 113190Y. [CrossRef]
29. Lei, Y.; He, X.; Yao, J.; Wang, T.; Wang, L.; Li, W.; Curran, W.J.; Liu, T.; Xu, D.; Yang, X. Breast tumor segmentation in 3D automatic breast ultrasound using Mask scoring R-CNN. *Med. Phys.* **2021**, *48*, 204–214. [CrossRef] [PubMed]

30. Cao, X.; Chen, H.; Li, Y.; Peng, Y.; Zhou, Y.; Cheng, L. Boundary loss with non-Euclidean distance constraint for ABUS mass segmentation. In Proceedings of the 2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI), Chengdu, China, 17–19 October 2020; pp. 645–650.
31. Cao, X.; Chen, H.; Li, Y.; Peng, Y.; Wang, S.; Cheng, L. Dilated densely connected U-Net with uncertainty focus loss for 3D ABUS mass segmentation. *Comput. Methods Programs Biomed.* **2021**, *209*, 106313. [CrossRef] [PubMed]
32. Zhou, Y.; Chen, H.; Li, Y.; Liu, Q.; Xu, X.; Wang, S.; Yap, P.-T.; Shen, D. Multi-task learning for segmentation and classification of tumors in 3D automated breast ultrasound images. *Med. Image Anal.* **2021**, *70*, 101918. [CrossRef] [PubMed]
33. Cheng, Z.; Li, Y.; Chen, H.; Zhang, Z.; Pan, P.; Cheng, L. DSGMFFN: Deepest semantically guided multi-scale feature fusion network for automated lesion segmentation in ABUS images. *Comput. Methods Programs Biomed.* **2022**, *221*, 106891. [CrossRef] [PubMed]
34. He, S.; Ji, D.; Ma, M.; Pan, X.; Xu, J.; Liu, F. BAMT-Net: A boundary-aware multi-task framework for 3D breast ultrasound lesion segmentation and classification. *Med. Biol. Eng. Comput.* **2026**, *64*, 2251–2267. [CrossRef] [PubMed]
35. Pan, P.; Chen, H.; Li, Y.; Cai, N.; Cheng, L.; Wang, S. Tumor segmentation in automated whole breast ultrasound using bidirectional LSTM neural network and attention mechanism. *Ultrasonics* **2021**, *110*, 106271. [CrossRef] [PubMed]
36. Zhou, Y.; Chen, H.; Li, Y.; Cao, X.; Wang, S.; Shen, D. Cross-model attention-guided tumor segmentation for 3D automated breast ultrasound (ABUS) images. *IEEE J. Biomed. Health Inform.* **2022**, *26*, 301–311. [CrossRef] [PubMed]
37. Zhang, H.; Xu, J.; Wang, M.; Li, Y. DenseATT-Net: Densely-connected neural network with intensive attention modules for 3D ABUS mass segmentation. In Proceedings of the 2022 7th International Conference on Image, Vision and Computing (ICIVC), Xi'an, China, 26–28 July 2022; pp. 348–353. [CrossRef]
38. Yang, X.; Li, X.; Qin, Y.; Wang, H.; Zhao, C.; Yin, Y. Interlayer information fusion-based and dual-attention improved U-Net for ABVS image sequence intelligent tumor segmentation. *Biomed. Signal Process. Control* **2024**, *98*, 106740. [CrossRef]
39. Liu, Y.; Yang, Y.; Jiang, W.; Wang, T.; Lei, B. 3D deep attentive U-Net with Transformer for breast tumor segmentation from automated breast volume scanner. In Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Virtual Conference, 31 October–4 November 2021; pp. 4011–4014. [CrossRef] [PubMed]
40. Liu, Y.; Li, J.; Yang, Y.; Zhao, C.; Zhang, Y.; Yang, P.; Dong, L.; Deng, X.; Zhu, T.; Wang, T.; et al. ABVS breast tumour segmentation via integrating CNN with dilated sampling self-attention and feature interaction Transformer. *Neural Netw.* **2025**, *187*, 107312. [CrossRef] [PubMed]
41. Zhuang, X.; Zhu, X.; Hu, H.; Yao, J.; Li, W.; Yang, C.; Wang, L.; Feng, N.; Xu, D. Residual Swin Transformer Unet with consistency regularization for automatic breast ultrasound tumor segmentation. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 3071–3075. [CrossRef]
42. Pan, P.; Li, Y.; Chen, H.; Zhao, X.; Chen, L.; Zhang, C. Gaussian Transformer and CNN segmentation method based on contrastive learning of boundary. In Proceedings of the 2023 8th International Conference on Image, Vision and Computing (ICIVC), Dalian, China, 27–29 July 2023. [CrossRef]
