

Design and Evaluation of a Breast Ultrasound Research Workspace Using VGG16 and Grad-CAM

Shivam Gupta¹, Saksham Singh², Mohd Sakib³, Shweta Roy⁴

¹Department of Computer Science
and Engineering

ABES Engineering College

Ghaziabad, India

sg1544693@gmail.com

²Department of Computer Science
and Engineering

ABES Engineering College

Ghaziabad, India

sakshamsingh0402@gmail.com

³Department of Computer Science
and Engineering

ABES Engineering College

Ghaziabad, India

pkarm460@gmail.com

⁴Department of Computer Science
and Engineering

ABES Engineering College

Ghaziabad, India

sguddr@gmail.com

CORRESPONDING AUTHOR

Shivam Gupta

Department of Computer Science
and Engineering

ABES Engineering College

Ghaziabad, India

Email;sg1544693@gmail.com

ABSTRACT Background

This audit followed a breast-ultrasound case through a Flask research workspace and checked the displayed functions against code, retained artifacts and final review records. Separate results were reproduced for multiple views of one lesion, together with original-resolution Grad-CAM, versioned annotations and reviewer material. Binary scoring used the saved frozen VGG16 classifier. The approved research collection contained 78 images; after duplicate, patient and lineage exclusions, the inspected query returned 12 matches. For segmentation, the residual U-Net reached a lesion-only Dice score of 0.2534 and failed to detect 12 of 20 lesions. Confirmed prediction supplied a single-label answer in only one of 31 cases. The input-screening experiment covered an insufficient range of negative sources. Segmentation, abstention and input screening consequently remain disabled, while the reproducible case, explanation, review and retrieval paths remain available as research functions.

Keywords—breast ultrasound, VGG16, Grad-CAM, medical image workspace, segmentation, model evaluation

How to cite this article: Shivam Gupta¹Saksham Singh²Mohd Sakib³Shweta Roy⁴. Design and Evaluation of a Breast Ultrasound Research Workspace. Int J Drug Deliv Technol. 2026;16(86s):8-13. DOI: 10.25258/ijddt.16.86s.2.

I. INTRODUCTION

Multiple non-interchangeable views of the same lesion may be obtained during a breast ultrasound examination. Each picture could seem different depending on the probe's location, the acquisition parameters, and the operator's method. Hence, rather than consolidating the visit into a single score, the workspace maintains each view, prediction, and explanation as an independent record.

Originally built as a Flask classifier, the program has since evolved into a case-based research environment. Users can browse results per image, annotate images, request reviews, and search an approved reference collection. A visit can hold numerous images. Segmentation, input filtering, and prediction abstention are also tested in separate versions of the repository.

They are not considered interface-ready just because they are in the codebase.

The maintained route from an input picture to the output displayed to a reviewer determines the project's reproducibility. By following that route, we were able to complete the preprocessing, retained model, output generation, and interface record keeping. The isolated app was launched using its fake account, and its pages were compared to previously stored verification materials. The same procedure was used to analyze the final artefacts obtained from the supplementary experiments. There is documentation of both successful and unsuccessful checks.

In order to determine if the saved build can replicate the evidence it provided, the audit checks to see if each experimental component met its release criteria. In

response to the first question, runtime checks and retained artefacts are utilised. Instead of training summaries, final review files address the second question. The application is identified as a research system rather than a diagnostic device, and this boundary keeps unfinished models outside the deployment claim.

II. RELATED WORK

The experimental selections in this workspace are guided by earlier investigations on breast ultrasonography. Classification outcomes are task and evaluation design dependent, as demonstrated by Yap et al. [1]. When a pixel-level mask, instead of a class score, is the desired output, U-Net comes into play [2]. While BUSI [3] and BUS-UCLM [4] make it possible to conduct repeated studies, such as our segmentation audit, the datasets vary depending on the type of collection. Neither source's performance can prove transferability to other centers or scanners.

