WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

MammoScan AI: An Explainable VGG-16 and SVM Framework for Breast Cancer Screening from Digital Mammograms

Rajesh Moraskar[1], Sujal Shahu[2], Prof. Ashish Trivedi[3]

[1]Dept. of Artificial Intelligence G H Raisoni University Amravati, India
[2]Dept. of Artificial Intelligence G H Raisoni University Amravati, India
[3]Dept. of Artificial Intelligence G H Raisoni University Amravati, India

ABSTRACT :

Breast cancer screening depends on the timely and consistent reading of mammograms, a task that is slow and varies between readers. This paper presents MammoScan AI, an end-to-end web application for automated benign/malignant screening of digital mammograms that runs on a CPU without a GPU. An ImageNet-pretrained VGG-16 network acts as a fixed feature extractor that yields 512-dimensional descriptors, and a Support Vector Machine (SVM), selected from four classical classifiers and an end-to-end CNN by AUC-ROC, produces the malignancy score. The score is calibrated into a probability and grouped into four BI-RADS-style risk levels; Grad-CAM heatmaps show the regions that drove each decision, an eight-step authenticity guard rejects inputs that are not mammograms, and HTML and PDF reports are generated for every scan. Trained on 5,000 augmented mammograms derived from INbreast and evaluated on 500 hold-out images, the deployed SVM reaches an accuracy of 65.4%, precision of 68.97%, recall of 56.0%, F1-score of 61.81% and AUC-ROC of 0.742, with about 140 ms inference per scan on a free cloud CPU instance. The modest recall and the limits of augmentation-based evaluation are discussed candidly. MammoScan AI is an academic demonstration and not a medical device.

Index Terms—breast cancer screening, mammography, deep learning, transfer learning, VGG-16, support vector machine, Grad-CAM, BI-RADS.

I. INTRODUCTION

Breast cancer is the most commonly diagnosed cancer among women, with about 2.3 million new cases and 685,000 deaths estimated worldwide in 2020 [1]. Screening mammography is the principal tool for finding it early, but every examination must be read by a trained radiologist, and in many regions that capacity is scarce. Early computer-aided detection (CAD) was not shown to improve screening accuracy [2], whereas deep learning [3] has produced systems whose performance is comparable to, and in some settings better than, that of radiologists [4]. Such results, however, typically rely on very large datasets, specialised hardware and models that return a verdict without explaining it.

This paper asks how much of that benefit can be delivered by a student-scale system with no GPU, no proprietary data and a free cloud instance. We present MammoScan AI, which uses a pretrained VGG-16 network only as a feature extractor and delegates the decision to a classical classifier. Around this core, we build the components that a bare classifier lacks: input validation, a calibrated risk presentation, a visual explanation and a printable report. The contributions are:

- a lightweight VGG-16 + SVM screening pipeline, benchmarked against Random Forest, Logistic Regression, Decision Tree and an end-to-end CNN on identical data;
- a calibration and BI-RADS-style risk-stratification layer with a bi-directional prediction bar;
- an eight-step authenticity guard, Grad-CAM explanations and HTML/PDF report generation in one deployable web application; and
- an explicit account of the limitations of the evaluation, including a low recall of 56.0%.

Section II reviews related work, Section III describes the system, Section IV reports the results, Section V discusses limitations, and Section VI concludes.

II. RELATED WORK

Deep Learning in Mammography

Lehman et al. found that CAD in digital screening mammography was not associated with improved accuracy [2]. Deep networks changed this picture: Ribli et al. trained a Faster R-CNN lesion detector and evaluated it on INbreast [5], Shen et al. trained an end-to-end classifier on screening mammograms [6], and McKinney et al. reported performance comparable to or better than that of radiologists on UK and US screening data [4]. The strongest results depend on large archives and long training on specialised hardware. INbreast [7] and CBIS-DDSM [8] are the usual public datasets; we use INbreast.

Transfer Learning and Classical Classifiers

Features learned on large natural-image collections such as ImageNet [9] transfer to other domains, with early and middle layers transferring best [10], [11]. VGG-16 [12] stacks 3 × 3 convolutions to sixteen weight layers and yields uniform feature maps that suit both feature extraction and Grad-CAM. On the resulting descriptors, a Support Vector Machine [13] finds the maximum-margin boundary, and Random Forests [14] combine many randomised trees; both are standard choices for moderate sample sizes.

