scientific reports

Check for updates

OPEN

An AI-enabled comprehensive breast ultrasound diagnostic system for low-resource settings without a sonographer or a radiologist

Emilio J. Ochoa[1], Luis C. Revilla[2], Stefano E. Romero[2], Giancarlo A. Guarnizo[2], Thomas J. Marini[3], Kevin J. Parker[4], Yu T. Zhao[3], Galen Brennan[3], Jonah Kan[5], Steven Meng[6], Ann Dozier[7], Anna Weiss[8,9] & Benjamin Castaneda[1]✉

Breast cancer is the most common non-skin related malignancy and the leading cause of cancer death in women. Mammography remains the gold standard for early detection; however, its accessibility is limited in low-resource settings due to cost and technical complexity. Ultrasound (US) is a viable alternative, but its implementation is hindered by the scarcity of trained radiologists and sonographers. Volume sweep imaging (VSI) has addressed the issue of US acquisition by enabling non-specialists to perform standardized scans. However, these still require expert interpretation, limiting their impact. To overcome this barrier, we propose a fully automated Breast VSI (VSI-B) system integrating artificial intelligence (AI) for segmentation and classification of breast lesions, aiming to provide an accessible diagnostic tool for low-resource environments. This study developed an AI-driven diagnostic system for VSI-B, combining a segmentation model (Attention U-Net 3D) with a classification model for lesion detection. A total of 98 patients with palpable breast lumps were included in the study. The dataset consisted of 392 VSI-B US videos and 2,100 classified frames. A new method was implemented to enhance mass identification by selecting key frames for analysis. A majority voting algorithm was used to optimize lesion classification. The system's performance was assessed based on sensitivity, specificity, and accuracy. Following a detection step that achieved 100% sensitivity and 93.6% specificity for cancer and no cancer patients, as well as 95.0% sensitivity and 63.0% specificity for mass and no mass patients, the DenseNet classification model reached 87% accuracy, 100% sensitivity, and 83% specificity. A majority voting algorithm optimized classification, yielding an AUC of 0.91. This study highlights the potential of an AI-enabled VSI-B system as a reliable diagnostic tool in low-resource settings. By integrating multi-modal segmentation and classification, the system automates breast lesion detection and stratification, reducing reliance on radiologists. The results suggest that this approach could enhance early breast cancer diagnosis and guide clinical decision-making, particularly in underserved regions.

Keywords VSI, Breast ultrasound, Breast cancer research, Artificial intelligence, Telemedicine

Approximately 2.1 million women are diagnosed with breast cancer worldwide each year, and approximately 685,000 deaths occur annually[1]. Breast cancer incidence and mortality are expected to increase, and it is

[1]Department of Biomedical Engineering, University of Rochester, Rochester, NY, USA. [2]Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Lima, Peru. [3]Department of Imaging Sciences, University of Rochester Medical Center, Rochester, NY, USA. [4]Department of Electrical Engineering, University of Rochester, New York, USA. [5]Department of Internal Medicine, Mayo Clinic, Rochester, MN, USA. [6]Department of Radiology, UT Southwestern Medical Center, Dallas, TX, USA. [7]Department of Public Health, University of Rochester Medica Center, Rochester, NY, USA. [8]Division of Surgical Oncology, Department of Surgery, University of Rochester Medical Center, Rochester, NY, USA. [9]Wilmot Cancer Center, University of Rochester Medical Center, Rochester, NY, USA. ✉email: bcastane@ur.rochester.edu

projected to reach more than 3.2 million new cases by 2030[2], with at least 60% of deaths in low- and middle-income countries (LMICs)[3].

Early diagnosis by mammography is often associated with the term screening in high-income countries, based on the assumption that mammography will always be required[4]. Unfortunately, use of mammography equipment in LMICs is limited by the shortage of human and economic resources, especially in rural zones. Mexico has previously reported that less than 10% of patients are diagnosed with early stage breast cancer whereas 47% are diagnosed with most advanced stages (III and IV)[5].In a recent study of breast cancer diagnosis delays in women in Peru, 93% of the subjects self-diagnosed their own breast cancer, with an average time of 407 days from the onset of symptoms to the start of treatment[6].Given the severity of this public health problem, novel solutions are urgently needed to improve the accessibility of medical imaging in LMICs, especially in rural regions. Early detection drastically improves outcomes: when breast cancer is diagnosed at Stage I—defined as a tumor 2 cm or smaller with no lymph node involvement (Stage IA) or a similarly small tumor with microscopic nodal metastases (Stage IB), the five-year breast cancer-specific survival rate is 98%–100%.By contrast, Stage III disease—characterized by locally advanced tumors with significant regional lymph node spread or invasion of the chest wall or skin, but without distant metastases—carries a five-year survival of only 66%–98%[7].

Point-of-care ultrasound (US) is a portable, noninvasive, and cost-effective medical imaging modality with the potential to improve diagnosis accessibility. However, deployment of this modality requires trained specialists for both acquisition and diagnosis with years of experience, increasing the risk of late diagnosis and, therefore, of severe or fatal complications[8]. To standardize and reduce the training for the acquisition process; Volume sweep imaging (VSI) is a streamlined and time-efficient asynchronous standardized protocol to enable the acquisition of US videos by inexperienced personnel and has been validated for obstetric, lung, right upper quadrant, thyroid and breast scanning[8–20].

In the breast VSI (VSI-B) protocol, the operator, after a 2-hour training, sweeps the US probe over the target region to obtain a complete volumetric acquisition. The video clips are then sent to a specialist for interpretation and diagnosis. These sweeps are standardized and based exclusively on external body landmarks. Figure 1 illustrates the VSI-B protocol, outlining the steps required to obtain a volumetric acquisition of the breast using an US probe. High levels of agreement between the VSI-B and the standard of care for mass visualization have been reported (Cohen's Kappa 0.95; CI: 0.89-1). However, there is still no automation process to interpret the images; VSI-B still relies on specialists which are experiencing broad shortages, resulting in an increased patient load for diagnosis and an immediate result even more challenging[8].

Artificial intelligence (AI) may be the answer to this problem. AI has demonstrated potential to improve the accuracy of breast cancer diagnosis by US, resulting in high sensitivity (85%; 95% CI 70–94%) and specificity (73%; 95% CI 56–85%) values using 2D US datasets[21].

The transition to incorporate AI models within medical equipment begins with systems or software to aid in the detection and characterization of lesions or tissue abnormalities[22]. A recent study evaluated the accuracy of Samsung's S-Detect model, a proprietary system requiring images from Samsung devices, for the classification of breast lesions using one US image per patient from the VSI-B dataset, selected by a physician. The results demonstrated a high level of agreement between the model's classifications of cancer, cysts, fibroadenomas, and lipomas, compared to the standard expert report ($\kappa = 0.73$). Furthermore, the model achieved a sensitivity of 100% in detecting the 20 malignant cases, along with a specificity of 86%[14].

Other state-of-the-art architectures have been proposed, such as the Attention U-Net and the Sharp attention U-Net[23]. Attention U-Net 3D developed important improvements over the classical model, such as the soft attention mechanism. These architectures have already been applied and tested in US imaging with the breast US images (BUSI) dataset[24]. However, it is important to note that the BUSI dataset is not specifically designed for low-resource settings; it consists only of segmented frames rather than video sequences, unlike the VSI-B protocol, which is better suited for such environments. Models for segmentation using the VSI protocol have been proposed in both obstetric[15,25,26] and breast[27,28] protocols. Systems that integrate both detection and classification of breast nodules have not yet been developed in low-resource areas.

The present research aimed to develop an automated system to support the segmentation, detection and classification of palpable breast lumps for the early diagnosis of breast cancer. Our system was designed with portable, inexpensive, VSI-B US data for conditions where a radiologist or sonographer is not available, so that a provisional recommendation can be determined where no other local option exists.

Materials and methods

System description

The proposed system for this project is shown in Fig. 2. A more detailed technical description is listed here in nine steps (Fig. 3). Step one involves direct intervention from healthcare personnel. Steps two to eight are an automated process:

1. Clinical evaluation, referral of the patient to medical imaging and acquisition of VSI-B protocol.
2. Preprocessing of the data for segmentation.
3. Segmentation with Attention U-Net 3D, followed.
4. Post-processing with a Gaussian mask. A discrimination algorithm defines whether it is a "non-mass" patient or a "mass" patient to pass to the binary classification stage.
5. Selection of four videos and five representative frames for each video.
6. Frame classification of lesions is conducted: "Possibly benign" or "Possibly malignant".
7. Threshold algorithm decision.
8. Binary results are delivered to support the classification for the physician's analysis.