Although they appear to be visually identical, Grad-CAM and segmentation actually address distinct concerns. Instead of using a measured lesion border as its coarse map, Grad-CAM makes use of gradients and convolutional features to pinpoint areas that had an impact on a class judgement [5]. That boundary is directly estimated via segmentation. For this reason, we keep the Grad-CAM view and the segmentation mask in separate files and assess their respective outputs independently.

Just because a label is accurate doesn't mean the score is reliable. Dataset changes can affect uncertainty estimates [7], and calibration is all about the relationship between scores and outcomes [6]. We conducted our abstention experiment because conformal prediction generates prediction sets based on specified assumptions [8]. Since a set with both labels can be accurate even without providing a class decision, we look at both label coverage and answered coverage.

Reporting guidelines are likewise followed by the evidence gate. In contrast to DECIDE-AI, which incorporates workflow and human aspects, CLAIM places an emphasis on provenance, reference standards, evaluation splits, and subgroup findings [10]. The hidden stratification technique alerts users that a total score might mask subgroup failure [11], while the model cards document the intended use, evaluation parameters, and constraints [12]. Thus, aggregate values are shown alongside malignant-lesion Dice and answered coverage, but experimental features that lack supporting evidence are blocked.

III. SYSTEM DESIGN

A. Case and image records

Some of the tools used to build the app include Flask, SQLAlchemy, OpenCV, and TensorFlow/Keras. Make sure the pictures of the visit actually show the same lesion before proceeding. Irreversible image records are created with every submitted view. The file is encrypted, validated for size and allowed types, and then hashed to prevent the owner's workspace from accessing an identical copy. Two distinct storage pillars support the public research library and private uploads.

Leasing the processing work ensures its execution. An older state cannot be overwritten by a recovered or delayed worker because a worker can only write a result while it still owns the current lease. Everything from the initial image route and model fingerprint to the processing phase, presented scores, raw output, and Grad-CAM assets are part of the recorded information. The app will show you both outcomes if two opinions are contradictory. In addition, it cannot conceal disagreements by averaging results or compute a diagnosis at the visit level.

In the report that is produced, this division persists. Even after adding an annotation, the report clearly shows which model fingerprint was used for the image. According to it, nobody has checked the output, the sigmoid values aren't calibrated probabilities, and the total score from connected views isn't certified. Avoiding the possibility of an exported report losing the associated conditions is achieved by generating these statements from the case record instead of manually adding them to a screenshot.

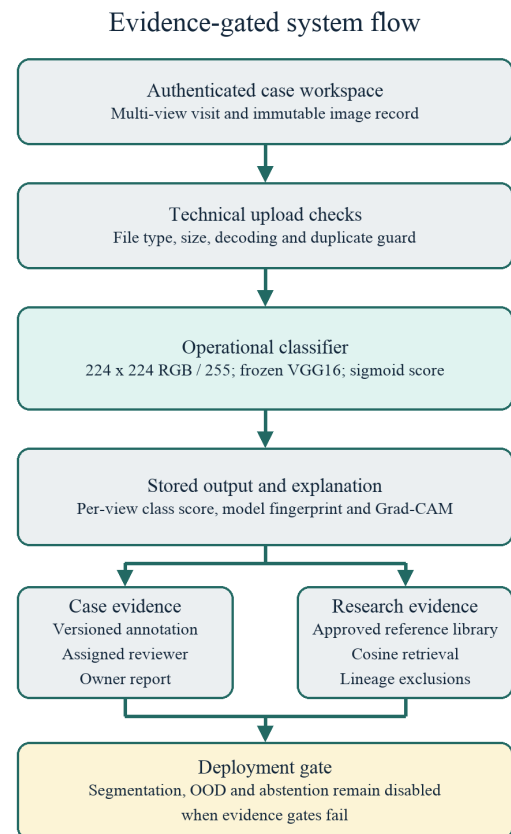


Fig. 1. Processing and review paths implemented in the current workspace.

B. Classification and Grad-CAM

Based on a frozen VGG16 convolutional foundation, the maintained classifier is a Keras Sequential model [13]. A single sigmoid output follows global average pooling, then thick layers with 512 and 256 units, and dropout. There are 15,108,929 parameters in the loaded artefact, with 394,241 of them indicated as trainable in the stored configuration. Prior to inference, images undergo resizing to 224 x 224 pixel refinement, RGB conversion, and division by 255.