Explainability and Calibration

Grad-CAM [15] uses gradients flowing into the last convolutional layer to localise the regions that matter for a prediction, without architectural changes. Classifier scores are not probabilities; Platt scaling [16] is the classical remedy for SVMs. Breast-imaging findings are reported with BI-RADS categories [17], so presenting a calibrated probability in bands resembling BI-RADS makes an output easier to place in clinical language.

III. PROPOSED SYSTEM

Overview

Fig. 1 shows the processing pipeline. An uploaded JPG or PNG mammogram first passes the authenticity guard, is resized to 224 × 224 pixels, and is converted by VGG-16 into a 512-dimensional feature vector. The SVM score is calibrated into a malignancy probability and mapped to a risk level. Grad-CAM overlays and reports are produced on request. The server is a Flask [18] REST application served by Gunicorn, and inference uses the LiteRT runtime [19] on Linux so that it fits within the 512 MB memory of a free cloud instance.

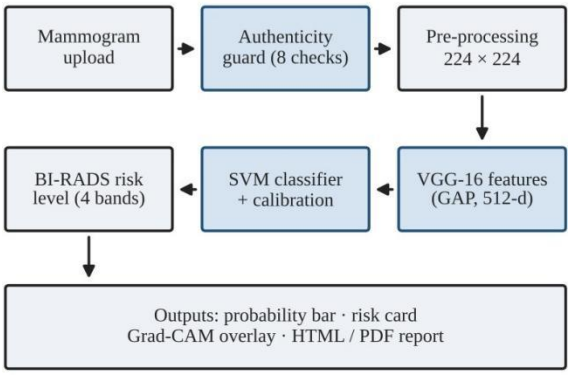


Fig. 1. Processing pipeline of MammoScan AI.

Dataset and Augmentation

Images come from INbreast [7]; the interface records 410 source images (310 benign, 100 malignant). To obtain a balanced set, the source images were expanded by augmentation into 5,000 training images (2,500 per class) and 500 validation images (250 per class). Transforms were chosen to mimic real examination variability: horizontal flip, rotation (±15°), CLAHE contrast enhancement [20], brightness and gamma jitter, Gaussian noise, elastic deformation and random zoom crop.

Feature Extraction

VGG-16 with ImageNet weights [12] is loaded without its classification head, and global average pooling over the last convolutional block yields a 512-dimensional vector per image. The network is not retrained; features are computed once and cached, which makes classifier experiments fast and CPU-only.

Classifiers and Model Selection

An SVM [13], a Random Forest [14], Logistic Regression and a Decision Tree were trained on the cached features using scikit-learn [21], and an end-to-end VGG-16 network served as a reference. With malignant as the positive class, accuracy, precision, recall and F1-score are defined from the counts of true and false positives and negatives:

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

$$P = TP / (TP + FP),\ R = TP / (TP + FN) \quad (2)$$

$$F1 = 2PR / (P + R) \quad (3)$$

The area under the ROC curve (AUC-ROC) is the probability that a random malignant image scores higher than a random benign one. Because it is independent of the decision threshold, it was used as the selection criterion.

Calibration and Risk Stratification

The raw SVM score is not a probability [16]. The application passes it through a custom non-linear calibration curve to obtain a malignancy probability, which is grouped into the four levels of Table I and described in BI-RADS terms [17]. A bar centred on 50% extends left in green for benign findings and right in red for malignant findings. The curve is designed for presentation and has not been fitted to clinical outcomes, so the probability is indicative only.

Authenticity Guard

A classifier answers for any image it is given, so every upload is checked by eight image-statistics tests before inference: (1) aspect ratio; (2) colour channels, which rejects colour photographs; (3) intensity distribution and contrast; (4) dark background ratio between 10% and 85%; (5) at least one dark corner; (6) a chest-wall border with at least 35% tissue coverage;
(7) hollow-object detection by contour hierarchy; and (8) edge density, which rejects mechanical objects. The first failed check returns HTTP 400 with the reason.

Explanation, Reports and Deployment

On request, a Grad-CAM heatmap is overlaid on the mammogram, and a styled HTML report or a formal PDF report (generated with fpdf2) is produced; both repeat that the result is not a diagnosis. The service exposes six endpoints (health, metadata, prediction, explanation and two report formats) and is deployed with a Render Blueprint using Gunicorn with two workers.

IV. EXPERIMENTAL RESULTS

Setup

All models were trained on the 5,000 training images (or their cached features) and evaluated on the 500 validation images, which the classifiers do not see during training. Experiments ran on CPU only.