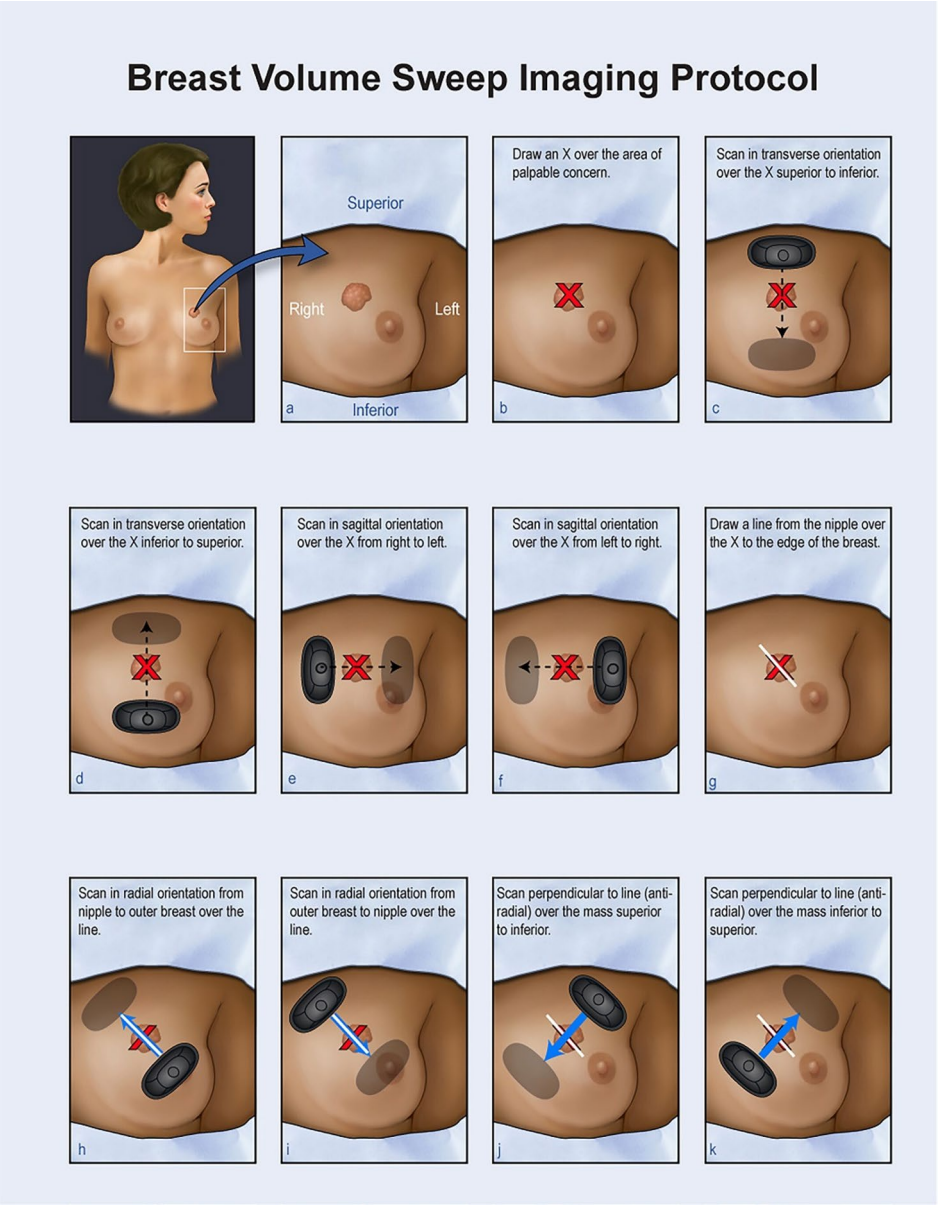

Fig. 1. Illustration of the standardized VSI-B protocol for breast mass evaluation. The process involves 8 US sweeps over the area of palpable concern. (**a**) The area of interest is identified. (**b**) An 'X' is marked over the palpable mass. (**c**–**f**) Transverse and sagittal sweeps are performed in superior-to-inferior, inferior-to-superior, right-to-left, and left-to-right directions. (**g**) A line is drawn from the nipple to the marked area. (**h**–**i**) Radial sweeps are performed along this line, from nipple to outer breast and back. (**j**–**k**) Anti-radial sweeps are done perpendicular to the line, covering superior-to-inferior and inferior-to-superior directions. This systematic approach enhances mass visualization[8].[https://doi.org/10.1002/jum.16047].

These steps will be described in further detail below, along with the one-time initialization of the data and training of the model.

VSI-B dataset acquisition and ground truth

US data was collected in a previous work by the University of Rochester, New York[8]; using the VSI protocol applied to breast as shown in Fig. 1. The VSI-B protocol was carried out by healthcare workers in clinics after a two-hour training. A hand-held US probe IQ+ (Butterfly Network, MA, USA) with a small organs pre-setting was used. This probe was connected to a tablet for acquisition. Results were compared with standard-of-care (SOC) US images from a Logiq e10 scanner (General Electric, MA, USA) or an Epiq 7G scanner (Phillips, Amsterdam)[8].

The sample set consisted of 160 patients with a total of 170 palpable breast lumps. For our evaluation, we included all patients with a cyst, cancer, fibroadenoma, or normal exam. As this is a preliminary study of the potential of AI, we excluded any discrepant cases between VSI and standard of care, any other category of

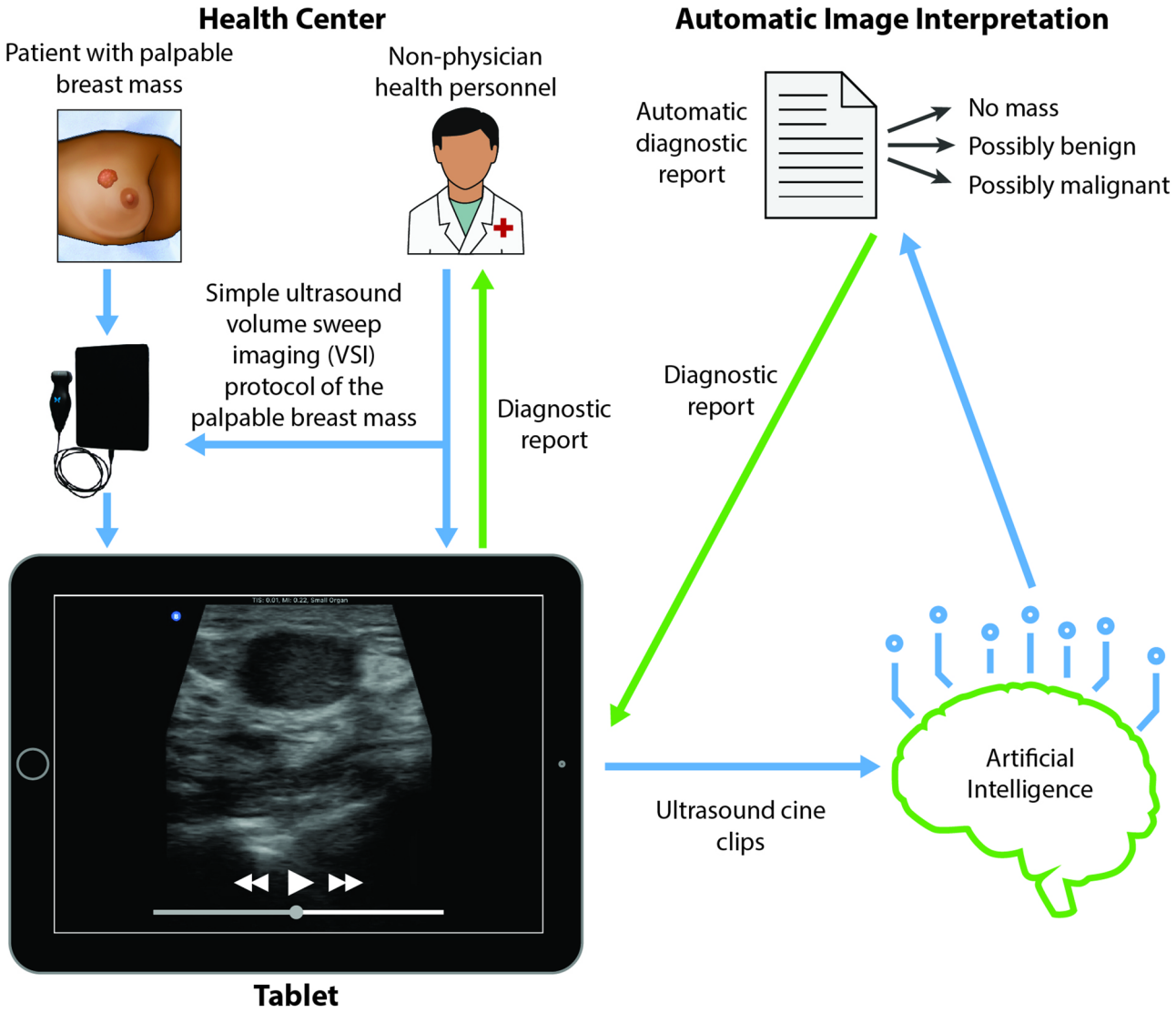

Fig. 2. Illustrated workflow of an automated breast lesion assessment system. A non-physician health worker performs a simple VSI-B protocol on a patient with a palpable breast mass, capturing US cine clips on a tablet. These clips are then uploaded to an artificial intelligence (AI) system for automatic image interpretation, producing a diagnostic report that categorizes the lesion as "no mass," "possibly benign," or "possibly malignant". Blue arrows represent the technical flow of information: ultrasound acquisition, image transfer, and automated interpretation. Green arrows represent the diagnostic feedback loop, showing how the final report is transmitted back to the health worker to guide clinical decision-making.

lesions, and cases where the lesion was unsegmentable. Pathology was determined by biopsy whenever possible. However, in cases where biopsy was not performed, BI-RADS 2 or 3 were considered benign. Patients categorized as "non-mass" in the ground truth did not present any masses on US imaging. Patients categorized with a "mass" had either a cyst, cancer, or fibroadenoma. For patients with a detected mass, "non-cancer" indicates that the mass was visible on US, but SOC determined that it was not malignant (non-mass, fibroadenoma, cyst). Finally, cases with malignancy were "cancer."

For patients with detected masses, four US videos were selected per patient, one of each type of sweep (transverse, sagittal, radial, and antiradial). Sweeps were chosen between each pair based on the sweep visualizing the largest portion of the mass. If both sweeps showed equal amounts of the mass, the first sweep was chosen. These videos were analyzed and labeled by a trained radiologist using the "Video Labeler App" tool from MATLAB R2021 (MathWorks, Natick, Massachusetts, USA), generating masks to delineate tumor shapes.

Table 1 provides a detailed breakdown of patients with confirmed diagnoses of cancer or non-cancer, including those with cysts and fibroadenomas, resulting in a total of 98 subjects.

Due to the variable length of the videos, ground truth masks were generated only on the areas where the tumor was located (Fig. 4). The size of the video frames with their labeled masks were originally 1696 × 1080 pixels. Frames with ground truth annotations assigned were interpolated in the intermediate frames since the

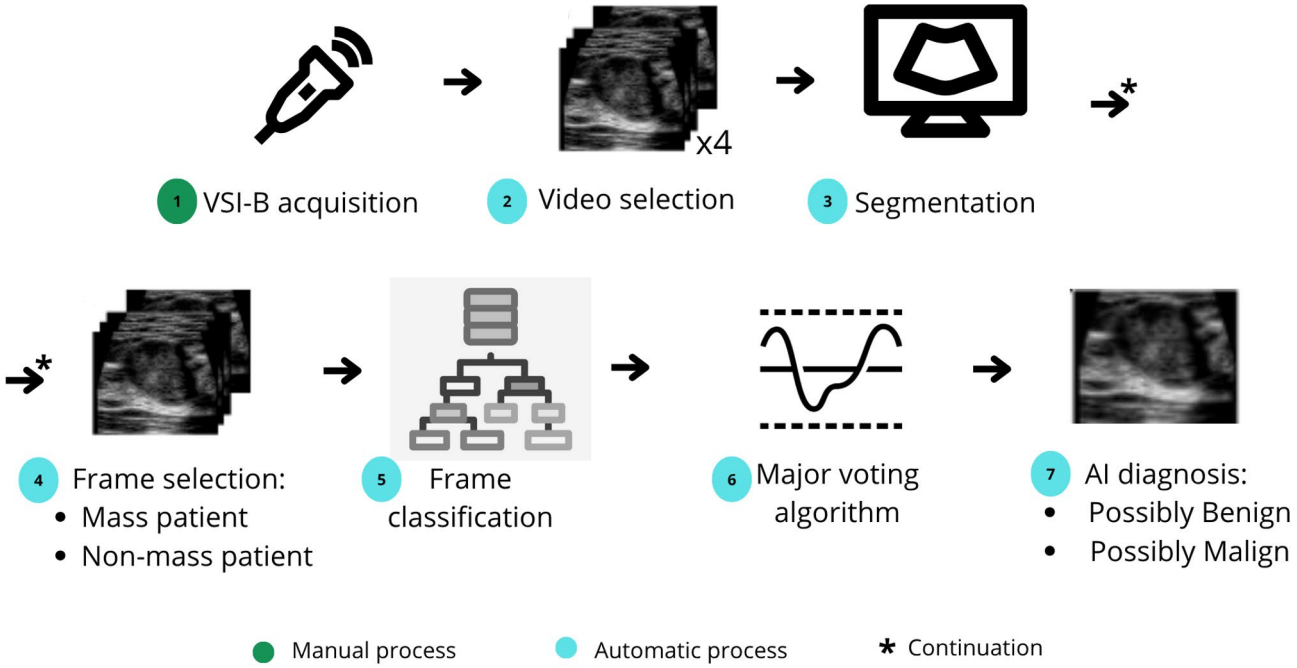

Fig. 3. Proposed technical system. Step 1 involves a direct intervention from the healthcare personnel, while steps 2–8 are an automated process.