The interface converts the sigmoid values into a displayed malignant score of 100s and a benign score of 100(1-s). It also shows the distance from the 0.5 decision threshold. The displayed values are not described as calibrated probabilities. Grad-CAM is calculated for the malignant output using the last available convolutional layer. The heatmap is resized to the dimensions of the uploaded raster before the coloured overlay is saved. The report and browser labels state that this output is an explanation rather than a lesion boundary.

C. Annotation and assigned review

Users are able to save polygons, notes, and assessments separately from the classifier record. With each save, a new version is created while keeping the previous image space's coordinates. In the absence of a scale, all measurements will be expressed in pixels. There is a distinct procedure for reviewer access: an operator assigns a specific case and authorises reviewer capacity. The owner's export may include the finalised version of the reviewer's notes, but the draft notes remain private. Access is controlled by the program; it does not validate a person's credentials.

When saving an annotation points, the browser returns them to their original source-image coordinates because of the ability to zoom and resize images. Since two pictures may depict the same lesion, it is not possible to transfer a contour from one view to the other; this is a crucial consideration in a multi-view scenario. Additionally, the version history safeguards against the reviewer contour being confused with a patient note. Despite having distinct authors and states, the two records might be shown side by side for comparison purposes.

D. Research image retrieval

Only documents that have been approved for research are available in the search library. The source, license, label provenance, patient key, and classifier fingerprint are all stored in each reference. Prior to ranking by cosine similarity, the code removes the query hash, exact duplicates, photos from the same patient, and related source lineages when a query is submitted. Returns no more than 12 samples. Rather than diagnostic probability, the page explains the value as feature similarity.



Fig. 2. Research retrieval page showing source and label information for approved references.

E. Reports and failure states

Assembling a case export is done from saved records instead of the current browser screen. Since this is the case, the report may consistently list the model's fingerprint, processing time, image result, and finalised human material. Notes written by reviewers and notes written by individuals retain their separate folders. The export indicates that the provider is unavailable rather than inserting a blank segmentation panel, which could be interpreted as a negative result.

The worker records phase changes so that a failed upload, model error or explanation error can be distinguished from a completed prediction. A retry does not erase the earlier failure state until the new lease has been accepted. These details do not improve model accuracy, but they make an erroneous or incomplete run easier to identify. For a research application, preserving this history is preferable to showing only the last visible label.

IV. DATA AND EXPERIMENTAL METHODS

A. Repository and data audit

About evenly split between benign and malignant directories, the local classification folders hold 900 validation files and 8,116 training files. For these files, we were unable to locate a patient-level manifest, a comprehensive source license, or a trustworthy membership record. Accordingly, the accuracy of their folder-based validation is not cited in the article. Images from the same patient or acquisition sequence appearing on both sides would make it difficult to interpret a large number from this split.

The BUS-UCLM release contains 683 images from 38 studies, together with RGB masks and labels derived from publisher metadata [4]. The repository audit compared publisher checksums, decoded the files and checked image-mask geometry. Four decoded duplicates and 13 geometry mismatches were removed. The second segmentation run also excluded the patients used for visual inspection during the rejected first run. The remaining split is shown in Table I.

Dataset labels were read from the supplied INFO.csv file instead of being inferred from file names. The resulting manifest records the patient group, source image, mask, split and exclusion reason. This was necessary because a file-level random split would place closely related views on both sides of the evaluation. The manifest also gives the final review a stable description of the cohort. Re-running the evaluation against a different list of images would no longer satisfy the stored review even if the model file name stayed the same.