43. Jeong, H.; Yoon, C.; Lim, H.; Won, J.; Luo, G.; Xu, M.; Kim, K.; Kim, N.; Kim, C. 3D deep adaptively efficient attention model to segment automated breast ultrasound. In Proceedings of the 2024 IEEE Ultrasonics, Ferroelectrics, and Frequency Control Joint Symposium (UFFC-JS), Taipei, Taiwan, 22–26 September 2024; pp. 1–4. [CrossRef]
44. Jeong, H.; Yoon, C.; Lim, H.; Won, J.; Kim, K.; Luo, G.; Xu, M.; Kim, N.; Kim, C. A 3D self-configuring hybrid transformer with multi-task learning for 3D automated breast ultrasound segmentation. *Comput. Biol. Med.* **2025**, *198*, 111263. [CrossRef] [PubMed]
45. Li, Y.; Ren, Y.; Cheng, Z.; Sun, J.; Pan, P.; Chen, H. Automatic breast ultrasound (ABUS) tumor segmentation based on global and local feature fusion. *Phys. Med. Biol.* **2024**, *69*, 115039. [CrossRef] [PubMed]
46. Cao, X.; Chen, H.; Li, Y.; Peng, Y.; Zhou, Y.; Cheng, L.; Liu, T.; Shen, D. Auto-DenseUNet: Searchable neural network architecture for mass segmentation in 3D automated breast ultrasound. *Med. Image Anal.* **2022**, *82*, 102589. [CrossRef] [PubMed]
47. Pan, P.; Li, Y.; Chen, H.; Sun, J.; Li, X.; Cheng, L. ABUS tumor segmentation via decouple contrastive knowledge distillation. *Phys. Med. Biol.* **2024**, *69*, 015019. [CrossRef] [PubMed]
48. Cao, X.; Chen, H.; Li, Y.; Peng, Y.; Wang, S.; Cheng, L. Uncertainty aware temporal-ensembling model for semi-supervised ABUS mass segmentation. *IEEE Trans. Med. Imaging* **2021**, *40*, 431–443. [CrossRef] [PubMed]
49. Li, Y.; Cheng, Z.; Sun, J.; Chen, H.; Pan, P.; Ren, Y. Multi-domain consistency constraint model for semi-supervised lesion segmentation in automatic breast ultrasound (ABUS). *Biomed. Signal Process. Control* **2024**, *98*, 106724. [CrossRef]
50. Pan, P.; Chen, H.; Li, Y.; Li, J.; Cheng, Z.; Wang, S. Dual teacher model for semi-supervised ABUS tumor segmentation. *Biomed. Signal Process. Control* **2024**, *87*, 105514. [CrossRef]
51. Huang, Y.; Chang, A.; Dou, H.; Tao, X.; Zhou, X.; Cao, Y.; Huang, R.; Frangi, A.F.; Bao, L.; Yang, X.; et al. Flip Learning: Weakly supervised erase to segment nodules in breast ultrasound. *Med. Image Anal.* **2025**, *102*, 103552. [CrossRef] [PubMed]
52. Zhou, Y.; Ye, M.; Ye, H.; Zeng, S.; Shu, X.; Pan, Y.; Wu, A.; Liu, P.; Zhang, G.; Cai, S.; et al. Semiautomated segmentation of breast tumor on automatic breast ultrasound image using a large-scale model with customized modules. *Sci. Rep.* **2025**, *15*, 17329. [CrossRef] [PubMed]

53. Wang, Q.; Chen, H.; Luo, G.; Li, B.; Shang, H.; Shao, H.; Sun, S.; Wang, Z.; Wang, K.; Cheng, W. Performance of novel deep learning network with the incorporation of the automatic segmentation network for diagnosis of breast cancer in automated breast ultrasound. *Eur. Radiol.* **2022**, *32*, 7163–7172. [CrossRef] [PubMed]
54. Zhang, J.; Huang, Y.-S.; Wang, Y.-W.; Xiang, H.; Lin, X.; Chang, R.-F. Automated whole-breast ultrasound tumor diagnosis using attention-inception network. *Phys. Med.* **2025**, *134*, 104989. [CrossRef] [PubMed]
55. Xiang, H.; Huang, Y.-S.; Lee, C.-H.; Chang Chien, T.-Y.; Lee, C.-K.; Liu, L.; Li, A.; Lin, X.; Chang, R.-F. 3-D Res-CapsNet convolutional neural network on automated breast ultrasound tumor diagnosis. *Eur. J. Radiol.* **2021**, *138*, 109608. [CrossRef] [PubMed]
56. Li, Y.; Li, C.; Yang, T.; Chen, L.; Huang, M.; Yang, L.; Zhou, S.; Liu, H.; Xia, J.; Wang, S. Multiview deep learning networks based on automated breast volume scanner images for identifying breast cancer in BI-RADS 4. *Front. Oncol.* **2024**, *14*, 1399296. [CrossRef] [PubMed]
57. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef] [PubMed]
58. Wang, H.; Guo, S.; Ye, J.; Deng, Z.; Cheng, J.; Li, T.; Chen, J.; Su, Y.; Huang, Z.; Shen, Y.; et al. SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. *arXiv* **2023**, arXiv:2310.15161.
59. Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nat. Commun.* **2024**, *15*, 654. [CrossRef] [PubMed]
60. Barkhof, F.; Abbring, S.; Pardasani, R.; Awasthi, N. Deep learning based tumor detection and segmentation for automated 3D breast ultrasound imaging. In Proceedings of the 2024 IEEE South Asian Ultrasonics Symposium (SAUS), Gujarat, India, 27–29 March 2024; pp. 21–24. [CrossRef]
61. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
62. Milletari, F.; Navab, N.; Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In Proceedings of the Fourth International Conference on 3D Vision (3DV), Stanford, CA, USA, 25–28 October 2016; pp. 565–571.
63. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
64. Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A.L.; Zhou, Y. TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2021**, arXiv:2102.04306.
65. Hatamizadeh, A.; Nath, V.; Tang, Y.; Yang, D.; Roth, H.R.; Xu, D. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries (BrainLes 2021)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2022; Volume 12962, pp. 272–284.
66. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.-Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 4015–4026.
67. Butoi, V.I.; Ortiz, J.J.G.; Ma, T.; Sabuncu, M.R.; Guttag, J.; Dalca, A.V. UniverSeg: Universal medical image segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023; pp. 21438–21451. [CrossRef]
68. Tian, F.; Zhai, J.; Gong, J.; Lei, W.; Chang, S.; Ju, F.; Qian, S.; Zou, X. SAM-MedUS: A foundational model for universal ultrasound image segmentation. *J. Med. Imaging* **2025**, *12*, 027001. [CrossRef] [PubMed]
69. Mallina, R.; Shareef, B. XBusNet: Text-Guided Breast Ultrasound Segmentation via Multimodal Vision–Language Learning. *Diagnostics* **2025**, *15*, 2849. [CrossRef] [PubMed]
70. Mongan, J.; Moy, L.; Kahn, C.E., Jr. Checklist for Artificial Intelligence in Medical Imaging (CLAIM): A guide for authors and reviewers. *Radiol. Artif. Intell.* **2020**, *2*, e200029. [CrossRef] [PubMed]
71. Maier-Hein, L.; Reinke, A.; Godau, P.; Tizabi, M.D.; Buettner, F.; Christodoulou, E.; Glocker, B.; Isensee, F.; Kleesiek, J.; Kozubek, M.; et al. Metrics reloaded: Recommendations for image analysis validation. *Nat. Methods* **2024**, *21*, 195–212. [CrossRef] [PubMed]
72. Rieke, N.; Hancox, J.; Li, W.; Milletarì, F.; Roth, H.R.; Albarqouni, S.; Bakas, S.; Galtier, M.N.; Landman, B.A.; Maier-Hein, K.; et al. The future of digital health with federated learning. *npj Digit. Med.* **2020**, *3*, 119. [CrossRef] [PubMed]
73. Yang, Z.; Chen, Y.; Huangfu, H.; Ran, M.; Wang, H.; Li, X.; Zhang, Y. Dynamic corrected split federated learning with homomorphic encryption for U-shaped medical image networks. *IEEE J. Biomed. Health Inform.* **2023**, *27*, 5946–5957. [CrossRef] [PubMed]
74. Chen, Y.; Yang, Z.; Shen, C.; Wang, Z.; Zhang, Z.; Qin, Y.; Wei, X.; Lu, J.; Liu, Y.; Zhang, Y. Evidence-based uncertainty-aware semi-supervised medical image segmentation. *Comput. Biol. Med.* **2024**, *170*, 108004. [CrossRef] [PubMed]
75. Chen, F.; Chen, L.; Kong, W.; Zhang, W.; Zheng, P.; Sun, L.; Zhang, D.; Liao, H. Deep semi-supervised ultrasound image segmentation by using a shadow aware network with boundary refinement. *IEEE Trans. Med. Imaging* **2023**, *42*, 3779–3793. [CrossRef] [PubMed]

76. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017*; PMLR: Cambridge, MA, USA, 2017; Volume 70, pp. 1321–1330.
77. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Machine Learning in Medical Imaging (MLMI 2017)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2017; Volume 10541, pp. 379–387.
78. Abraham, N.; Khan, N.M. A novel focal Tversky loss function with improved attention U-Net for lesion segmentation. In Proceedings of the IEEE 16th International Symposium on Biomedical Imaging (ISBI), Venice, Italy, 8–11 April 2019; pp. 683–687.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.