Variable	Category	*N* (%)
Age		40 ± 14*
Sex	Female	98 (100%)
Race	African American Hispanic White	20 (20.0%) 5 (5.0%) 73 (75.0%)
BMI		29 ± 7*
Time since apparition		595 ± 1325*
Breast with palpable mass	Left Right	50 (51.0%) 48 (49.0%)
Pathology	No mass Benign Malign	38 (39.0%) 47 (48.0%) 13 (13.0%)
Pain	Yes	44 (45.0%)

Table 1. Characteristics of the sample studied ($N = 98$). *Median (IQR).

doctor segmented the ground truth by skipping one frame[29,30]. Black masks were created as background in regions without labels. These steps produced masks for the entire video, allowing for a comparison between the medical ground truth and the predicted outcomes. For non-mass patients, entirely black labels were generated for all the videos.

Preparation of videos for segmentation

To preprocess the segmentation dataset, we selected the four labeled sweeps per patient. To keep the "no-mass" cohort balanced, we likewise chose four sweeps for each of those patients and applied black masks to them, yielding a total of 392 videos (98 patients). In the 2D domain, 1696 × 1080 pixels were cropped to 912 × 912 to only consider the US images and not the Butterfly IQ + interface. After this, a resize to 128 × 128 was proposed due to computational limitations. The video length was resized to 128 frames to obtain 3D volumes, ensuring an average resize between all the videos (average length of the videos: 222 frames). Contrast limited adaptive histogram equalization (CLAHE) was applied as a preprocessing step. A K-fold cross-validation (k = 10) was performed as a method of training and validation to maximize data utilization while obtaining stable performance estimates across folds in a relatively small video dataset. To evaluate the model's performance, two metrics were utilized: the Dice coefficient and accuracy, as specified in Eqs. 1 and 2, respectively. The Dice coefficient, as shown in Eq. 1, quantifies the similarity between two sets A and B, where A represents the predicted segmentation mask

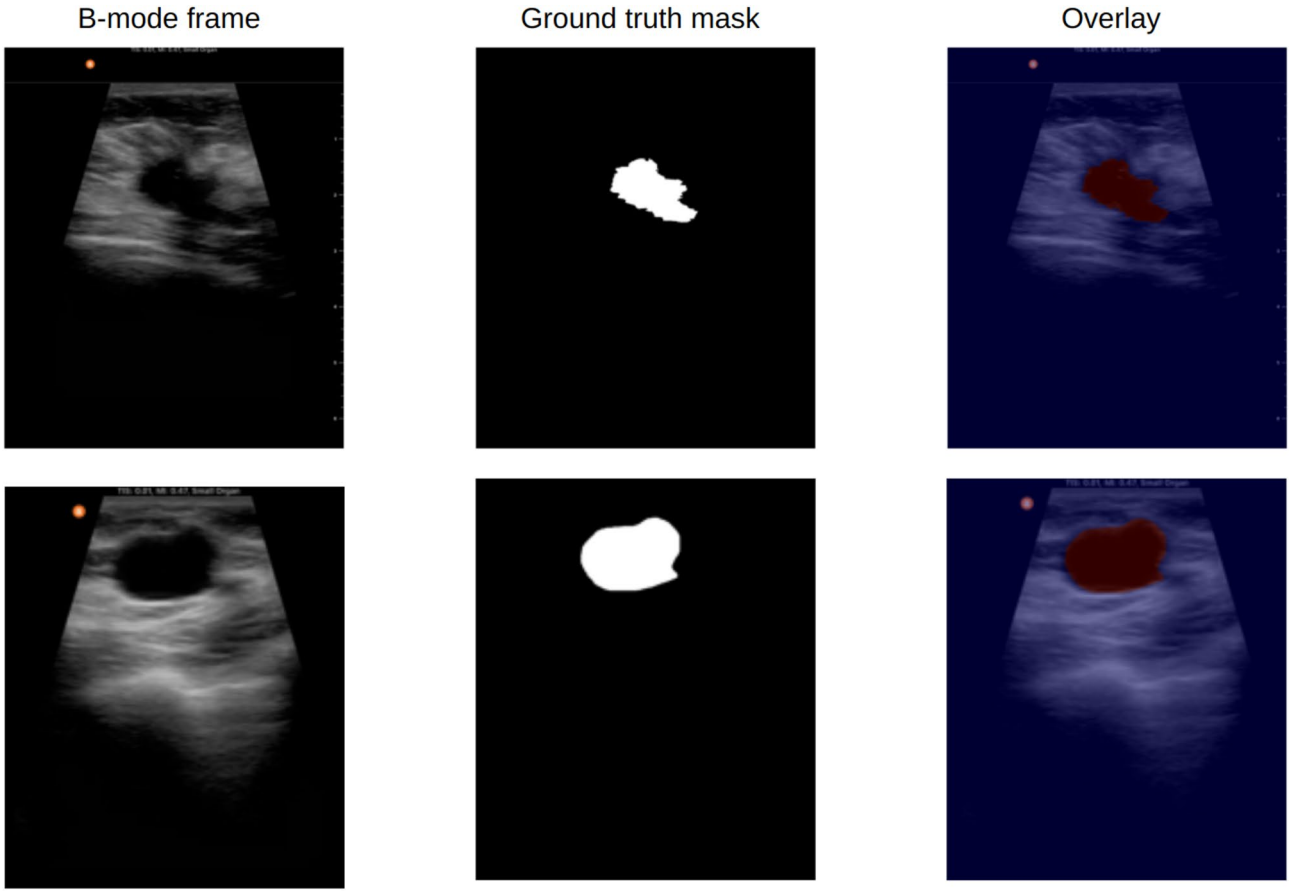

Fig. 4. B-mode image, physician-segmented mask, and overlay in two different patients. Top: segmented tumor represents a malignant case in the sixth sweep (radial). Bottom: benign tumor in the eighth sweep (perpendicular).

and *B* represents the ground truth segmentation mask. Accuracy, on the other hand, refers to the proportion of total predictions that the model correctly classifies as either mass or non-mass patients.

$$Dice = \frac{2 \parallel A \cap B \parallel}{\parallel A \parallel + \parallel B \parallel} \tag{1}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{True\ Positive + True\ Negative + False\ Positive + False\ Negative} \tag{2}$$

Segmentation

Tumor segmentation was performed by Attention U-Net 3D (Fig. 5), an architecture tested on the BraTS 2019 dataset (Multimodal Brain Tumor Segmentation Challenge dataset)[31]. As a 3D attention map was obtained, average pooling for channel correlation was performed in parallel. Finally, skip-connections were fused to reduce sparsity and singularity, also improving generic learning and segmentation prediction. Attention U-Net 3D hyperparameters included: 90 epochs, a batch size of 1, Binary Cross Entropy Dice loss with 0.5 weight, and a net depth of 4 layers (16 to 128 filters; 256 filters in the bridge). Training was processed on a 3080 GPU (NVIDIA, Santa Clara, California, USA) using PyTorch Lightning 2.0[32] and Python 3.9.12.

Mask postprocessing

To improve the tumor detection process using US imaging, the analysis began with a plot of the normalized mask area versus frame, where 100% indicated that a full mask covered the entire image. This plot was independent by volume: if a frame contained two separate masks, they were treated as distinct volumes and represented by two individual area measurements. This approach, illustrated in Fig. 6, enabled a precise analysis of each potential tumor site as an independent entity, ensuring that the data was not conflated.

To enhance the accuracy of predictions, a 3D Gaussian window was applied to this plot. Figure 6c shows this Gaussian voxel and its outputs, which played a vital role in smoothing the data outputs from our Attention U-Net 3D model, reducing temporal and spatial noise and emphasizing areas with the highest segmentation probability, assigning greater weight to the central frames—where the likelihood of correctly identifying a tumor was higher—while reducing the influence of peripheral frames, which were more prone to noise. By multiplying

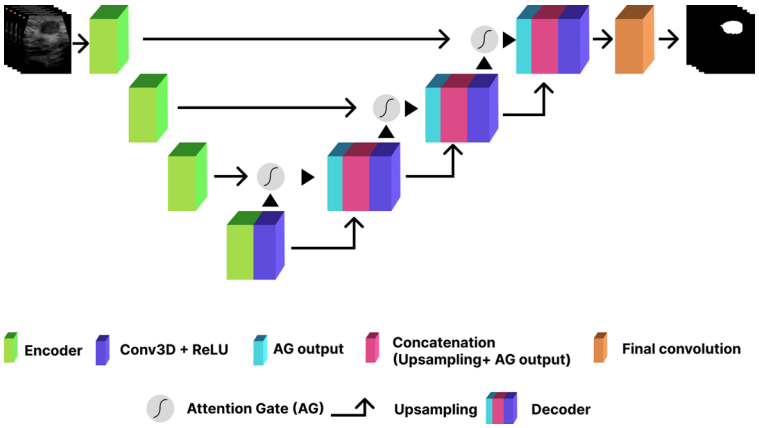

Fig. 5. Diagram depicting the architecture of the 3D Attention U-Net model, designed for volumetric medical image segmentation. The model processes 3D US volumes as input, applying multiple encoding and decoding blocks with attention mechanisms to enhance the segmentation of relevant structures. The encoding path captures spatial and contextual information through convolutional and downsampling operations, while the decoding path progressively reconstructs the segmented volume using upsampling and concatenation. Attention gates are integrated to suppress irrelevant regions and focus on salient anatomical features, improving the precision of segmentation.

the Gaussian voxel with the segmented mask, noise was effectively minimized, allowing for the extraction of more relevant information.

After applying the Gaussian filter, every data point below a 0.01% normalized area was used to eliminate small, irrelevant areas considered as noise, further refining the data. The threshold value of 0.01% was tuned based on a range from 0.005% to 0.04%; the highest detection accuracy was achieved at 0.01%. This threshold was particularly important because it corresponded to a tumor diameter of approximately 0.2 cm, a size generally not considered clinically significant for breast cancer. This approach allowed the focus to remain on more substantial and potentially hazardous tumor indications, ensuring that the analysis was both precise and clinically relevant without being overly conservative.

Finally, after thresholding, a volume that exhibits the highest peak area in the plot was selected. Figure 7 shows this pivotal selection to identify the most significant volumetric data in a 2D plot, representing the most probable tumor presences. The protocol is explicitly designed for assessment around a single clinically palpable lesion per acquisition. When clinical suspicion indicates multiple palpable lesions, the protocol specifies separate VSI acquisitions for each lesion. Therefore, our current model assumes the presence of a single lesion per VSI scan and is not intended for segmentation and separation of multiple simultaneous lesions.