Table I. Patient-Grouped Split Used For Segmentation

Split	Images	Benign	Malignant	Normal
Train	414	21	108	247
Validation	121	6	27	80
Final	51	5	12	31

B. Segmentation experiment

The second experiment used a residual U-Net with a 128 by 128 input and 16 base filters. Its skip-connected encoder-decoder follows the U-Net pattern, while the convolutional blocks use residual connections associated with residual learning [2], [14]. Letterbox resizing preserved the aspect ratio, bilinear interpolation was used for images and nearest-neighbour interpolation was used for masks. Training combined binary cross-entropy with Dice loss. Lesion and normal samples were balanced during training, while horizontal flipping was limited to the training split. The optimizer was Adam with a learning rate of 0.0003, batch size 8 and seed 42.

Training stopped after 20 of the requested 35 epochs because the validation loss no longer improved within the configured patience. The final mask threshold was not chosen on the final cohort. A frozen search on the validation cohort considered lesion Dice, the fraction of completely missed lesions and false positives on normal images. The selected threshold was 0.15. We report lesion-only Dice and intersection-over-union, lesion-detection sensitivity, complete misses, false positives on normal images and results for benign and malignant lesions separately.

The deployment review defined its acceptance criteria before reading the final output. Lesion-only Dice had to reach 0.60, lesion-only IoU 0.45 and lesion-detection sensitivity 0.90. Completely missed lesions could account for no more than 10 percent of lesion images, and mean Dice had to reach 0.50 in both the benign and malignant groups. Normal-image behaviour was examined separately so that correct empty masks could not compensate for missed lesions. The review also checked that the weight fingerprint and preprocessing description matched the artifact loaded by the provider

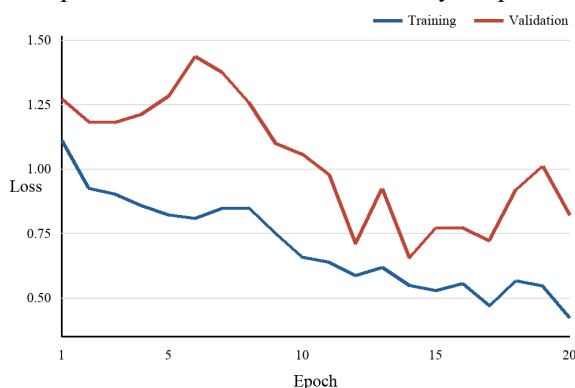


Fig. 3. Training and validation loss stored for the 20 completed residual U-Net epochs.

C. Input screening and abstention

The input-screening experiment used a 28 by 28 logistic source discriminator. BUS-UCLM images formed the positive class, while CT and chest radiograph images from MedMNIST v2 formed the available negative classes [15]. This was a feasibility test, not a comprehensive out-of-distribution study. It did not include screenshots, natural photographs, mammography, MRI, ultrasound from another organ or images containing vendor overlays.

The initial classifier remained unaltered in the abstaining experiment. There were a total of 31 photos utilised for

the final evaluation, 16 for temperature fitting, and 25 for conformal calibration. Despite a small patient population and certain associated perspectives, patient groups were distinct throughout these phases. Coverage of labels and answers were both recorded. An answered case was a collection of predictions with only one label, while an unlabelled set was considered to be an abstention.

D. Reproduction checks

We opened the isolated demonstration in a browser using the synthetic account retained with the project. The check covered visit creation, a two-view case, separate image results, the Grad-CAM page and research retrieval. Screenshots used in this paper were captured from that running instance. We also loaded the model artifact directly and read its layer structure and parameter counts.

The automated test suite was executed again throughout the audit. A total of 29 tests were run, with 19 passing and 19 failing. One module import was unsuccessful because of an error. The ML-readiness route was typically registered after a Flask blueprint was already in use, which led to the failures. The article does not reiterate a previous claim that the complete suite passes since this regression affects the present software claim. No model was retrained while preparing this paper. The numerical results were read from final artifacts whose hashes match the associated review files. This avoids an easy but serious mistake: changing a threshold or selecting another checkpoint after seeing the final cohort. The browser checks and screenshots test a different question. They show that the retained application can display the stored result and its limitations; they do not add samples to the model evaluation.

E. Interpretation rules

The numerical checks were interpreted against the task the feature would perform in the interface. Segmentation required acceptable lesion overlap and few complete misses; correct empty masks were reported but could not replace lesion performance. Input screening required coverage of realistic unsupported sources, not only a high area under the curve on two consecutive negative datasets. Abstention required the procedure to answer a useful share of cases as well as include the correct label. These rules were fixed in the review files before the final status was assigned.