Classifier Comparison

Table II and Fig. 2 compare the five models. The SVM obtained the highest AUC-ROC (0.742) and was deployed. The end-to-end VGG-16 CNN obtained slightly higher accuracy (66.4%), recall (58.4%) and F1-score (63.48%) but a lower AUC-ROC (0.711). With 500 validation images, the standard error of an accuracy near 65% is about 2.1 percentage points, so these accuracy and recall differences are within sampling uncertainty; the SVM was chosen because it ranks cases better and is far cheaper to train, store and run on a CPU.

TABLE II
BENCHMARK ON THE 500 VALIDATION IMAGES

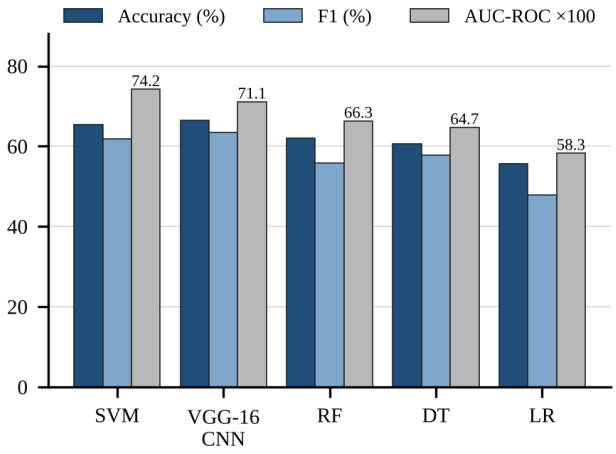
Model	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)	AUC
SVM (deployed)	65.4	68.97	56.0	61.81	0.742
VGG-16 end-to-end	66.4	69.52	58.4	63.48	0.711
Random Forest	62.0	66.67	48.0	55.81	0.663
Decision Tree	60.6	62.21	54.0	57.82	0.647
Logistic Regression	55.6	57.95	40.8	47.89	0.583

TABLE I

Level	Malignancy probability	Description
Low	< 20%	Routine screening (BI-RADS 1–2)
Moderate	20–49%	Intermediate band
Elevated	50–74%	Intermediate band
High	≥ 75%	High suspicion (BI-RADS 5)

RISK LEVELS USED BY THE APPLICATION

Fig. 2. Accuracy, F1-score and AUC-ROC (×100) of the five models.



Confusion Matrix

Table III gives the confusion matrix of the deployed SVM. Specificity is 187/250 = 74.8%, the false-positive rate is 25.2%, and the negative predictive value is 187/297 = 63.0%. The system makes more false negatives (110) than false positives (63), which is the less desirable direction for screening.

TABLE III
CONFUSION MATRIX OF THE SVM (500 IMAGES)

	Pred. benign	**Pred. malignant**
Actual benign	187 (TN)	63 (FP)
Actual malignant	110 (FN)	140 (TP)

Runtime and Verification

On the free Render instance, the LiteRT runtime analyses a scan in about 140 ms. On a Windows development machine with the TensorFlow Lite fallback, a scan takes about 420 ms after a first model load of about 7 s. An automated suite of eight end-to-end tests (health, metadata, interface, malignant and benign samples, invalid inputs, Grad-CAM and PDF report) passes on the application; these tests verify software behaviour and not classification accuracy.

System Outputs

Fig. 3 shows the interface for a benign validation sample: the prediction bar reports 98.0% benign and 2.0% malignant, the risk card shows Low risk (routine screening, BI-RADS 1–2), and the Grad-CAM panel overlays the heatmap on the original mammogram.

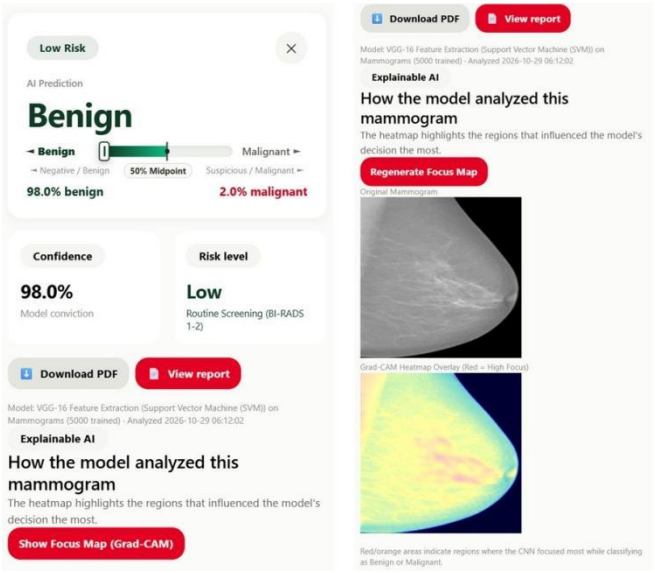


Fig. 3. (a) Prediction result with risk level and report buttons. (b) Grad-CAM heatmap overlay.