To address the issue of false segmentations in non-mass patients and refine the classification of patients as either mass or non-mass, the post-processing discrimination algorithm in Fig. 8 was implemented. This discrimination algorithm evaluates the number of masks generated per video for each patient and decides which patient will be detected as a non-mass or mass patient based on predefined criteria. Initially, for each patient, the presence of segmented masks is counted across the videos. The algorithm evaluates the presence of segmentation masks across four predictions per patient. If two or more predictions are classified as "empty" (no mass detected), the patient is classified as a "non-mass" patient. Conversely, if fewer than two predictions are empty, the patient is classified as having a "presence of mass" and proceeds to the classifier for further analysis. This threshold was selected after validation across different thresholds to balance sensitivity and specificity.

Frames selection for classification

All selected frames maintained the original resolution for classification. Since classification was a 2D task, similarity between the ground truth of the same patient was addressed through a data cleaning process. For this purpose, 5 frames per video per patient were initially selected (all selected frames had been labeled ground-truth): the frame with the largest tumor area, the two previous frames, and the two subsequent frames. Previous and subsequent frames were randomly selected with the condition of exceeding 80% of the tumor area size of the frames. The target assignment for these frames was based on histopathology results, with cancer frames considered malignant and all other frames considered benign. Final data for training consisted of 2100 frames. The resize of 224×224 pixels was used as it is a standard input dimension for many convolutional neural networks (CNNs) and has been empirically determined to offer a balance between computational efficiency and sufficient detail retention for effective feature extraction in image classification tasks. Leave-one-out cross-validation (LOOCV) was performed as the training dataset was split into different folds (one per patient), resulting in 60 folds as it is more suitable for limited sample sizes and ensures complete independence between training and testing data by leaving out all data from one patient at a time.

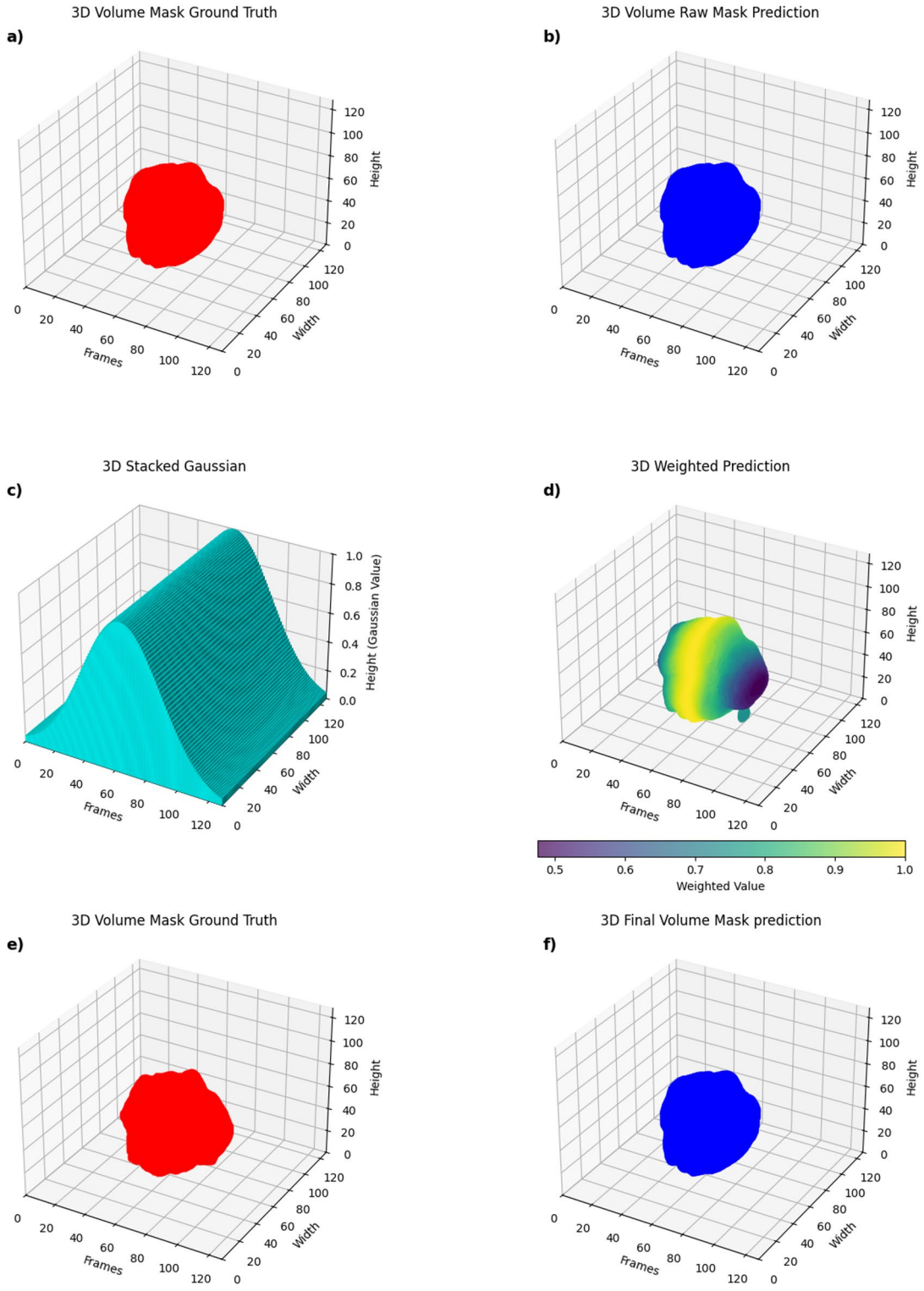

Fig. 6. Mask postprocessing with a 3D stacked Gaussian filter. Panels (a) and (e) show the original 3D volume ground truth mask for comparison (same figure). Panel (b) presents the raw 3D volume mask prediction obtained directly from the model, which exhibits initial irregularities. Panel (c) demonstrates the 3D stacked Gaussian filter applied to the mask, where the Gaussian values are distributed across the volume, as indicated by the smooth surface. Panel (d) visualizes the weighted prediction after applying the Gaussian filter, highlighting the refined probabilistic values with a colormap that scales from 0.5 to 1.0. Finally, panel (f) displays the postprocessed 3D volume mask prediction.

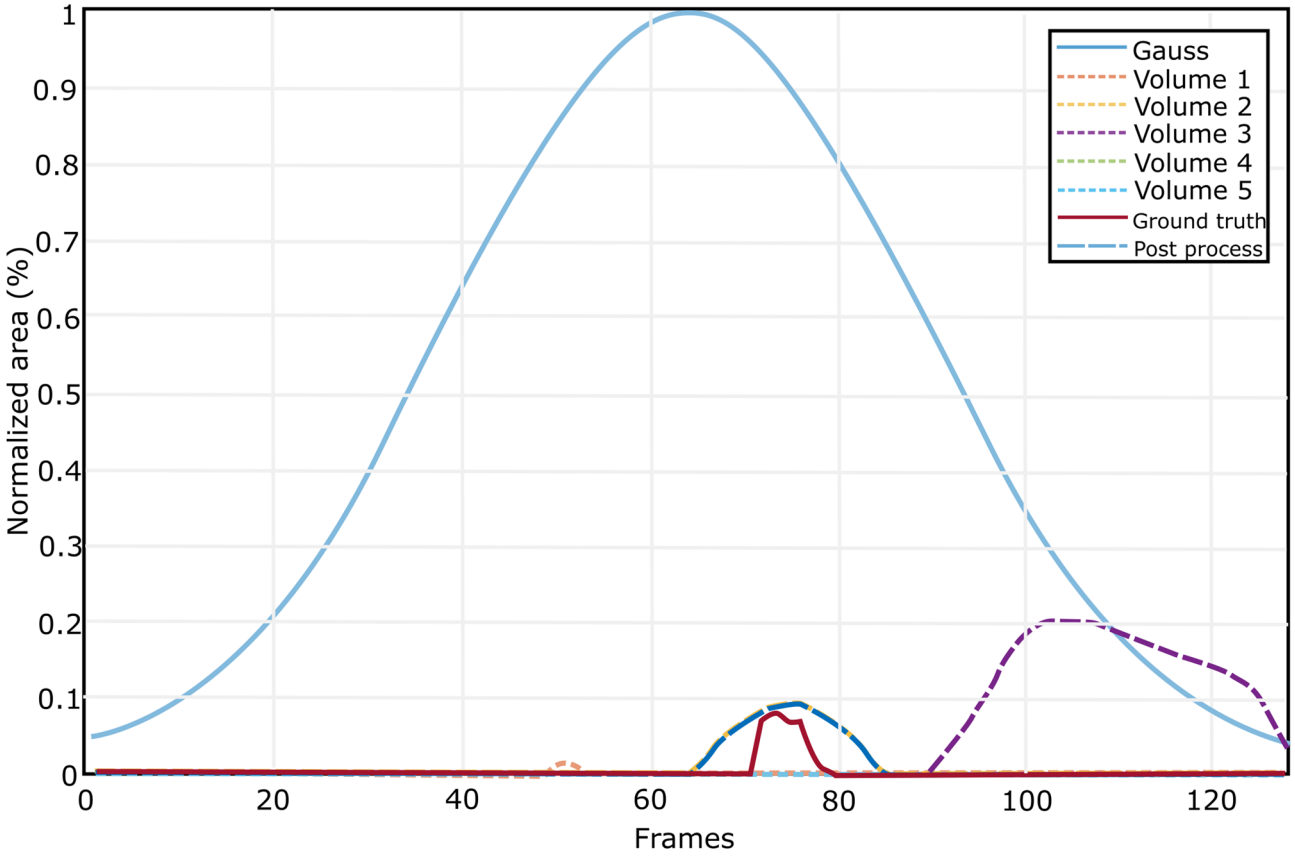

Fig. 7. Normalized mask area vs. frame after post-processing in a 2D representation. While five segmented volumes were originally detected, the post-processing step correctly identifies the ground-truth volume as the only valid mask. Volumes 1, 2, and 4 are removed using a 0.01% threshold, Volume 3 is discarded based on the Gaussian probability plot, and Volume 5 remains as the final desired mask.

Frame classification

Training was carried out on five CNN architectures, commonly used in deep learning (DL) for image recognition tasks, particularly in medical imaging: MobileNet, DenseNet, ResNet, Vgg16, and Inception[33–37]. MobileNet utilizes depthwise separable convolutions to significantly reduce the number of parameters, making it efficient for mobile and embedded vision applications. DenseNet connects each layer to every other layer in a feed-forward fashion, which improves gradient flow and encourages feature reuse. ResNet employs residual learning by introducing shortcut connections that bypass one or more layers, allowing for the training of very deep networks. Vgg16, known for its simplicity, uses 16 layers with small 3 × 3 convolutions to increase depth and feature extraction capabilities. Inception, also known as GoogLeNet, incorporates multi-scale convolutions within the same layer to capture various spatial patterns and reduce computational cost. These architectures have been effectively utilized in various medical imaging applications, such as tumor detection, organ segmentation, and disease classification, due to their ability to accurately extract and interpret complex features from medical images[38]. A comparative analysis of accuracy, sensitivity, and specificity was made to select the best model for the final evaluation strategy.