V. RESULTS AND DISCUSSION

A. Working application paths

To validate, the live demo replicated the two saved raw classifier outputs. It was 856 by 606 pixels for both the heatmaps and the overlays, much like the source photos. There were 78 vetted citations in the library's active classifier fingerprint. Twelve results remained after excluding duplicates, patients, and lineages from the examined query. Repeatability of the application's path and its provenance records are supported by these checks. Clinical accuracy cannot be measured by these. Separation of results was also verified on the multi-view page. Each card kept its unique score, model fingerprint, and link to the inspection site, yet both photos stayed viewable under one visit. User annotations, the Grad-

CAM view, the source image, and reviewer materials could all be accessed without affecting the model's output. Although this is only a little improvement to the user interface, it fixes a prevalent reporting problem where human interpretation is rewritten as though it were the initial algorithm result.

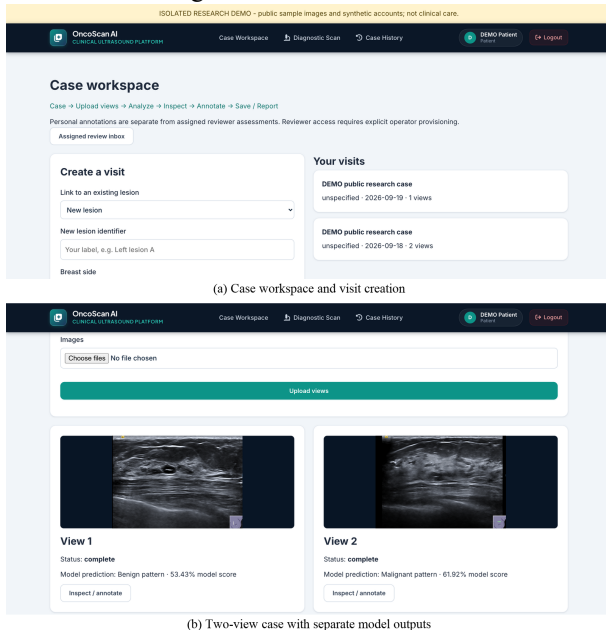


Fig. 4. Fresh demo views: visit creation and a case containing two ultrasound images.

There were 31 photos from 5 individuals in a feasibility set that solely included exterior lesions. Eleven pictures (or 35.48% of the total) had the unchanging classifier right. Rather than taking this as an indication of correctness, we view it as a caution. Several photos of the same patient are correlated, and it is not possible to reconstitute the original training membership adequately to demonstrate independence. Language that portrays the score as reliable across external data is nevertheless rejected by the result.

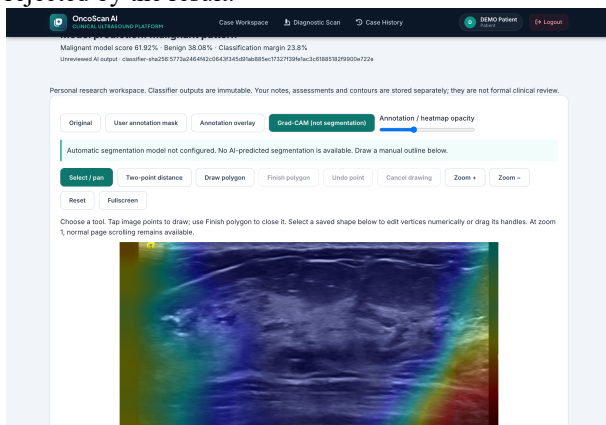


Fig. 5. Model output and original-resolution Grad-CAM displayed by the isolated demo.

B. Classifier error pattern

Three valid malignant predictions, five missed ones, eight accurate benign predictions, and fifteen false positives were included in the 31-image cohort when the stored threshold was set at 0.5. Thus, a slight bias against one class cannot account for the mistake. Due to the fact that each patient in the cohort has many views, the count

fluctuates within patient groups, and both false positives and false negatives might happen. The counts for transparency at the picture level are recorded in Table II, but no confidence range is given since a consistent population estimate cannot be supported by five single patients.