V. DISCUSSION AND LIMITATIONS

The results show that a fixed ImageNet feature extractor with an SVM head is a workable, CPU-friendly baseline for mammogram classification, matching an end-to-end network within the uncertainty of the experiment at a small fraction of the cost. The engineering around the model, namely the guard, risk bands, heatmap and reports, changes what a user can do with an answer without changing its accuracy.

Several limitations must be stated plainly. First, recall is 56.0%: 110 of 250 malignant validation images are classified as benign, so the system must not be used to rule out cancer. Second, the training and validation images were produced by augmenting a small source set (410 images), so they may not be fully independent, and the figures should be read as a comparison of methods and not as clinical performance; a patient-level split before augmentation is needed. Third, the calibration curve is not fitted to outcomes, and Grad-CAM has not been compared with radiologists' annotations. Fourth, only single JPG/PNG images

are analysed, and all data come from one database. No study with radiologists has been carried out.

VI. CONCLUSION AND FUTURE WORK

We presented MammoScan AI, a complete, explainable and deployable screening aid built on VGG-16 features and an SVM, achieving an AUC-ROC of 0.742 and an accuracy of 65.4% on 500 hold-out images at about 140 ms per scan on a CPU. The results demonstrate feasibility, not clinical accuracy. Future work will adopt patient-level splitting on larger multi-source datasets, fine-tune the network and tune a recall-oriented operating point, accept DICOM input and multiple views, add lesion localisation, and evaluate the system in a reader study with radiologists.

ACKNOWLEDGMENT

The authors thank the creators of the INbreast database and the developers of the open-source tools used in this work.

REFERENCES :

1. H. Sung et al., "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA Cancer J. Clin.*, vol. 71, no. 3, pp. 209–249, May 2021.
2. C. D. Lehman et al., "Diagnostic accuracy of digital screening mammography with and without computer-aided detection," *JAMA Intern. Med.*, vol. 175, no. 11, pp. 1828–1837, Nov. 2015.
3. Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.
4. S. M. McKinney et al., "International evaluation of an AI system for breast cancer screening," *Nature*, vol. 577, pp. 89–94, Jan. 2020.
5. D. Ribli, A. Horváth, Z. Unger, P. Pollner, and I. Csabai, "Detecting and classifying lesions in mammograms with deep learning," *Sci. Rep.*, vol. 8, Art. no. 4165, 2018.
6. L. Shen, L. R. Margolies, J. H. Rothstein, E. Fluder, R. McBride, and W. Sieh, "Deep learning to improve breast cancer detection on screening mammography," *Sci. Rep.*, vol. 9, Art. no. 12495, 2019.
7. I. C. Moreira et al., "INbreast: Toward a full-field digital mammographic database," *Acad. Radiol.*, vol. 19, no. 2, pp. 236–248, Feb. 2012.
8. R. S. Lee et al., "A curated mammography data set for use in computer-aided detection and diagnosis research," *Sci. Data*, vol. 4, Art. no. 170177, 2017.
9. J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE CVPR*, 2009, pp. 248–255.
10. J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," in *Proc. NIPS*, 2014, pp. 3320–3328.
11. S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345–1359, Oct. 2010.
12. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, 2015. [Online]. Available: https://arxiv.org/abs/1409.1556
13. C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
14. L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
15. R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE ICCV*, 2017, pp. 618–626.
16. J. C. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in Large Margin Classifiers*. Cambridge, MA, USA: MIT Press, 1999, pp. 61–74.
17. C. J. D'Orsi et al., *ACR BI-RADS Atlas: Breast Imaging Reporting and Data System*, 5th ed. Reston, VA, USA: American College of Radiology, 2013.
18. Pallets Projects, "Flask documentation." [Online]. Available: https://flask.palletsprojects.com/
19. Google AI Edge, "LiteRT overview." [Online]. Available: https://ai.google.dev/edge/litert
20. K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. San Diego, CA, USA: Academic Press, 1994, pp. 474–485.
21. F. Pedregosa et al., "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.