Major voting algorithm

To make a final decision on the classification of all tested frames per patient, a majority voting algorithm was designed. Different decision thresholds (20–50%) were used to assess the feasibility of the proposed pipeline for VSI-B application. This major voting algorithm determined the classification as malignant if a specific number of frames, exceeding the threshold, were classified as malignant. Prior to this, an analysis of various cases was conducted using only the best-performing classification model (as determined by accuracy, sensitivity, and specificity metrics) to establish the optimal decision threshold, ensuring the reliability and accuracy of the classification method. A major voting algorithm was evaluated by measuring the validation metrics for each threshold and using the ROC-AUC (receiver operating characteristic area under the curve) metric. Binary labels were added to each patient: “Possibly Benign”, “Possibly Malignant”.

Results

From the original 170 cases, $n = 58$ miscellaneous lesions were removed from the current analysis. Additionally, $n = 4$ cases where VSI-B did not identify a mass seen on standard of care, $n = 2$ cases of DCIS without a sonographic

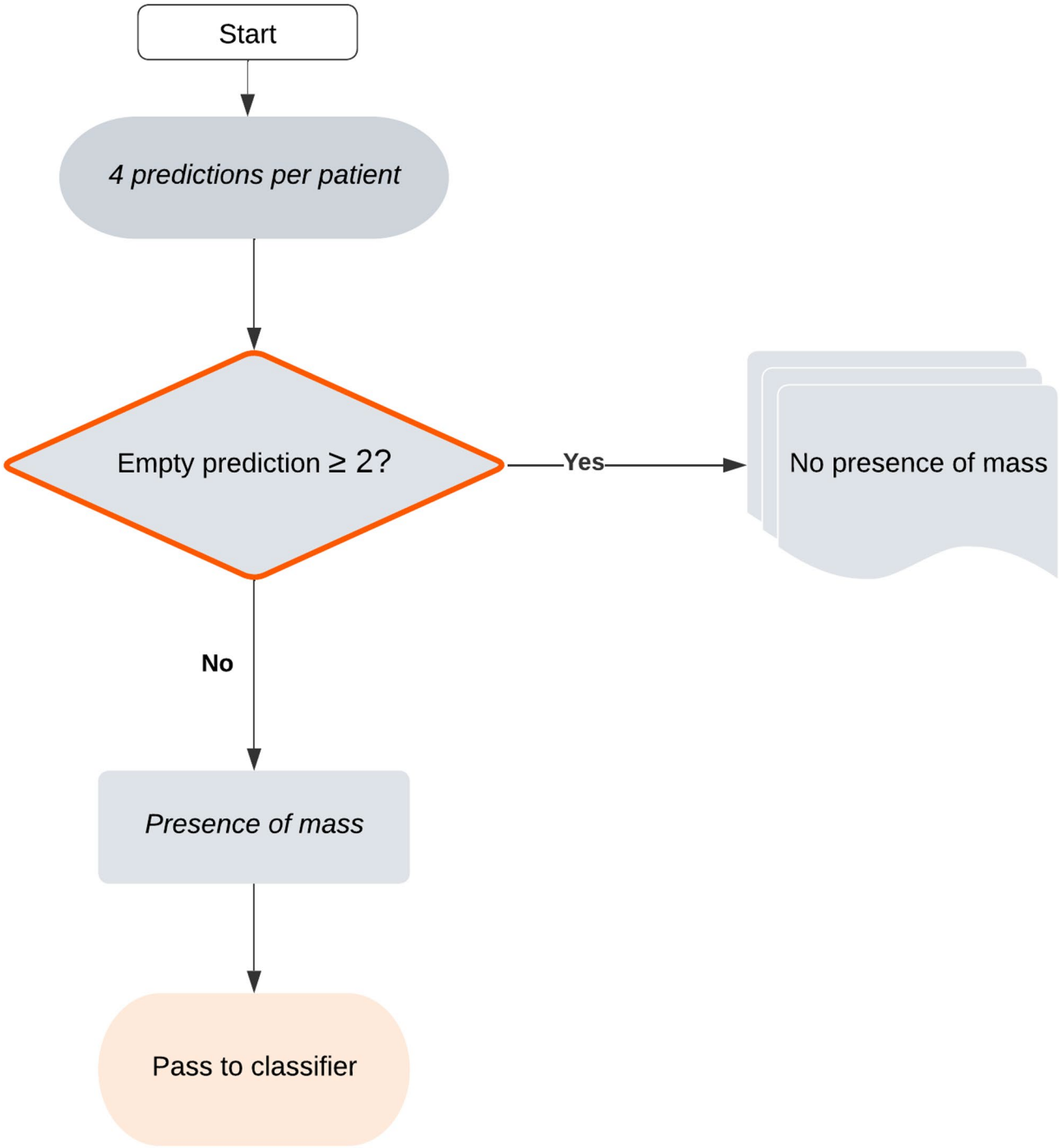

Fig. 8. Discrimination algorithm of segmentations before classifying breast US frames. Patients who had two or more videos with empty masks were considered as non-mass patients and discarded for the classification stage.

mass, and $n = 8$ cases of fibroadenomas or malignancy where there was difficulty in segmenting a mass lesion were removed from analysis. After exclusions, there were 98 total subjects, $n = 38$ cases no mass and $n = 60$ mass lesions. Table 1 provides a detailed breakdown of patients with confirmed diagnoses of cancer or non-cancer, including those with cysts and fibroadenomas.

Segmentation

Table 2 presents the results for the segmentation model. The Dice score improved from 53.7% (95% CI: 47.9–61.4%) to 62.9% (95% CI: 55.3–70.5%) after the 1D Gaussian window application with a threshold of 0.01%. Notably, the variance across folds was 0.71%. A paired t-test confirmed the statistical significance of this improvement ($p = 0.0216$).

Dice Coefficient	Dice (%) ± SD
Without post-processing	53.7 ± 39.1
With post-processing	62.9 ± 38.5

Table 2. Dice coefficient comparison for segmentation results using the post-processing technique.

Before classification, a clear improvement was observed with the post-processing in Fig. 9. Discrimination algorithm gave not only the possible mass/non-mass patient, but in case of a mass patients the maximum area mask to be analyzed (Fig. 10) with 4 more frames per video. The best result was obtained with the threshold of 2 or more unmasked videos to be considered non-mass patients. Similarly, the existence of 2 or more videos with masks to categorize patients with mass improved the evaluation metrics. Assessment in detecting breast lumps at patient level reached 95.0% sensitivity and 63.0% specificity. Considering only mass patients; Cancer and non-cancer detection had 100% sensitivity and 93.6% specificity (Table 3).

Classification

Classification models showed high sensitivity, accuracy, and specificity. Figure 11 shows that among the models selected, DenseNet reached 87.2% accuracy (95% CI: 76.5–93.5%), 100% sensitivity (95% CI: 78.5–100%), and 81.8% specificity (95% CI: 68.4–90.4%). Furthermore, the patient-level variance of correct classification across LOOCV folds was 11.3%. This DenseNet-based model had the highest overall metrics except for specificity, where Inception reached 87.9% specificity (95% CI: 75.4–94.5%). DenseNet, MobileNet, and Inception were the only architectures surpassing 80% accuracy and specificity. ResNet had the lowest performance with 61.7% accuracy (95% CI: 49.1–72.9%) and 50.0% sensitivity (95% CI: 26.8–73.2%), while Vgg16 had the lowest specificity, 60.6% (95% CI: 46.2–73.4%).

Major voting algorithm

Test patients from the LOOCV were evaluated with DenseNet architecture to assess the major voting algorithm for classification across different thresholds, as Figs. 12 and 13 depict. The threshold of 20% had the highest metrics. In comparison, the rest of thresholds tested (30%-50%), had lower results. The ROC-AUC of threshold of 20% was 0.91. These results indicated a superior performance of the DenseNet architecture at these thresholds.

Discussion

This work explored the potential of an automated system for breast lump diagnosis using VSI-B and AI. The integration of this fully automated system can serve as a valuable decision-support tool by providing clear patient management recommendations based on detected findings. For non-mass cases, the system suggests no findings of concern. When a benign mass is detected, it could recommend a follow-up in six months. In cases where a potential cancerous nodule is identified, the model does not directly diagnose cancer but rather advises referral to a hospital for further evaluation. This structured recommendation system helps complete the missing care pathway in LMICs by offering a streamlined approach to patient management, eliminating the need for acquisition and manual image selection by a specialist. By integrating pathology-based classifications, primary care assessments, and imaging findings, the model enhances diagnostic workflows, ensuring that patients receive timely and appropriate care. These findings suggest that combining 3D segmentation with 2D classification on VSI-B images may provide a robust and reproducible approach for breast cancer screening in resource-limited settings.

Prior work with the VSI-B protocol, include the use of a WATUNet, has demonstrated promising performance by applying a 2D breast cancer segmentation approach to US frames[28]. Results with WATUNet serves as an important benchmark for our research, offering a well-established reference point for VSI-based segmentation accuracy. Our previous study also assessed breast tumor segmentation comparison using VSI-B US protocol previously, where using a 2D multi-input attention U-Net achieved 72.45% of Dice coefficient[27]. However, both studies proposed 2D methods, that often overlook the temporal and volumetric information present in clinical US videos. Using videos for the segmentation task provides the incorporation of temporal information from multiple frames. This temporal attention approach leverages the greater volume of data in videos compared to single images, potentially improving segmentation accuracy.

The proposed approach was fully automated, from segmentation to classification, streamlining the diagnostic process and reducing the need for manual intervention. In contrast, S-Detect relies on a physician to identify a single "representative" frame. Restricting the analysis to a single frame may omit variations that appear elsewhere in the clip, potentially skewing diagnostic accuracy[14]. Although achieving relevant results in segmentation metrics, previous works such as obtained with WATUNet and Ochoa et al. focused solely on segmenting the videos and did not evaluate classification at the patient level[27,28]. Performance metrics, particularly sensitivity and specificity, are critical in assessing the clinical utility of diagnostic models. Our approach attains a sensitivity of 100% and a specificity of 83%—including 100% specificity for non-mass cases—indicating strong reliability in correctly identifying malignant lesions and minimizing false positives, even in a diverse patient population. S-Detect with VSI-B reported a sensitivity of 100% and a specificity of 86%, metrics that are like those achieved in this study. Achieving similar performance levels to a clinical benchmark is especially significant in healthcare and telemedicine, as it suggests that our fully automated, video-based pipeline could offer a viable alternative for remote or resource-limited settings.