Table II. Image-Level Confusion Counts In The External Feasibility Set

Reference	Predicted benign	Predicted malignant
Benign	8	15
Malignant	5	3

C. Segmentation result

The residual U-Net obtained an all-image mean Dice score of 0.7072 on the final cohort. That number looks favourable until the cohort composition is examined. Thirty-one of the 51 images were normal, and both the reference and predicted masks were empty for all 31. Empty agreement contributed a Dice score of one to each of those images.

On the 20 images that contained a lesion, mean Dice fell to 0.2534 and mean IoU to 0.2010. Only eight lesions were detected, giving a sensitivity of 0.40, and 12 lesions were missed completely. Mean Dice was 0.4223 for the 12 benign lesions. All eight malignant lesions received empty predictions, so their mean Dice was zero. Six of the seven quantitative conditions in the deployment review failed.

The chosen threshold explains part of the pattern but does not rescue the model. At 0.15, the validation search favoured fewer normal-image false positives while retaining some lesion response. On the final cohort it produced correct empty masks for every normal image, yet this apparent specificity came with poor lesion sensitivity. Lowering the threshold after observing the final cohort would invalidate the evaluation protocol. A new threshold would require another validation decision and a new untouched test cohort.

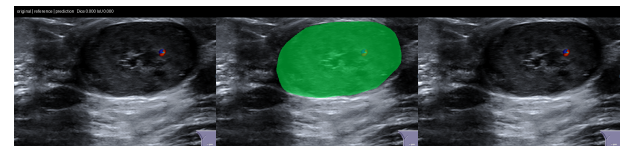


Fig. 6. Final-cohort malignant example: source image, reference mask and empty prediction.

D. Screening and abstention results

While 12 out of 150 CT scans were accepted by the input-screening model, none of the 149 chest radiographs were. All eighty BUS-UCLM pictures were accepted. Although there was one positive ultrasound source and two negative categories, the computed AUROC was 0.9969. It doesn't specify which picture formats are really supported by browsers. Accordingly, input screening was left disabled throughout the review. Every one of the thirty-one last instances had the right label in the conformal prediction sets. There was just one set that had a single label. The remaining 30 had both designations, resulting in a 96.77 percent abstention rate and a 3.23% replied coverage. As an interactive decision-aid, it is useless, but safe in the limited sense that the technique rarely commits to an answer. The

configured gate additionally kept abstention disabled because it required at least 60% replied coverage.

Table III. Final Results and Current Application Status

Component	Observed result	Status
Classifier	11/31 external lesion images correct	Research output
Segmentation	Lesion Dice 0.2534; sensitivity 0.40	Disabled
Input screening	12/150 CT accepted; narrow source coverage	Disabled
Abstention	1/31 cases answered	Disabled
Retrieval	78 indexed; 12 filtered results	Enabled

E. What the aggregate scores conceal

One aggregate metric might be deceiving, as the segmentation result demonstrates. Even though it fails to segregate instances that require it, a model that forecasts an empty mask can do well on a dataset where normal images predominate. A second issue is shown by the label-specific results: no malignant lesion was detected. You wouldn't know about either of the failures if you just reported the total Dice score.

The uncertainty experiments show the same tendency. Although one label coverage is mathematically desirable, it was accomplished by providing both labels for nearly all images. Also, commonly used browser inputs were not examined by the source discriminator, but the small evaluation sets were effectively separated. Each product claim is still unsupported by the experiments in the repository, even when the experiments themselves are valid.

F. Software and workflow implications

To overcome this discrepancy, the app asks for a review record with a fingerprint that corresponds to the model artefact. The segmentation provider will not allow normal construction if the review is either not present, refused, or linked to another weight file. For input screening, a comparable fail-closed rule is defined. This is better than a caution in the project documentation alone as a modification to the interface in the future cannot covertly reveal the rejected model.