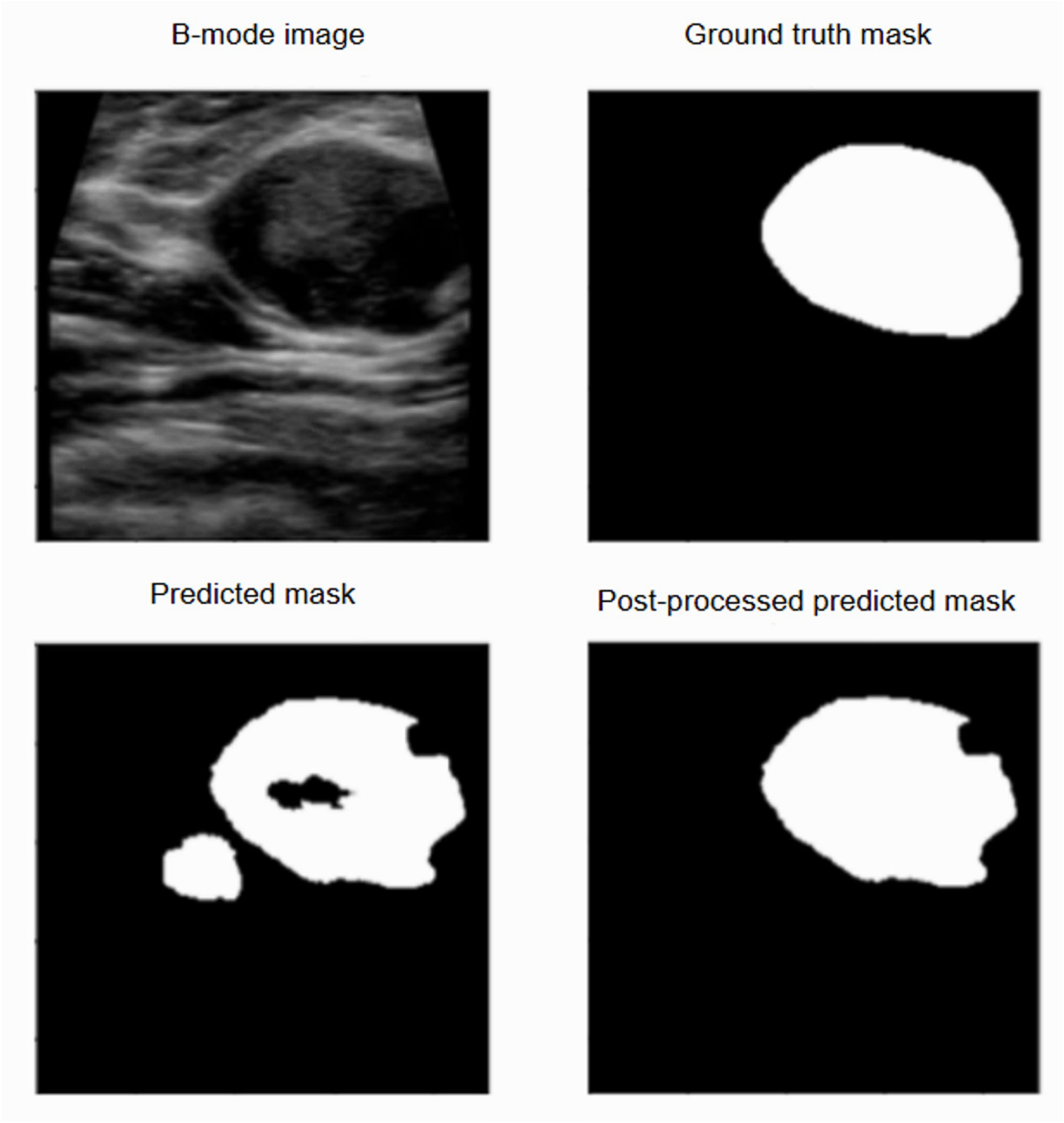

Fig. 9. Post-processing pipeline. Top-left: Obtained frame. Top-right: Ground truth mask. Bottom-left: Prediction generated by neural network. Bottom-right: Prediction post-processed.

This seemingly lower non-mass detection specificity is due to the inherent nature of the VSI-B protocol, which originally consider that all patients have a palpable breast lump, but this not necessarily means a sonographic correlation. Consequently, the initial dataset was predominantly composed of patients with a mass (either benign or malignant). The inclusion of palpable lumps without sonographic correlation (non-mass cases in the sweep) in the training dataset is a novel aspect of this study, allowing for the development of a more comprehensive diagnostic system that accounts for all three possibilities: the presence of a benign mass, a malignant mass, or no mass at all. Increasing the number on non-mass cases would improve the specificity value for this category.

In terms of classification with VSI-B protocol, an exploratory 2D classification approach was proposed using S-detect software, but the methodology focused on only a single frame per patient[14]. However, this selection of the frame was physician-dependent, leading to a manual process rather than an automated one. While this suits the process for this already trained system, it potentially omitted important temporal and volumetric details contained in the full US clip. Besides this, there was not a validation of the S-Detect system on non-mass

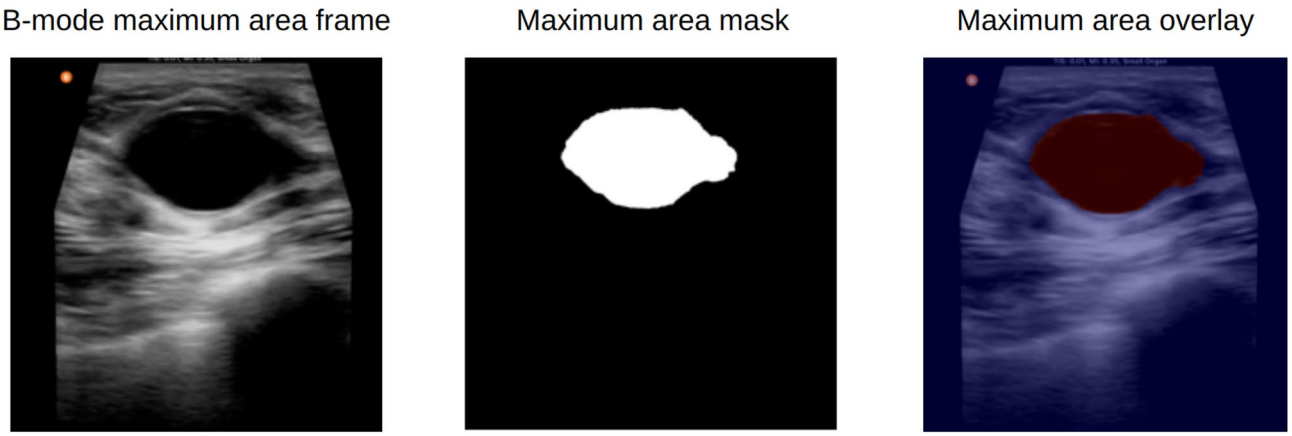

Fig. 10. Cropped Frame acquired with Butterfly IQ + with the maximum area segmented. Left: Original US image of a benign patient. Center: Mask predicted by Attention-U-Net 3D. Right: Overlay of the patient's US frame with larger area of the predicted tumor mask.

Detection	Specificity (%)	Sensitivity (%)
Mass/ non-mass	63.0	95.0
Cancer/ no cancer	93.6	100

Table 3. Model's specificity and sensitivity metrics at patient level for breast detection on mass patients versus non-mass patients, and cancer patients versus non-cancer patients. In the case of mass detection, the specificity of 63% might seem relatively low; however, this does not pose a critical issue within the diagnostic workflow. Since patients identified as "having a mass" proceed to the cancer classifier, all misclassified cases at this stage have been ultimately re-categorized as benign, which is beneficial as it minimizes the risk of missing malignant cases.

cases acquired with VSI-B protocol, leaving a gap in fully capturing the spectrum of breast abnormalities[14]. Recognizing this shortfall, we explicitly included non-mass patients, aiming to offer a more comprehensive approach that addresses real-world clinical variability.

Generalizability is a key requirement for AI in healthcare, which is why the present study employed a patient-wise cross-validation method based on our previous study[27], which ensured that the model is validated on unseen patient data, thereby reducing the risk of overfitting, and enhancing generalizability. On the other hand, WATUNet used random frames for train-test splitting, a simpler method that might reduce generalizability, and needed manual revision to prevent this issue and overfitting[28]. S-Detect system has an already trained model, so this methodology did not apply for that project[14].

Successful deployment of the complete pipeline requires careful consideration of deployment logistics, such as training and hardware requirements. The VSI-B acquisition protocol can be mastered in under two hours, consistently yielding high inter operator agreement among non-specialist personnel[8]. More complex protocols were successfully tested in low-resource regions such as obstetric and right upper quadrant, showing that is possible to deploy breast VSI training in a limited setting. To enable image capture, storage, transmission and processing in resource limited settings, we used a telemedicine device called the medical box, which features a simplified user interface and relies on CPU mode processing[39]. On this platform the end-to-end workflow completes in approximately four minutes per patient, demonstrating that the pipeline can be deployed cost effectively in low resource settings without specialized computational resources.

While high-end cart-based ultrasound systems generally deliver superior resolution, advanced imaging modalities (e.g., spectral Doppler, elastography) and greater depth penetration, enhancing AI-driven accuracy, POCUS handheld devices like the Butterfly iQ + trade some of these capabilities for portability and affordability[40]. Field evaluations have also noted that POCUS devices tend to have reduced imaging performance due to processing power requirements and occasional overheating during continuous use[40,41]; without affecting the clinical diagnosis. Despite these limitations, the Butterfly iQ + remains a viable solution in low-resource settings, and our medical box platform's ultrasound-agnostic design enables seamless adoption with high-end machines in better-resourced environments without modification of the end-to-end workflow.

In comparison with Automated Breast Ultrasound Systems (ABUS), our proposed VSI-B approach presents key advantages specifically tailored for low-resource settings. Recent clinical studies have reported sensitivity and specificity for ABUS of approximately 83% and 91%, respectively, underscoring its diagnostic efficacy as a complementary tool for breast cancer detection[42]. However, ABUS systems entail high costs (equipment prices can reach approximately $300,000) and demand specialized infrastructure with stable electricity supply, complicating their deployment in rural or resource-limited areas[43].

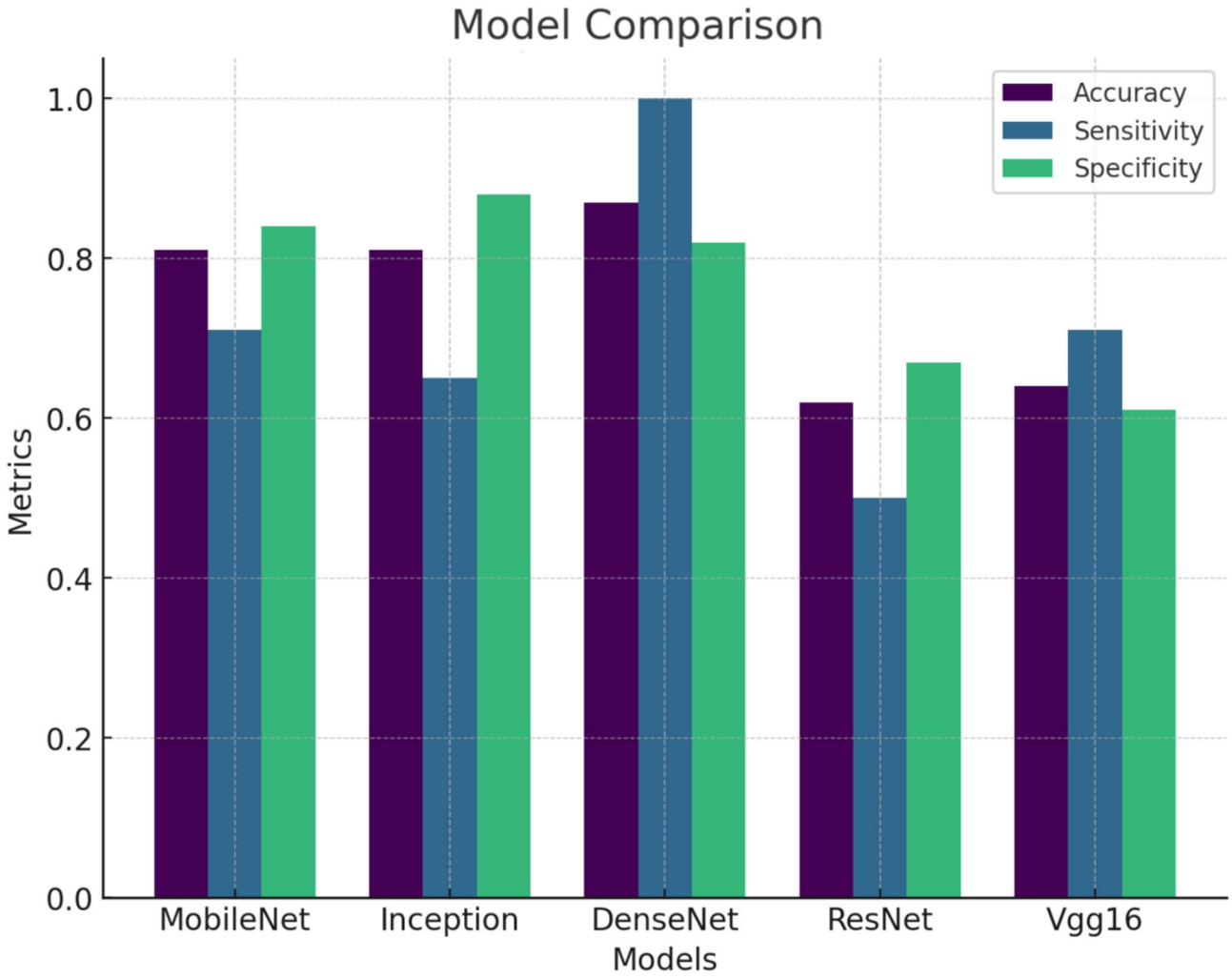

Fig. 11. Classification metrics for U-Net segmentation, comparing the 5 architectures, with a threshold of 20%.

Regarding future steps our study presents several limitations. First, the dataset was collected in a single center with a specific acquisition protocol and equipment, which may limit the representativeness of broader clinical populations. Second, while the model was validated on an independent test set, its generalizability to untrained operators or different ultrasound devices remains untested, which could impact real-world applicability. Third, the risk of automation errors, especially false negatives in low-quality images or atypical cases, necessitates that the AI system be used as an assistive tool rather than a standalone diagnostic method. Prospective multi-center studies and external validations are needed to address these limitations before clinical deployment. Increasing the volume of patients in the dataset is likely to improve the specificity and reliability of the AI, especially in difficult cases. In this study, we focus on cysts, fibroadenomas, and cancers but other types of mass lesions such as abscess, fat necrosis, skin lesions, lipomas and multiple lesions per patient would also benefit from analysis with AI. At the time of this study, we did not have enough of these other types of lesions to perform meaningful analysis. Thus, it is important to recognize that our current study will have some limitations in terms of generalizability that should be addressed in future work. Also, a larger number of ablation studies are necessary to calibrate and optimize the process of applying a threshold on the Gaussian mask, as it may prevent the detection and segmentation of masses smaller than 0.001% of the image area. Key frame selection and 3D model comparison can be proposed. While redundancies in the data, such as segmenting only 4 of the 8 clips, may contribute to precision by reducing noise and focusing on key frames, this can also lead to longer processing times. It is crucial to assess whether the additional frames are enhancing the model's accuracy or simply increasing computational load without significant gains in precision. Balancing redundancy with time efficiency will be key in refining the VSI-B-based system.

Conclusion

VSI-B provides a potentially low-cost, portable, and non-invasive method for breast imaging, which can be crucial in areas with limited access to medical imaging. The integration of AI with VSI-B could enhance the reach of VSI-B by automating the detection and classification process, thus reducing the dependency on highly

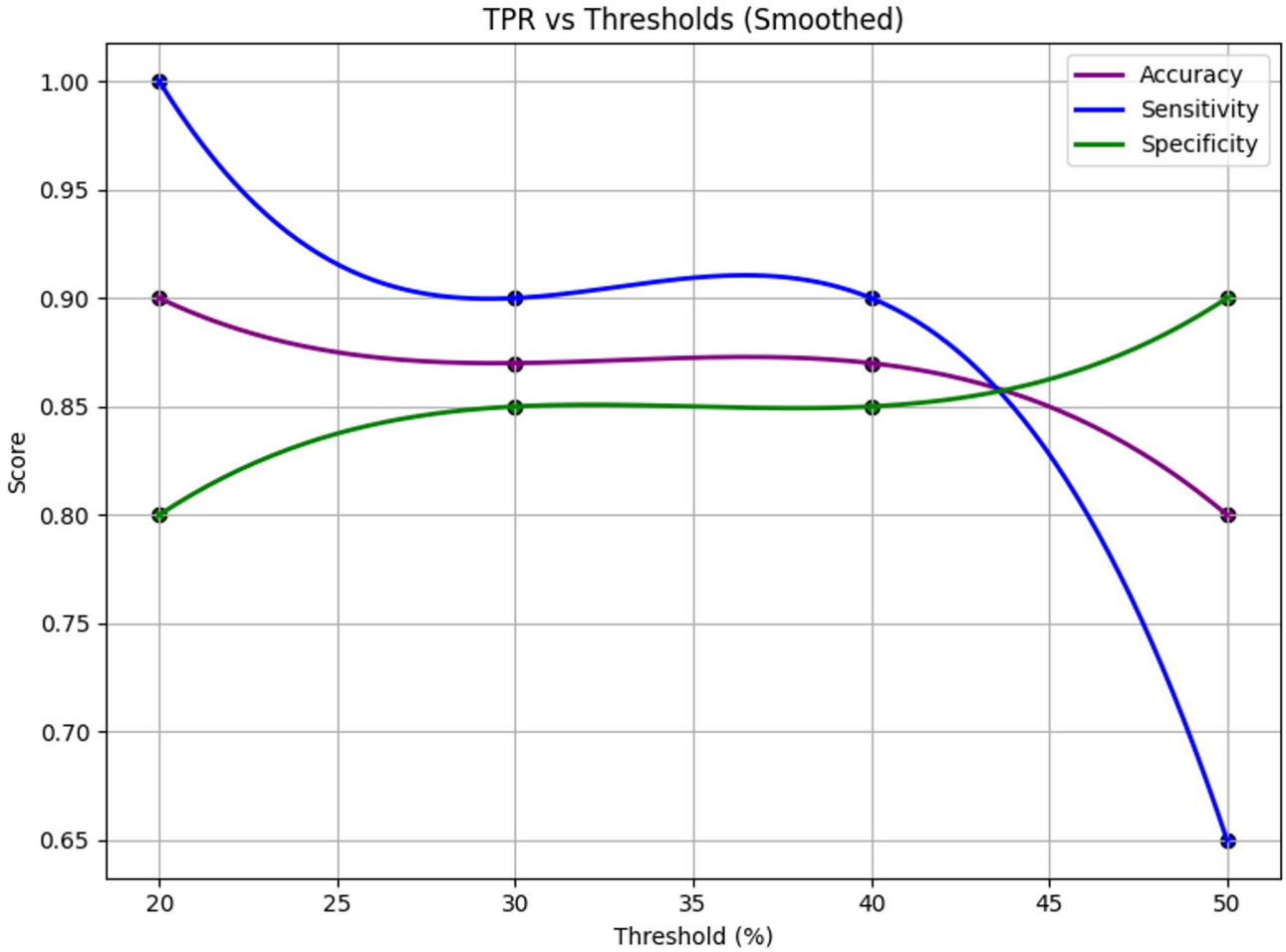

Fig. 12. DenseNet classification metrics across different thresholds. Threshold of 20% had the highest scores in accuracy and sensitivity.

trained radiologists. This model, with its high sensitivity and specificity for malignancy, supports early detection and accurate diagnosis of breast cancer, improving patient outcomes through timely and effective treatment. The combination of VSI-B and AI in this study demonstrates a promising approach to addressing the challenges of breast cancer diagnosis.

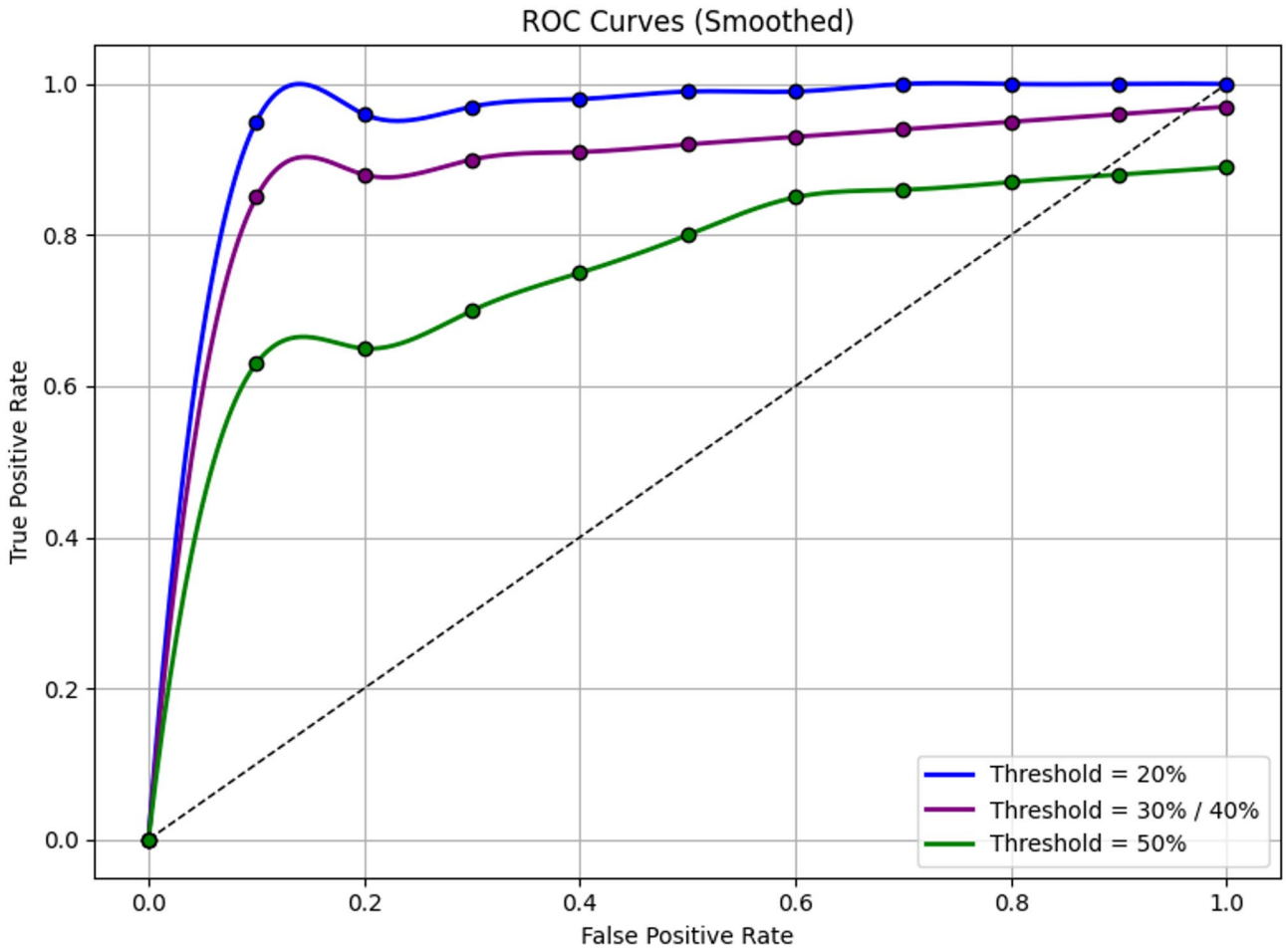

Fig. 13. ROC curves of the DenseNet across different thresholds. The best result is found with the threshold of 20% with an AUC of 0.91.