The release decision is also affected by the present test regression. Although the browser demo and saved experimental metrics are both usable, the whole automated suite does not pass. Prior to proceeding with the release, the route registration needs to be fixed. Then, instead of inheriting the application initialisation order from an earlier import, run the suite in separate processes to test it.

There is also the matter of upkeep to consider. It is not possible to establish that the surrounding preprocessing code is unchanged using the fingerprint check, but it does prevent replacing a model file without a matching review. Thus, in addition to the weight hash, the configuration keeps track of picture size, colour handling, normalisation, and threshold. This whole inference contract should be tested with fixed input-output fixtures in a future version. Even if the weights of the neural network don't change, that would still detect a preprocessing modification.

VI. LIMITATIONS

A patient-level training manifest could not be linked to the available classifier folders. The diagnostic sensitivity, specificity, and calibration cannot be established by the separate classifier check because it only used photos from five cases. Information from a single center and scanner is provided by BUS-UCLM. Despite not being included in the v2 development process, the final v2 patients were part of the failed v1 training run. So, instead of being a separate external validation set, the last cohort is an internal holdout.

Technical issues, not editorial ones, cause some interface constraints. While reviewer authorization does grant access to the application, it does not validate the reviewer's qualifications. Unless a picture scale is accessible, measurements recorded in pixels do not have any physical significance. Even though a comparable reference image could be useful for inspection, it can't tell you what a new case is. You can't utilise Grad-CAM to validate a cohort since it can divert attention away from the lesion.

All of our privacy assessments were limited to the local study environment. The demo makes use of public BUS-UCLM photos and synthetic accounts, while the main workspace maintains database and media roots independently. Although that segmentation is developmentally appropriate, it does not constitute evidence of conformity with a clinical data-governance scheme. For production use, you'll require an independent security assessment, written retention rules, access reviews, incident handling, consent, or similar legal justification. On their own, the login and case assignment pages cannot guarantee any of these things.

Having a licensed classifier dataset that is organised by patient, a reference standard that is documented, and a truly independent test cohort would all be necessary for a more robust review. To investigate differences in scanner performance and acquisition methods, data from multiple centers would be required. Screening inputs should incorporate actual browser errors and external ultrasound pictures. The last assessment cannot happen without first planning the calibration and subgroup analysis. Predetermined acceptance criteria in line with CLAIM and DECIDE-AI should likewise be used in any subsequent clinical workflow research [9], [10]. It is not sufficient to infer hidden failure groups from average scores; instead, direct research is required [11].

VII. CONCLUSION AND FUTURE WORK

The preserved classifier case workflow is the part of the project that can be reproduced. Every view in the standalone demo maintained its own score and high-resolution Grad-CAM picture. The saved fingerprints and verification data were also matched with versioned annotations, assigned review, report generation, and lineage-filtered retrieval.

The auxiliary models should not have been activated based on the same evidence. Although informal prediction delivered both labels for 30 out of 31 cases, the source discriminator covered too few realistic input types, and there were many missed lesions in the final

segmentation cohort. Disabling each component means it failed at least one stored acceptance criteria. These results are presented as experimental results, not as proof of how well the treatment worked in the clinic. For more robust assertions, a reliable software baseline and patient-arranged independent review are necessary. In order to say that the whole test suite passed, the route-registration problem must be fixed. The same release should then be confirmed by running clean-process and checking in with the browser multiple times. Additionally, it is necessary to test the classifier threshold, Grad-CAM dimensions, resizing, colour conversion, normalisation, and fixed fixtures. Examining the models on appropriately sourced external data should only be done once such controls have been implemented.

A. Near-term software work

Assembling the Flask application begins with the first software task: registering the readiness routes. Then, the blueprint is attached. Then, it is safe to test the application construction multiple times. Multiple application instances should be created simultaneously during a regression test. Someone else has to swap out an artefact from the evaluated model with another file and make sure the provider won't load the old one. A fixture that covers the model fingerprint, the result for each view, and the finalised reviewer section in the exported document is also needed for report generation.