Data availability

The code used for this study is publicly available on GitHub at https://github.com/castanedalab/Breast-AI-model.The database used for the analysis in this study was provided by the University of Rochester and contains proprietary information. Due to the sensitive nature of the data and confidentiality agreements, the raw data cannot be made publicly available. However, anonymized and aggregated data that support the findings of this study is available from the corresponding author on reasonable request. Requests for data access should be directed to the corresponding author and will be reviewed to ensure compliance with confidentiality and intellectual property agreements.

Received: 13 March 2025; Accepted: 18 November 2025
Published online: 19 December 2025

References

1. Francies, F. Z., Hull, R., Khanyile, R. & Dlamini, Z. Breast cancer in low-middle income countries: Abnormality in splicing and lack of targeted treatment options. *Am. J. Cancer Res.* **10** (5), 1568–1591 (2020).
2. Ginsburg, O. et al. Breast cancer disparities among women in low- and middle-income countries. *Curr. Breast Cancer Rep.* **10** (3), 179–186 (2018).
3. Torre, L. A., Islami, F., Siegel, R. L., Ward, E. M. & Jemal, A. Global cancer in women: burden and Trends. cancer epidemiology, biomarkers & prevention: a publication of the American association for cancer research. *Cosponsored Am. Soc. Prev. Oncol.* **26** (4), 444–457 (2017).
4. Duggan, C. et al. The breast health global initiative 2018 global summit on improving breast healthcare through Resource-Stratified phased implementation: methods and overview. *Cancer* **126** (S10), 2339–2352 (2020).
5. Unger-Saldaña, K., Peláez-Ballestas, I. & Infante-Castañeda, C. Development and validation of a questionnaire to assess delay in treatment for breast cancer. *BMC Cancer*. **12** (1), 626 (2012).
6. Romanoff, A. et al. Association of previous clinical breast examination with reduced delays and Earlier-Stage breast cancer diagnosis among women in Peru. *JAMA Oncol.* **3** (11), 1563–1567 (2017).
7. Weiss, A. et al. Validation study of the American joint committee on cancer eighth edition prognostic stage compared with the anatomic stage in breast cancer. *JAMA Oncol.* **4** (2), 203–209 (2018).

8. Marini, T. J. et al. Breast ultrasound volume sweep imaging: A new horizon in expanding imaging access for breast cancer detection. *J. Ultrasound Med.* **42** (4), 817–832 (2023).
9. Erlick, M. et al. Assessment of a brief standardized obstetric ultrasound training program for individuals without prior ultrasound experience. *Ultrasound Q.* (2022).
10. Toscano, M. et al. Diagnosis of pregnancy complications using blind ultrasound sweeps performed by individuals without prior formal ultrasound training. *Obstet. Gynecol.* **141**(5). (2023).
11. Marini, T. J. et al. Lung ultrasound volume sweep imaging for pneumonia detection in rural areas: piloting training in rural Peru. *J. Clin. Imaging Sci.* **9**:35 .
12. Marini, T. J. et al. Lung ultrasound volume sweep imaging for respiratory illness: a new horizon in expanding imaging access. *BMJ open. Respiratory Res.* ;**8**(1). (2021).
13. Marini, T. J. et al. New ultrasound telediagnostic system for Low-Resource areas. *J. Ultrasound Med.* **40** (3), 583–595 (2021).
14. Marini, T. J. et al. No sonographer, no radiologist: assessing accuracy of artificial intelligence on breast ultrasound volume sweep imaging scans. *PLOS Digit. Health*. **1** (11), e0000148 (2022).
15. Arroyo, J. et al. No sonographer, no radiologist: New system for automatic prenatal detection of fetal biometry, fetal presentation, and placental location. *PloS One*. **17** (2), e0262107 (2022).
16. Marini, T. J. et al. Sustainable volume sweep imaging lung Teleultrasound in Peru: Public health perspectives from a new frontier in expanding access to imaging. *Front. Health Serv.* **3**, 1002208 (2023).
17. Toscano, M. et al. Testing telediagnostic obstetric ultrasound in peru: A new horizon in expanding access to prenatal ultrasound. *BMC Pregnancy Childbirth*. **21** (1), 328 (2021).
18. Marini, T. J. et al. Testing telediagnostic right upper quadrant abdominal ultrasound in peru: A new horizon in expanding access to imaging in rural and underserved areas. *PloS One*. **16** (8), e0255919 (2021).
19. Marini, T. J. et al. Testing telediagnostic thyroid ultrasound in peru: a new horizon in expanding access to imaging in rural and underserved areas. *J. Endocrinol. Investig.* (2021).
20. Marini, T. J. et al. Volume sweep imaging lung Teleultrasound for detection of COVID-19 in peru: A multicentre pilot study. **12**(10):e061332. (2022).
21. Zhou, L. Q. et al. Lymph node metastasis prediction from primary breast cancer US images using deep learning. *Radiology* **294** (1), 19–28 (2020).
22. Browne, J. L. et al. AI: can it make a difference to the predictive value of ultrasound breast biopsy? *Diagnostics (Basel)* ;**13**(4). (2023).
23. Khaledyan, D., Marini, T. J., Baran, M., O'Connell, T. & Parker, A. Enhancing breast ultrasound segmentation through fine-tuning and optimization techniques: Sharp attention UNet. *PloS One*. **18** (12), e0289195 (2023).
24. Al-Dhabyani, W., Gomaa, M., Khaled, H. & Fahmy, A. Dataset of breast ultrasound images. *Data Brief.* **28**, 104863 (2020).
25. Arroyo, J. et al. *Automatic Fetal Presentation Diagnosis from Ultrasound Images for Rural Zones: Head Location as an* (SPIE, 2021).
26. Saavedra, A. C. et al. (eds) Automatic ultrasound assessment of placenta previa during the third trimester for rural areas. 2020 IEEE International Ultrasonics Symposium (IUS). (2020).
27. Ochoa, E. J. et al. editors (eds) itors. A comparison between deep learning architectures for the assessment of breast tumor segmentation using VSI ultrasound protocol. *46th Annual Int. Conf. IEEE Eng. Med. Biology Soc. (EMBC)* **2024** IEEE (2024).
28. Khaledyan, D. et al. WATUNet: a deep neural network for segmentation of volumetric sweep imaging ultrasound. *Mach. Learning: Sci. Technol.* **5** (1), 015042 (2024).
29. Schenk, A., Prause, G. & Peitgen, H-O. (eds) Efficient Semiautomatic Segmentation of 3D Objects in Medical Images2000; Berlin: Springer.
30. Raya, S. P. & Udupa, J. K. Shape-based interpolation of multidimensional objects. *IEEE Trans. Med. Imaging.* **9** (1), 32–42 (1990).
31. Islam, M. et al. (eds) Brain Tumor Segmentation and Survival Prediction Using 3D Attention UNet2020; Cham: Springer International Publishing.
32. Falcon, W. The PyTorch Lightning Team, 2023. Pytorch lightning.
33. He, K., Zhang, X., Ren, S. & Sun, J. (eds) Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2016).
34. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (eds) Densely connected convolutional networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2017).
35. Szegedy, C. et al. (eds) Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition; (2015).
36. Howard, A. G. et al. Mobilenets: efficient convolutional neural networks for mobile vision applications. *ArXiv Preprint arXiv* :170404861. (2017).
37. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint arXiv* :14091556. (2014).
38. Tang, Y-X. et al. Automated abnormality classification of chest radiographs using deep convolutional neural networks. *NPJ Digit. Med.* **3** (1), 70 (2020).
39. Marini, T. J. et al. New ultrasound telediagnostic system for low-resource areas: Pilot results from Peru. *J. Ultrasound Med.* **40** (3), 583–595 (2021).
40. Salimi, N. et al. Ultrasound image quality comparison between a handheld ultrasound transducer and mid-range ultrasound machine. *POCUS J.* **7** (1), 154 (2022).
41. Burleson, S. L. et al. Evaluation of a novel handheld point-of-care ultrasound device in an African emergency department. *Ultrasound J.* **12** (1), 53 (2020).
42. Dan, Q., Zheng, T., Liu, L., Sun, D. & Chen, Y. Ultrasound for breast cancer screening in resource-limited settings: current practice and future directions. *Cancers* **15** (7), 2112 (2023).
43. Company, G. E. *Commercial Reimbursement and Utilization for Invenia ABUS (Automated Breast Ultrasound)* (General Electric Company, 2019).

Acknowledgements

This work was funded by the Dirección de Fomento de la Investigación from Pontificia Universidad Católica del Perú, through grant CAP 2023-F-0020, project PI1094.

Author contributions

EJO, LCR, SER, and GAG were involved in the AI development, study design, data analysis and manuscript writing. TJM and KJP were involved in the AI development, study design, data acquisition, data analysis, and manuscript writing. YZ, GB, JK, and SM were involved the data acquisition, image segmentation, data analysis, and manuscript writing. AD and AW were involved in the study design, data interpretation, and manuscript production. BC oversaw the entire study and was involved in data analysis, AI development, and manuscript production.

Funding
This work was partially funded by the Dirección de Fomento de la Investigación from Pontificia Universidad Católica del Perú, through grant CAP 2023-F-0020, project PI1094.

Declarations

Competing interests
The authors declare no competing interests.

Conflict of interest
The authors have no conflict of interest.

Ethics and use of human participants statement
This study was approved by the University of Rochester Research Subjects and Review Board. (Study 00005262). Informed consent was obtained from every participant. The acquisition process adhered to all relevant guidelines and regulations established by the University of Rochester and was conducted in accordance with the principles outlined in the Declaration of Helsinki. To ensure compliance with HIPAA guidelines, all video clips underwent a thorough anonymization process.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-29554-6.

Correspondence and requests for materials should be addressed to B.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.