B. Model evaluation plan

A manifest should be generated before training any new classifier. The patient's information should be included, along with their institution, device (or scanner), reference standard, and desired split (if applicable). All images taken of the same patient should remain in a single split. Model selection, threshold selection, and visual examination should not have any bearing on the external cohort's origin. Sensitivity, specificity, calibration, confidence intervals, and findings for relevant patient and acquisition groups should be included in the report in addition to accuracy, since accuracy alone is insufficient.

It is recommended to wait until the data path is steady before returning to segmentation. Decisions like using a different loss or a bigger input size in a subsequent trial are still part of the development process. Prior to opening the next final cohort, the acceptance criteria should be clarified. There should still be a consolidated reporting of lesion-only overlap, full misses, malignant-lesion outcomes, and false positives on normal imaging. Similar self-control is necessary for abstention: label coverage serves a purpose only if both the replied coverage and the error rate among answered cases are tolerable.

C. Boundary for later clinical study

Technical validation is not enough to go beyond a research demonstration. No held-out metric can account for the scientific, workflow, and adoption concerns brought up by clinical usage [16]. For a prospective study to be conducted, it is necessary to identify the target audience, the workflow step at which the output is generated, the possible next steps, and the

procedure for handling errors. After then, it's up to human reviewers to figure out if the interface makes modifications to decisions securely and in a fashion that users can comprehend. Keep this build marked as a research workspace and note its constraints next to each model output until that task is finished.

REFERENCES

- [1] M. H. Yap et al., "Automated breast ultrasound lesions detection using convolutional neural networks," *IEEE J. Biomed. Health Inform.*, vol. 22, no. 4, pp. 1218-1226, 2018, doi: 10.1109/JBHI.2017.2731873.
- [2] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *MICCAI, LNCS 9351*, pp. 234-241, 2015, doi: 10.1007/978-3-319-24574-4_28.
- [3] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in Brief*, vol. 28, art. 104863, 2020, doi: 10.1016/j.dib.2019.104863.
- [4] N. Vallez, G. Bueño, O. De.niz, M. A. Rie.nda, and C. Pastor, "BUS-UCLM: Breast ultrasound lesion segmentation dataset," *Scientific Data*, vol. 12, art. 242, 2025, doi: 10.1038/s41597-025-04562-3.
- [5] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, pp. 618-626, 2017.
- [6] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proc. 34th ICML, PMLR 70*, pp. 1321-1330, 2017.
- [7] Y. Ovadia et al., "Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift," in *Advances in Neural Information Processing Systems 32*, 2019.
- [8] A. N. Angelopoulos and S. Bates, "Conformal prediction: A gentle introduction," *Found. Trends Mach. Learn.*, vol. 16, no. 4, pp. 494-591, 2023, doi: 10.1561/22000000101.
- [9] J. Monga, L. Moy, and C. E. Kahn, Jr., "Checklist for Artificial Intelligence in Medical Imaging: A guide for authors and reviewers," *Radiol. Artif. Intell.*, vol. 2, no. 2, e200029, 2020, doi: 10.1148/ryai.2020200029.
- [10] B. Vasey et al., "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI," *Nat. Med.*, vol. 28, pp. 924-933, 2022, doi: 10.1038/s41591-022-01772-9.
- [11] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Re, "Hidden stratification causes clinically meaningful failures in machine learning for medical imaging," *Proc. ACM Conf. Health Inference Learn.*, pp. 151-159, 2020, doi: 10.1145/3368555.3384468.
- [12] M. Mitchell et al., "Model cards for model reporting," in *Proc. FAT**, pp. 220-229, 2019, doi: 10.1145/3287560.3287596.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015, arXiv:1409.1556.
- [14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, pp. 770-778, 2016.

- [15] J. Yang et al., "MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Scientific Data*, vol. 10, art. 41, 2023, doi: 10.1038/s41597-022-01721-8.
- [16] C. J. Kelly, A. Karthikesalingam, M. Suleyman, G. Corrado, and D. King, "Key challenges for delivering clinical impact with artificial intelligence," *BMC Med.*, vol. 17, art. 195, 2019, doi: 10.1186/s12916-019-1426-2.